# ITISE 2019

## International Conference on Time Series and Forecasting

# PROCEEDINGS OF PAPERS

Volumen 1

# ITISE 2019
# International Conference on Time Series and Forecasting

## Proceedings of Papers
## 25-27 September 2019
## Granada (Spain)

**Editors and Chairs**
Olga Valenzuela
Fernando Rojas
Héctor Pomares
Ignacio Rojas

**I.S.B.N: 978-84-17970-78-9**
**Legal Deposit: Gr 1209-2019**
**Edit and Print:** Godel Impresiones Digitales S.L.

# Preface

We are proud to present the set of final accepted papers for the 6th International conference on Time Series and Forecasting (ITISE 2019) held in Granada (Spain) during September, 25th-27th, 2019.

The ITISE 2019 seeks to provide a discussion forum for scientists, engineers, educators and students about the latest ideas and realizations in the foundations, theory, models and applications for interdisciplinary and multidisciplinary research encompassing disciplines of computer science, mathematics, statistics, forecaster, econometric, etc, in the field of time series analysis and forecasting.

The aims of ITISE 2019 is to create a friendly environment that could lead to the establishment or strengthening of scientific collaborations and exchanges among attendees, and therefore, ITISE 2019 solicits high-quality original research papers (including significant work-in-progress) on any aspect time series analysis and forecasting, in order to motivating the generation, and use of knowledge and new computational techniques and methods on forecasting in a wide range of fields.

The list of topics in the successive Call for Papers has also evolved, resulting in the following list for the present edition:

1. **Time Series Analysis and Forecasting**.
    - Nonparametric and functional methods
    - Vector processes
    - Probabilistic Approach to Modeling Macroeconomic Uncertainties
    - Uncertainties in forecasting processes
    - Nonstationarity
    - Forecasting with Many Models. Model integration
    - Forecasting theory and adjustment
    - Ensemble forecasting
    - Forecasting performance evaluation
    - Interval forecasting
    - Econometric models
    - Econometric Forecasting
    - Data preprocessing methods: Data decomposition, Seasonal adjustment, Singular spectrum analysis, Detrending methods, etc.
2. **Advanced method and on-Line Learning in time series**.
    - Adaptivity for stochastic models
    - On-line machine learning for forecasting
    - Aggregation of predictors
    - Hierarchical forecasting
    - Forecasting with Computational Intelligence
    - Time series analysis with computational intelligence

- Integration of system dynamics and forecasting models

3. **High Dimension and Complex/Big Data**.

   - Local Vs Global forecast
   - Techniques for dimension reduction
   - Multiscaling
   - Forecasting Complex/Big data

4. **Forecasting in real problem**.

   - Health forecasting
   - Telecommunication forecasting
   - Modelling and forecasting in power markets
   - Energy forecasting
   - Financial forecasting and risk analysis
   - Forecasting electricity load and prices
   - Forecasting and planning systems
   - Real time macroeconomic monitoring and forecasting
   - Applications in: energy, finance, transportation, networks, meteorology, health, research and environment, etc.

After a careful peer review and evaluation process (each submission was reviewed by at least 2, and on the average 2.9, program committee members or additional reviewer). In this proceedings we are presetting the abstract of the contribution to be presented during ITISE-2019 (accepted for oral, poster or virtual presentation,according to the recommendations of reviewers and the authors' preferences).

In this edition of ITISE, we are honored to have the following invited speaker:

1. Prof. Per Bjarte Solibakke, Professor and Associate Dean for Education, Faculty of Economics, Norwegian University of Science and Technology — NTNU Department of International Business.Vice Dean for Education, Faculty of Ecdonomics and Management, Department of International Business.

2. Prof. Thorsten Lehnert, Full professor of Finance. Luxembourg School of Finance (LSF)

3. Prof. Dieter Nautz, Professor Freie Universitt Berlin. Fachbereich Wirtschaftswissenschaft. Chair of Econometrics.

4. Prof. Dr. Stephan Schlter, Professor. Faculty of Mathematics, Natural and Economic Sciences University of Applied Sciences Ulm .

5. Prof. J. Hinaunye Eita, Professor and Head of Academics School of Economics College of Business and Economics . University of Johannesburg .

This new edition of ITISE was organized at the Universidad de Granada, with the help of the Spanish Network Time Series (RESET). We wish to thank to our main sponsor the institutions Faculty of Science, Dept. Computer Architecture & Computer Technology and CITIC-UGR from the University of Granada for their support. We wish also to thank to the Dr. Veronika Rosteck and Dr. Eva Hiripi, Springer, Associate Editor, for their interest in the future editing a book series of Springer from the best papers of ITISE 2019.

We would also like to express our gratitude to the members of the different committees and to the reviewer for their support, collaboration and good work.

September, 2019
Granada

ITISE Editors and Chairs
Olga Valenzuela
Fernando Rojas
Hector Pomares
Ignacio Rojas

.

# Program Committee

| | |
|---|---|
| Tatyana Afanaseva | Ulyanovsk State Technical University |
| Dorel Aiordachioaie | University Dunarea de Jos of Galati |
| Cagdas Hakan Aladag | Hacettepe University |
| Jose M. Amigo | Universidad Miguel Hernandez |
| Josu Arteche | University of the Basque Country UPV/EHU |
| Marcel Ausloos | GRAPES |
| Rosangela Ballini | IE - DTE - UNICAMP |
| Oresti Banos | University of Granada |
| Josep Lluís Carrion-I-Silvestre | Universitat de Barcelona |
| German Castellanos | Universidad Nacional de Colombia |
| João P. S. Catalão | University of Porto |
| Miguel Damas | University of Granada |
| Lee Chang-Yong | Kongju National University |
| Marijana Cosovic | University of East Sarajevo, Faculty of Electrical Engineering |
| Pierpaolo D'Urso | Sapienza University of Rome |
| Ricardo de A. Araújo | Laboratório de Inteligência Computacional do Araripe / Instituto Federal do Sertão Pernambucano |
| Lola Gadea | University of Zaragoza |
| Alberto Guillen | University of Granada |
| Jesus Gonzalo | U. Carlos III de Madrid |
| Ferda Halicioglu | Istanbul Medeniyet University |
| Luis Javier Herrera | University of Granada |
| Tzung-Pei Hong | Department of Computer Science and Information Engineering, National Univesity of Kaohsiung |
| Plamen Ch. Ivanov | Boston University |
| Ivan Izonin | Lviv Polytechnic National University |
| Samrad Jafarian-Namin | Yazd University |
| Vinayakam Jothiprakash | |
| Emina Junuz | Faculty of information technologies |
| Sreekanth K J | KISR |
| Rebecca Killick | Lancaster University |
| Alexey Koronovskiy | Saratov State University, Faculty of Nonlinear Processes |
| Dalia Kriksciuniene | Vilnius University |
| Elmar Lang | University of Regensburg |
| Hooi Hooi Lean | Universiti Sains Malaysia |
| Thorsten Lehnert | Luxembourg School of Finance |
| Chunshien Li | Department of Information Management, National Central University |
| Carlos Lima | University of Brasilia |
| Hui Liu | Central South University, China & University of Rostock, Germany |
| Songan Mao | Purdue University |
| Francisco Martínez-Álvarez | Universidad Pablo de Olavide |
| Janusz Miśkiewicz | University of Wrocław |
| Antonio Montañés | |
| Miquel Montero | Universitat de Barcelona |

| | |
|---|---|
| Fionn Murtagh | University of Huddersfield |
| Guy Mélard | Université libre de Bruxelles |
| P. C. Nayak | National Institute of Hydrology |
| Juan M. Palomo-Romero | University of Córdoba |
| Eros Pasero | Politecnico di Torino |
| Fernando Perez De Gracia | Universidad de Navarra |
| Irina Perfilieva | University of Ostrava |
| Hector Pomares | University of Granada |
| María Dolores Pérez Godoy | Departamento de Informática. Universidad de Jaén |
| Vadlamani Ravi | IDRBT, Hyderabad |
| Antonio Jesús Rivera Rivas | Departamento de Informática. Universidad de Jaén |
| Paulo Rodrigues | Banco de Portugal |
| Ignacio Rojas | University of Granada |
| Heather Ruskin | Dublin City University |
| Kalle Saastamoinen | National Defence University of Finland |
| Reza Sadeghi | Department of Computer Science and Engineering, Kno.e.sis Research Center, Wright State University |
| Francois Schmitt | CNRS |
| Thanasis Sfetsos | NCSR Demokritos |
| Leonid Sheremetov | Mexican Petroleum Institute |
| Yixiao Sun | University of California San Diego |
| Leopold Sögner | Institute for Advanced Studies |
| Ryszard Tadeusiewicz | AGH University of Science and Technology, Krakow, Poland |
| Mohsen Talebsafa | University of Texas at Arlington |
| Chor Foon Tang | Universiti Sains Malaysia |
| Alicia Troncoso | Universidad Pablo de Olavide |
| Mehdi Vafakhah | Tarbiat Modares University |
| Olga Valenzuela | University of Granada |
| Dimitris Varoutas | National and Kapodistrian University of Athens, Facutly of Informatics & Telecommunications |
| Claudia Villalonga | Universidad Internacional de La Rioja |
| Martin Wagner | Faculty of Statistics, Technical University Dortmund |
| Michael Wolf | University of Zurich |
| Slawomir Zadrozny | Systems Research Institute, Polish Academy of Sciences |

# Table of Contents

**Session A.3: Energy forecasting (Part I)**



**Session B.3: Forecasting theory and adjustment**



**Session A.4: Dimension reduction techniques in Time Series**



**Session B.4:Real macroeconomic monitoring and forecasting (Part I)**

**Session B.7: Forecasting with Many Models**



**Session A.8: Econometric models (Part II)**



**Session B.8: Financial forecasting and risk analysis**



**Session A.9: Forecasting Complex/Big data (Part I)**



**Session B.9: Vector processes**

# WHY IS THE MARKET SKEWNESS-RETURN RELATIONSHIP NEGATIVE?

THORSTEN LEHNERT*

May 2019

**Abstract**

The observed negative relationship between market skewness and excess return or the negative price of market skewness risk in the cross-section of stock returns is somewhat counterintuitive when we consider the usual interpretation of e.g. option-implied skewness as an indicator of jump risk or downside risk. One possible explanation for this inconsistency is that there are factors affecting option-implied market skewness other than jump risk in the stock market. In this paper, I find that price pressure associated with "crowded trades" of mutual funds is an important endogenous factor. Given that retail investors are prone to herding, the directional trading of mutual funds is correlated, and their collective actions can generate short-term price pressure on aggregate stock prices. Short sellers systematically exploit these patterns not only in the equity lending market, but also in the options market. In line with this economic channel, I find that firstly, the significant negative relationship between market skewness and returns becomes insignificant, once I control for price pressure. Secondly, the negative relationship is only present for the "bad" downside component of risk-neutral skewness, associated with out-of-the-money put options. For the "good" upside component of risk-neutral skewness, associated with out-of-the-money call options, the relationship is always positive. Thirdly, price pressure affects the skewness-return relationship, which can be clearly distinguished from the impact of flows on the volatility-return relationship in terms of the leverage effect.

Keywords: Mutual Funds, Index Options, Fund Flows, Short Sellers, Risk-neutral Skewness.
JEL-Classification: G12, C15.

*Thorsten Lehnert is at the Luxembourg School of Finance, University of Luxembourg, 6, rue Richard Coudenhove-Kalergi, 1359 Luxembourg, Luxembourg, tel +352466644-6941; fax +352466644-6835. E-mail address: thorsten.lehnert@uni.lu .

# Two Algorithms to Identify Outliers in Large Climate Time Series

Stephan Schlueter[1] and Milena Kresoja[2]
(1) University of Applied Sciences Ulm (Germany), (2) Institut ekonomskih nauka (Serbia)

## 1 Abstract

Nowadays, most technical devices generate and store data in the form of time series. They are widely used in climatology, where e.g. water levels are collected for dike construction, in agriculture or the energy sector, for water management, and many other industries. Such time series help farmers, for example, to calculate the risk of frost in April, or an energy company to estimate the photovoltaic production, which - among other things - depends on the temperature.

Most of the surface climate data such as air temperature, wind speed, or solar radiation are recorded in high frequency by automated instruments at many sites. These series are likely to contain outliers or irregular patterns caused e.g. by measurement errors, transmission errors, or systems changes, which, as a result, leads to false model specification, skewed parameter estimates, and forecasting errors. Data quality is crucial, which is exacerbated by the fact that climate is a complex system showing elements like high dimensionality, multiple seasonality, and serial dependence. From a practitioner's point of view, setting up a model is the easy part, it's the data preprocessing step which requires most attention and is the most time-consuming part of data analysis. Hence, having a reliable preprocessing method is essential.

Literature offers a huge variety of methods for detecting anomalies in time series. For an overview, please refer to Gupta et al. (2014). Here we propose two further ones especially designed to treat clime time series. To overcome the problem of complexity of most algorithms, we first introduce a rather simple approach for anomaly detection based on autoregressive cost updates (ACU). The algorithm is easy to implement and to communicate, and it detects anomalies without the necessity of calibrating a model. Besides, we test a more sophisticated alternative, which is based on a method from signal processing called wavelet transform (see e.g. Mallat, 2003). Here we split up a time series into a linear combination of different frequencies and identify irregularities among the highest frequencies. We apply the model to temperature data. However, the model can be easily adapted to other climate time series such as humidity, precipitation, wind speed, but also to time series generated in other fields and applications. The performance of our proposed algorithms is tested on synthetic time series and benchmarked with existing preprocessing algorithms. Thereby we see that there is no method that performs best in all scenarios and for all tested quality criteria. If a time series is supposed to contain mostly solitary outliers, we recommend to apply the ACU algorithm. Otherwise, for multiple consecutive outliers, the wavelet-based algorithm is the best choice among the tested alternatives.

Additionally, the algorithms are applied to a multivariate data set of temperatures at various sites in the City of Novi Sad, Serbia. We thereby especially focus on identifying irregular patterns that are not clear outliers, i.e. hard to identify. We see that both algorithms effectively handle large data sets, which is the basis for applying further algorithms e.g. for clustering (i.e. dimension reduction) or modeling the data.

**Literature:**

Gupta M, Gao J, Aggarwal C, Han J (2014). Outlier detection for temporal data. Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1); pp. 1-129.

Mallat S. A wavelet tour of signal processing, 2nd ed. . Academic Press: Manchester, 2003..

# Divisia Monetary Aggregates for a Heterogeneous Euro Area

Maximilian Brill*, Dieter Nautz† and Lea Sieckmann†

†Freie Universität Berlin
Department of Economics

*University of Antwerp
Department of Economics

This version: May 16, 2019

We introduce a Divisia monetary aggregate for the euro area that accounts for the heterogeneity across member countries both, in terms of interest rates and the decomposition of monetary assets. In most of the euro area countries, the difference between the growth rates of the country-specific Divisia aggregate and its simple sum counterpart is particularly pronounced before recessions. The results obtained from a panel probit model confirm that the divergence between the Divisia and the simple sum aggregate has a significant predictive content for recessions in euro area countries.

*Keywords*: Monetary aggregation, Euro area Divisia aggregate, Recessions.

*JEL classification*: E51, E32, C43

*University of Antwerp, Department of Economics, Prinsstraat 13, B-2000 Antwerp, Belgium. E-mail: maximilian.brill@uantwerpen.be.

†Freie Universität Berlin, Department of Economics, Boltzmannstrasse 20, 14195 Berlin, Germany. Tel.: +49(0) 30838 51399. E-mail: dieter.nautz@fu-berlin.de. E-mail: lea.sieckmann@fu-berlin.de.

# 1 Introduction

The role of money for monetary policy analysis has changed remarkably in recent years. In the early years of the European Monetary Union, for example, the European Central Bank (ECB) placed a lot of emphasis on the prominent role of monetary aggregates for its monetary policy analysis. The ECB even published a reference value for money growth in order to explain its interest rate decisions. Yet, this prominent role of money has never been beyond controversy. On the one hand, the empirical literature raised doubts on the stability of money demand and, thus, on the information content of monetary aggregates for inflation and output. On the other hand, the theoretical literature assumed that monetary policy is fully reflected in interest rates and money virtually disappeared from standard macro models. In accordance with the declining role of money for both, monetary theory and monetary policy practice, the ECB downplayed the role of monetary aggregates for its interest rate decisions, see e.g. European Central Bank (2003) or Constâncio (2018).

Since the outbreak of the financial crisis, however, there has been a renewed interest in the analysis of monetary aggregates. With interest rates at the zero lower bound, central banks increasingly use monetary aggregates to assess the effectiveness of their unconventional monetary policy measures. However, traditional simple sum aggregates may not accurately measure the quantities of monetary services and the availability of liquidity. Following Barnett (1980), monetary analysis should be based on Divisia aggregates where different monetary components, like currency and time-deposits, are weighted by their individual and time-varying opportunity cost. In contrast to their simple sum counterparts, Divisia aggregates account for the substitution effects between different types of monetary assets. There is increasing empirical evidence that Divisia aggregates contain useful information for the real economy. Early evidence of superior forecasting ability of U.S. Divisia aggregates for output relative to simple sum aggregates is provided by Schunk (2001). More recently, Belongia and Ireland (2015) show that Divisia aggregates can improve output forecasts for the United States. Barnett and Chauvet (2011) observe that U.S. monetary aggregates and their Divisia counterparts diverge particularly during times of high uncertainty indicating

that this divergence can be used as a signal for impending recessions.

A small but increasing number of central banks publish Divisia aggregates, including the Bank of England (Hancock, 2005) and the Federal Reserve Bank of St. Louis (Anderson and Jones, 2011). Divisia monetary aggregates for the United States are also provided by the Center of Financial Stability (CFS), see Barnett et al. (2013). Stracca (2004) made a first attempt to compute a Divisia monetary aggregate for the euro area. Assuming that euro area countries have already converged, he applied a single euro area wide interest rate for each of the monetary assets. More recently, Darvas (2015) proposed a Divisia aggregate for the euro area under similar homogeneity assumptions. However, since the run-up to the great recession, there has been a significant degree of heterogeneity in the level of interest rates and the composition of monetary assets in the euro area. Therefore, this paper proposes a new euro area wide Divisia aggregate that allows for both, country-specific interest rates and heterogeneous monetary developments.[1] To that aim, we follow Barnett (2007) who developed a theory for monetary aggregation across countries.

Our results show that country-specific monetary developments should not be ignored in the euro area. Particularly since the outbreak of the financial crisis, user cost and expenditure shares of monetary assets and, thereby, Divisia aggregates differ significantly across euro area countries. In line with the findings of Barnett and Chauvet (2011) obtained for the U.S., the divergence between simple sum and Divisia aggregates seems to be particularly pronounced around recession periods. Therefore, we employ a panel probit analysis to investigate whether the divergence between simple sum and Divisia aggregates can predict recessions in individual euro area countries.

The rest of the paper is structured as follows. Section 2 recalls how to compute Divisia aggregates in a heterogeneous currency union. Section 3 presents and discusses the data. Section 4 analyzes the Divisia aggregates and its components at a country level. The focus of Section 5 is on the resulting euro area wide aggregate. Section 6 investigates the predictive content of monetary aggregates for recessions and Section 7 concludes.

[1] In doing so, we partly build on Barnett and Gaekwad (2018) and Chen and Nautz (2015) with, however, some important deviations regarding country and data selection, see Section 3 for more details.

## 2 Monetary Aggregation

### 2.1 Simple Sum Aggregates

Defining and measuring the amount of money in the economy is not straightforward. On the one hand, monetary aggregates differ because they include different types of assets. While narrow aggregates may include only currency in circulation and overnight deposits, broader measures additionally consider short term savings deposits or even debt securities. Table 1 shows the various types of monetary assets that are used by the ECB and many other central banks.

Table 1 Monetary Aggregates

| Monetary Asset | M1 | M2 | M3 |
|---|---|---|---|
| Currency in circulation | x | x | x |
| Overnight Deposits | x | x | x |
| Deposits with agreed maturities of up to 2 years | | x | x |
| Deposits redeemable at notice of up to 3 month | | x | x |
| Repurchase agreements | | | x |
| Money market fund shares/units | | | x |
| Debt securities with a maturity of up to two years | | | x |

Notes: The Table presents the components of the three common monetary aggregates in the euro area, following the definition by European Central Bank (2012).

On the other hand, it is not obvious how different asset types should be aggregated. The widely-used monetary aggregates M1, M2 and M3 simply add up the asset quantities implying that different monetary assets are treated as perfect substitutes. Simple sum aggregates do not take into account the different degrees of liquidity provided by its components. Therefore, simple sum monetary aggregates do not change even in the presence of large shifts in their composition and, thus, in the availability of money. Consider, for example, a situation where time deposits are withdrawn on a large scale and completely changed into cash. In this extreme scenario, the liquidity conditions of the economy change dramatically

but the simple sum monetary aggregate remains unchanged. Disregarding differences in opportunity costs and therefore the substitution effect between monetary assets may lead to a distorted picture of liquidity services available in the economy. According to Belongia and Ireland (2014, p.5), the only question about simple sum aggregates is the *magnitude* of their measurement error.

## 2.2 Divisia monetary aggregates

Barnett (1980) applies aggregation and statistical index number theory to derive the optimal aggregate measure of liquidity services. The Divisia aggregate incorporates the concept of user costs developed by Barnett (1978), which can be interpreted as the opportunity costs of a monetary asset, i.e. how much a consumer is willing to give up in order to hold a certain asset. The assets are weighted accordingly, with more liquid assets receiving a higher weight. Specifically, the Divisia aggregate $D_t$ is defined in terms of its growth rate by:

$$\ln D_t - \ln D_{t-1} = \sum_i v_{it}(\ln M_{it} - lnM_{it-1}), \tag{1}$$

where $M_{it}$, the quantity of monetary asset $i$ in period $t$, is weighted by $v_{it}$, the two-period average of its expenditure share $s_{it}$:

$$s_{it} = \frac{p_{it}M_{it}}{\sum p_{it}M_{it}}. \tag{2}$$

$p_{it}$ denotes the user cost of asset $i$ in period $t$ in discrete time:

$$p_{it} = \frac{R_t - r_{it}}{R_t + 1} \tag{3}$$

where $r_{it}$ denotes the rate of return on asset $i$ in period $t$ and $R_t$ is the benchmark rate. The benchmark rate is the expected yield on a pure investment, i.e. an asset that provides no services other than its yield. The user cost can therefore be interpreted as the interest which is given up in order to hold a liquid monetary asset.

There are two cases where a Divisia and its corresponding simple sum aggregate provide the same information and will move in parallel. First, Divisia and simple sum aggregates can only differ if the underlying monetary assets are actually heterogeneous, i.e. if different assets have different opportunity cost $(p_{it})$. In recent years, however, deposit rates $(r_{it})$ have converged to zero in many euro area countries for most of the monetary assets. As a result, opportunity cost of different assets coincide (Eq. (3)) and the growth rates of Divisia and simple sum aggregates can be expected to be similar. Second, irrespective of the user cost, Divisia and simple sum aggregates grow with the same rate if the various monetary assets $(M_{it})$ grow with identical rates, see Eq. (1). By contrast, the difference between a Divisia index and its simple sum counterpart should be particularly pronounced in uncertain times when the composition of money holdings may change significantly. Consequently, Barnett and Chauvet (2011) suggest that the divergence between the Divisia and its simple sum counterpart could be a useful indicator for recessions.

### 2.3 Divisia monetary aggregates in a currency union

The previous subsection discussed monetary aggregation within a single country. Let us now turn to monetary aggregation across countries in order to define a Divisia aggregate for a currency union. Barnett (2007) developed a theory for the aggregation across countries assuming different degrees of homogeneity. At the one end of the scale, he considers a perfectly homogenous currency union where money demand characteristics and user costs for each monetary asset coincide across countries. This assumption may be less critical for the pre-crisis period when both, short- and long-term interest rates were very similar across the euro area. However, in the run-up to the great recession and during the European debt crisis longer-term interest rates diverged significantly between crisis- and non-crisis countries. In such periods, benchmark rates and, thereby, user cost for the same type of monetary asset could be very different across euro area countries. At the other end of the scale, Barnett (2007) considers a multi-country area with distinct currencies and time-varying exchange rates. In the following, we apply this model to the case of a currency union. Thus, while

the exchange rate is constant, the member countries of the currency union are still heterogeneous because the growth rates of certain monetary assets and the corresponding user cost are allowed to vary between countries.

In line with Barnett (2007), the construction of the area wide aggregate proceeds in two steps. In a first step, Divisia aggregates $D_k$ for each individual country $k$ are defined according to Equation (1). In a second step, the country-specific Divisia indices are aggregated to the area wide Divisia index $DMU$ as follows:

$$\ln DMU_t - \ln DMU_{t-1} = \sum_k V_{kt} \left[ \ln \left( h_{kt} D_{kt} \right) - \ln \left( h_{kt-1} D_{kt-1} \right) \right] \tag{4}$$

In accordance with (1), the area wide Divisia aggregate $DMU$ is defined in terms of its growth rates which are the weighted sum of the country-specific Divisia growth rates. The country weights are the two-period averages ($V_{kt}$) of the countries expenditure shares ($S_{kt}$)

$$S_k = \frac{D_k \Pi_k^* h_k}{\sum D_k \Pi_k^* h_k} \tag{5}$$

where we suppressed time-subscripts for notational convenience. Based on (2) and (3), $\Pi_k^*$ denotes the quantity-weighted average of the real user cost and, thus, measures the opportunity cost of holding a unit of $D_k$ in country $k$. Note that the expenditure share $S_{kt}$ depends on a country's price level, the composition of monetary assets and the level of country-specific interest rates. $h_k$ denotes country's $k$ population share. In contrast to e.g. user cost, population shares (like other measures of economic size, including the GDP share) did not change significantly over the last 15 years in the euro area. Therefore, changes in the size of a member country play no important role for the evolution of the euro area Divisia aggregate.

# 3 Data

While Darvas (2015) provides a Divisia aggregate for the monetary union under the assumption of homogeneous countries, there is still no publicly available Divisia aggregate that takes into account the heterogeneity of the euro area. In the following, we compute a euro area wide Divisia aggregate by adopting the heterogeneous country approach of Barnett (2007). The data for the Divisia computation is publicly available from the ECB Statistical Data Warehouse.[2]

## 3.1 Countries under consideration

In the following, we compute a Divisia monetary aggregate for the first 12 countries (EA-12) that adopted the Euro. For these countries all data series are available on a monthly basis from January 2003 onward. The data employed in the current paper end in December 2018 but we plan to provide updates on our website on a monthly basis. The 12 euro area countries under consideration account for more than 95% of the unions population and more than 97% of GDP, compare Table 2.

Barnett and Gaekwad (2018) calculate a Divisia aggregate for a different set of countries including Estonia, Finland, France, Germany, Ireland, Italy, Luxembourg, Malta, the Netherlands, Slovakia, and Slovenia. Note that this group of countries covers a significantly lower share of the euro area, both in terms of population and GDP. A further advantage of using EA-12 countries is that they have adopted the Euro already in 2003. Therefore, the EA-12 index does not require any assumptions about exchange rates.

## 3.2 Monetary assets and transactions data

In the rest of the paper, the focus is on computing M2 Divisia aggregates, i.e. the country-specific and EA-12 wide Divisia aggregate that correspond to the simple sum aggregate

[2]For a full list of the data see Table A.1.

Table 2 The relative size of euro area countries

| Country | Adoption of Euro | Population Share in % (2018) | GDP share in % (2017) |
|---|---|---|---|
| Austria | 1999-01-01 | 2.58 | 3.30 |
| Belgium | 1999-01-01 | 3.34 | 3.92 |
| Finland | 1999-01-01 | 1.61 | 2.00 |
| France | 1999-01-01 | 19.69 | 20.45 |
| Germany | 1999-01-01 | 24.26 | 29.25 |
| Ireland | 1999-01-01 | 1.42 | 2.62 |
| Italy | 1999-01-01 | 17.71 | 15.39 |
| Luxembourg | 1999-01-01 | 0.18 | 0.49 |
| The Netherlands | 1999-01-01 | 5.01 | 6.58 |
| Portugal | 1999-01-01 | 3.01 | 1.74 |
| Spain | 1999-01-01 | 13.66 | 10.41 |
| Greece | 2001-01-01 | 3.14 | 1.61 |
| **EA-12** | | **95.61** | **97.76** |
| Slovenia | 2007-01-01 | 0.61 | 0.38 |
| Cyprus | 2008-01-01 | 0.25 | 0.17 |
| Malta | 2008-01-01 | 0.14 | 0.10 |
| Slovakia | 2009-01-01 | 1.59 | 0.76 |
| Estonia | 2011-01-01 | 0.39 | 0.21 |
| Latvia | 2014-01-01 | 0.57 | 0.24 |
| Lithuania | 2015-01-01 | 0.82 | 0.38 |
| **EA-19** | | **100** | **100** |

Notes: In the euro area, population shares and GDP shares did not change significantly over the past 20 years. The presented numbers refer to 2018 and 2017, respectively.

M2.[3] M2 consists of four types of assets: i) currency in circulation, ii) overnight deposits, iii) deposits with agreed maturity of up to two years and iv) deposits redeemable at notice of up to three month. The computation of a Divisia index requires for each monetary asset country-specific data for its volume and the corresponding interest rate.

For each of the four monetary assets, volumes are published as monetary financial institution (MFI) balance sheet statistics, for which a detailed description can be found in European Central Bank (2012). Note that the ECB also provides estimates for country-specific currency in circulation based on a country's share in the ECB's capital. Deposits might exit or enter

[3]Since M1 considers only two types of assets, the difference between the M1 Divisia and M1 is only small. M3 Divisia is not considered in the current paper due to data availability problems but is the subject of future efforts.

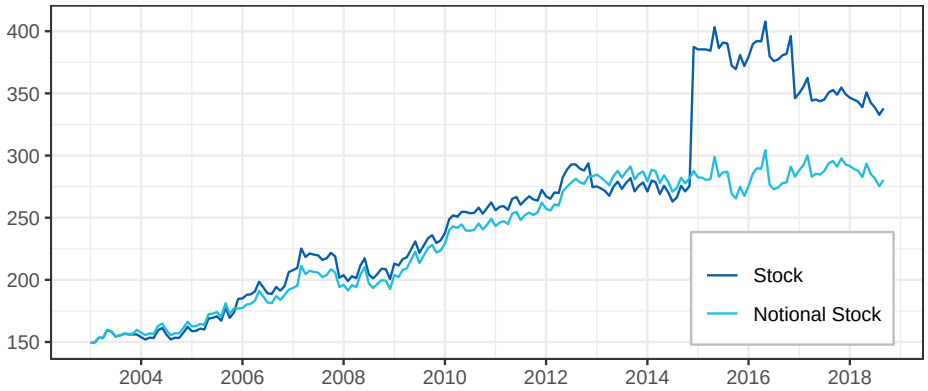
the market. In fact, the level of certain deposits drop to zero at some point in time in some countries. In order to avoid growth rates of minus infinity, we follow Barnett et al. (2013) and calculate growth rates only if deposits are non-zero in two consecutive periods.

The level data provided by the ECB has a further problem that might be less obvious but is still critical for the computation of Divisia aggregates. The problem is that the level data are not adjusted for breaks and shifts due to reclassification or reevaluation of assets. However, simple reclassifications of assets do not represent changes in liquidity and, therefore, should not affect the Divisia aggregate. In the euro area, the shifts in the levels of monetary assets resulting from a simple reclassification of deposits are partly huge. Ignoring this issue of the ECB's level data can lead to spurious shifts in the Divisia aggregate, compare Barnett and Gaekwad (2018). Following Darvas (2015), this problem can be solved using the ECB's transactions data, as defined in the regulation ECB/2013/33: Financial transactions are computed by the ECB as the difference between stock positions at end-of-month reporting dates, from which the effect of changes that arise due to influences other than transactions is removed. For each monetary asset, these transactions can be used to compute the index of notional stock (European Central Bank, 2012). In the following, this index is applied to compute a Divisia aggregate that controls for reclassifications or other breaks unrelated to financial transactions.[4]

The importance of using transaction data for the computation of a Divisia index is illustrated in Figure 1 which shows the unadjusted level and the index of the notional stock of overnight deposits in the Netherlands. In December 2014, the Netherlands introduced a new reporting framework (De Nederlandsche Bank, 2018) which had no effects on transactions and the amount of liquidity. Yet, the reclassification implied a sharp increase in the level of overnight deposits. Note that this spurious realloaction of monetary assets would distort the year-to-year growth rates of the Divisia aggregate for a whole year. Similar level shifts due to reallocations of monetary assets can be seen in Ireland, Spain, Italy and France.

[4]Specifically, the index of the notional stock of a monetary asset $S_i$ in period $t$ is defined as $I_{it} = I_{it-1}(1 + \frac{T_{it}}{S_{it-1}})$ where $T_i$ is the transaction volume of asset $S_i$. The ECB selects a base value of 100, which is not applicable for the Divisia index because the level of a component matters for the calculation of its weights. Following the procedure proposed in European Central Bank (2012), the base value is the level of the corresponding

Figure 1 Stock and index of notional stock for overnight deposits in the Netherlands




Notes: In December 2014, the Netherlands introduced a new reporting framework which led to a large increase in overnight deposits (De Nederlandsche Bank, 2018) (stock) that had no effects on the amount of liquidity. The Figure shows the unadjusted level data (stock) and the shift-adjusted index of notional stock of overnight deposits used in the computation of the Divisia aggregate.

### 3.3 Interest rates

The country-specific own rates of return ($r_i$) for the monetary assets are taken from the MFI interest rate statistics.[5] In accordance with the literature, the interest rate for currency in circulation is assumed to be zero. Since there is no data available for the interest rates on *outstanding* amounts of deposits, we use the interest rates on new business. Missing values are imputed using a linear regression on the overnight deposit rate, see Barnett et al. (2013) and Fisher et al. (1993).

The choice of the benchmark rate ($R$) is less obvious. In theory, the benchmark rate is the rate of return on a pure investment asset that provides no liquidity services on its own and is capital-certain. The assets sole purpose is the transfer of wealth from one period to the next, but such an asset does not exist in reality. User costs of zero would imply the asset to be a free good which is not plausible. In order to ensure that user cost of monetary assets are above zero ($\frac{R-r_i}{1+R} > 0$), the benchmark rate has to be strictly larger than the monetary assets own rates of return. Therefore, a natural candidate for the benchmark rate is the upper envelope of the monetary assets own rates of return plus a liquidity premium. Stracca (2004) includes a risk premium on 60 basis points while the Divisia indices provided by the Fed of St. Louis

monetary asset in the base period January 2003.

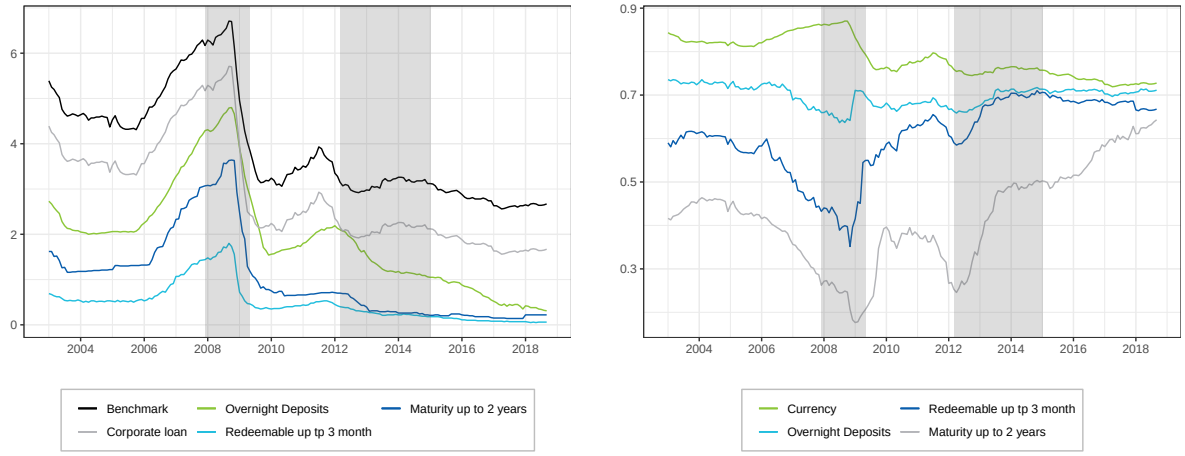
[5]European Central Bank (2017) gives a detailed description of the data and of the methods used to collect it.

use 100 basis points, see Anderson and Jones (2011). Both studies conclude that Divisia growth rates are not sensitive to the magnitude of the liquidity premium.

The upper envelope approach with the liquidity premium is a practical but rather *ad-hoc* solution of the non-negativity problem of the benchmark rate. Therefore, the literature suggests alternative candidates for the benchmark rate which are more closely related to economic theory. In particular, Darvas (2015) approximates the benchmark rate by bank debt with longer maturities than those included in the monetary aggregate. He finds them to be larger than the monetary assets own rates and accepts the downside that long-run bank debts are not risk-free. Barnett et al. (2013), following a suggestion from Offenbacher and Shemesh (2011), decide to stay in the risk-neutral setting and include the low risk corporate loan rate in the calculation of the upper envelope. This is because banks would not pay out a higher interest rate on short-term deposits than they earn with short-term loans. Barnett et al. (2013) only resolve to the upper envelope approach with liquidity premium of 100 basis point in periods where the corporate loan rate is not available.

In order to define an appropriate benchmark rate for the euro area, we follow Barnett and Gaekwad (2018) and consider the interest rate on loans up to one year maturity as the corporate loan rate. However, in contrast to the United States (Barnett et al., 2013) and Israel (Offenbacher and Shemesh, 2011), corporate loan rates in the euro area do not always exceed the monetary assets own rates. Thus, a liquidity premium of 100 basis points is added to the upper envelope of the own rates and the loan rate to ensure positive user costs. In order to illustrate our approach for defining the benchmark rate, Figure 2 displays the interest rates and the implied user cost for Finland. In normal times, the upper envelope of the interest rates is the corporate loan rate implying that the benchmark rate is the loan rate plus 100 basis points. For several months in 2009, however, the corporate loan rate was below the rate for longer-term deposits. In this period, the longer-term deposit rate is the upper envelope and, thus, the Finnish benchmark rate is computed as the longer-term deposit rate plus 100 basis points. While Barnett and Gaekwad (2018) add the liquidity premium only for those periods where the loan rate does not exceed the own rates, we find it more plausible to add the liquidity premium in each period.

Figure 2 Benchmark Rate and User Cost for Finland




(a) Benchmark Rate ($R_t$) (b) User cost ($p_t$)

Notes: The user costs are calculated according to Equation 3. The benchmark rate is defined as the upper envelope of the monetary assets own interest rates and the interest rate on loans up to one year maturity plus a liquidity premium of 100 basis points. The shaded areas indicate recession periods.

# 4 Divisia Monetary Aggregates at the Country Level

Divisia aggregates depend on both, interest rates and the composition of monetary assets. Before we further aggregate to the EA-12 Divisia index, this section investigates the behavior of the various components of the M2 Divisia aggregate at the country level. The aim of the analysis is twofold. On the one hand, we explore when and why one should expect economically relevant differences between the behavior of Divisia and simple sum monetary aggregates *within* a country. On the other hand, we are interested in the main drivers of heterogeneity in monetary developments, i.e. when and why the behavior of country-specific Divisia aggregates differs *across* EA-12 countries.

## 4.1 User cost

If user cost were always identical for all monetary assets, growth rates of Divisia and corresponding simple sum aggregates would be also identical. In this case, there would be no additional information content of Divisia aggregates. Therefore, it is worth emphasizing that user cost of different monetary assets differ significantly within and across EA-12

countries, compare Figure A.1 in the Appendix.

The Finnish data displayed in Figure 2 can be used to illustrate when and why user cost may change within EA-12 countries. Changes in user cost require that the own rates of monetary assets and the benchmark rate grow at different rates. Figure 2 shows that the user costs for overnight deposits and currency in circulation have been fairly stable over the past 15 years. By contrast, the user cost for the two types of longer-term deposits included in M2 display remarkable down- and upswings around the two recession periods (marked by the shaded areas). The drop in the user cost of longer-term deposits in the run-up to the great recession can be observed for all EA-12 countries. Interestingly, the second drop, probably related to the European debt and banking crisis, is particularly pronounced in Greece and Ireland.

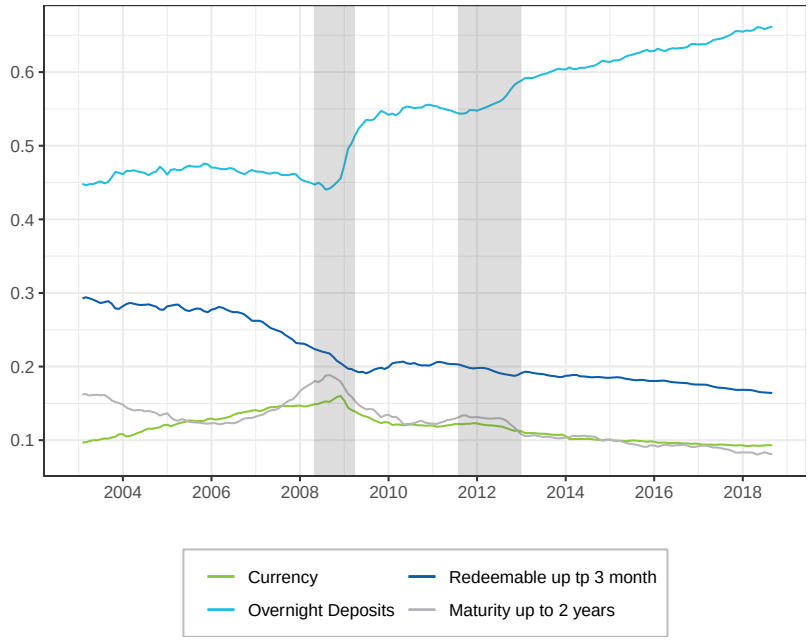
### 4.2 Expenditure shares

The weight of a monetary asset used in the computation of the Divisia aggregate depends on its expenditure share and thus on both, the user cost and the volume of the monetary asset. The expenditure shares differ significantly across the EA-12 countries, compare Figure A.2 in the Appendix. The large and persistent differences in the level and the dynamics of expenditure shares strongly suggest that a euro area Divisia aggregate should not be based on the assumption of homogeneous member countries.

In spite of the significant heterogeneity across EA-12 countries, there are a few stylized facts about the size and evolution of expenditure shares that are worth noting. First, the expenditure share of currency in circulation is small (around 10%) and rather stable over time for most of the EA-12 countries. The major exception is Greece where the currency weight has steadily increased since the outbreak of the financial crisis to more than 20%. Second, for most of the EA-12 countries, overnight deposits take the highest expenditure share across monetary assets. The exception is now Belgium where the weights of three-month deposits are particularly high. For most countries, however, the weight of overnight deposits range between 50% (France) and 70% (Italy). Third, the expenditure share of overnight deposits is

typically upward trending, particularly since the outbreak of the financial crisis, see e.g. Figure 3 for the expenditures shares in Germany. The German example further illustrates the fourth stylized fact, namely that major shifts in expenditure shares are related to recession periods.

Figure 3 Expenditure Shares in Germany



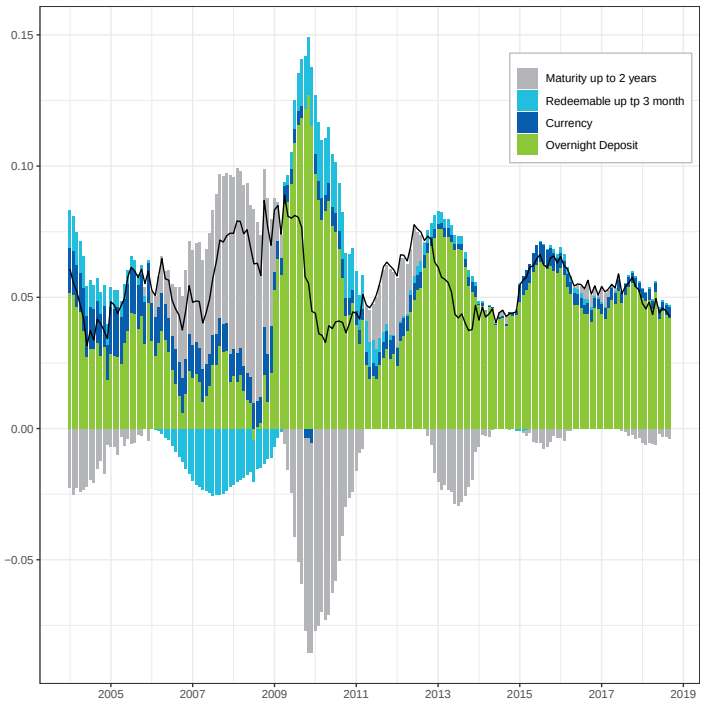

Notes: The weight of a monetary asset used in the computation of the Divisia aggregate depends on its expenditure share, compare Equation (2) in Section 2.2. The Figure shows the expenditure shares of the monetary assets included in the German M2 Divisia aggregate. Shaded areas indicate recessions.

For all EA-12 countries, the nearly constant expenditure share of currency implies an inverse relationship between the expenditure share of overnight deposits and the weight of the two remaining types of longer-term deposits, i.e. three-month deposits and deposits with a maturity up to two years. The relative importance of both types of longer-term deposits varies remarkably across EA-12 countries. In some countries, including e.g. Germany and Spain, the expenditure share of three-month deposits is large but decreasing. In other countries, including Austria, Greece and Portugal, three months deposits play no role such that their weight in the Divisia aggregate is virtually zero.

### 4.3 Monetary components and Divisia growth

The analysis of expenditure shares provided insights about the *relative* importance of monetary components for the Divisia aggregate. Expenditure shares, however, cannot reveal the *absolute* importance of a monetary asset, i.e. to what extent an observed change in the Divisia aggregate can be attributed to the underlying monetary components. To that aim, we adopt the approach of the CFS who regularly decomposes the contributions of the monetary components to the growth rate of the U.S. Divisia index.

Figure 4 Components contribution in Germany




Notes: The Figure shows the annual growth rate of the German M2 Divisia aggregate and how the four types of monetary assets contribute to it.

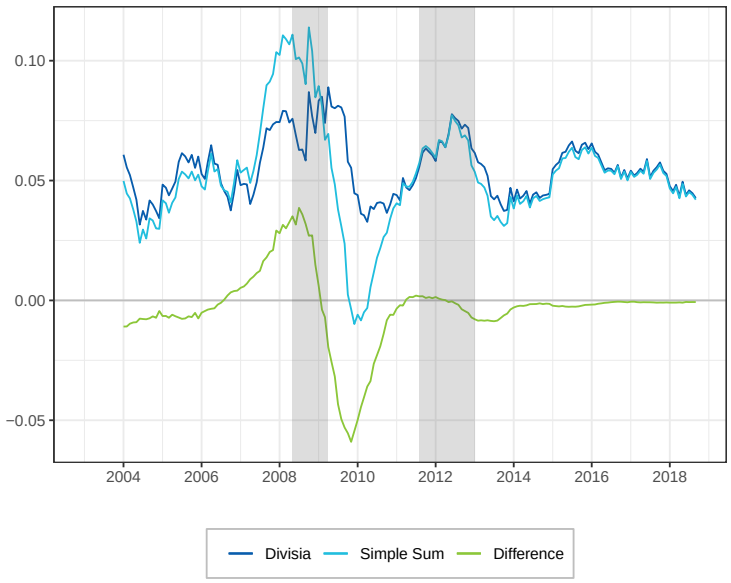
We calculated the contributions of the four M2-related monetary assets to the growth of the Divisia aggregate for all EA-12 countries, see Figure A.3. In order to illustrate the usefulness of this tool, Figure 4 shows how the various monetary assets contributed to the annual growth rates of the German Divisia aggregate. Note that the conclusion that can be drawn from this analysis shed further light on the stylized facts derived for the expenditure shares. Figure 4 shows that i) the contribution of currency to the growth of the Divisa index is small

and stable. ii) Typically, the contribution of overnight deposits is by far the largest. iii) The dominant role of overnight deposits for the growth rate of the Divisia index is particular pronounced after the financial crisis. iv) During recessions, positive growth rates of overnight deposits are partly compensated by negative growth rates of longer-term deposits.

### 4.4 The divergence between simple sum and Divisia aggregates

Let us now compare the country-specific Divisia aggregate with its simple sum counterpart. For each of the EA-12 countries, both monetary aggregates are shown in the Figure A.4.

Figure 5 Growth Rates of Divisia and simple sum monetary aggregate in Germany




Notes: The Figure shows the annual growth rates of the German M2 Divisia aggregate, its simple sum counterpart and the difference between the two growth rates. Shaded areas indicate recessions.

Figure 5 shows the year-to-year growth rates of German M2, the M2 Divisia aggregate, and their *divergence* defined as the difference between the two growth rates. Similar to the other EA-12 countries, the growth rates of German M2 and its Divisia counterpart were very similar before 2007. In fact, M2 and the related Divisia aggregate conveyed broadly the same information about the liquidity situation in the economy in the rather calm pre-crisis period. However, Divisia and simple sum aggregates tend to grow very differently in more turbulent times. According to Figure 4, the non-zero divergence around recessions

can be explained by a reallocation of monetary assets from short- to longer-term deposits and *vice versa*. In line with Barnett and Chauvet (2011), the crisis-induced substitution from less liquid to more liquid monetary assets suggests that the difference between Divisia and simple sum growth rates could have a predictive content for recessions.

## 5 The Divisia Monetary Aggregate for the Euro Area

Let us now use the Divisia aggregates computed at the country level to compute the EA-12 Divisia monetary aggregate. The Divisia EA-12 aggregate is the weighted sum of the country-specific Divisia aggregates, compare (5). The weight of a country can be interpreted as its expenditure share.

Table 3 Country Weights in the euro area Divisia Index

| | Country | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AT | BE | DE | ES | FI | FR | GR | IE | IT | LU | NL | PT |
| Weight | 2.6 | 3.4 | 25.1 | 15.8 | 1.6 | 21.1 | 3.4 | 1.3 | 17.7 | 0.2 | 5.1 | 2.7 |

Notes: The Table shows the average expenditure shares (in %) used as weights in the euro area wide M2 Divisia aggregate for each of the EA-12 countries, including Austria (AT), Belgium (BE), Finland (FI), France (FR), Germany (DE), Greece (GR), Ireland (IE), Italy (IT), Luxembourg (LU), the Netherlands (NL), Portugal (PT), Spain (ES). For more details on the derivation of expenditure share, see Equation (5) in Section 2.3.

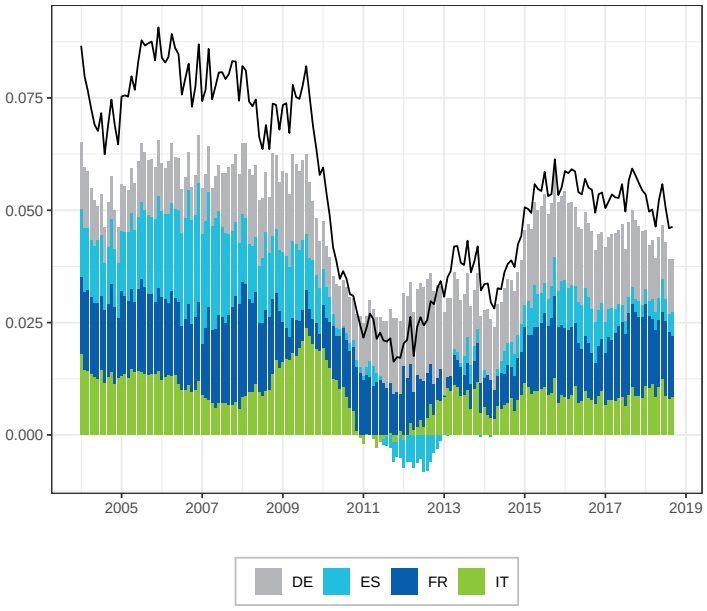
Table 3 shows that the average expenditure shares of the EA-12 countries are very close to the corresponding shares in population or GDP, compare Table 2. As a consequence of the weighting scheme, euro area wide monetary aggregates will hardly respond to monetary developments in small countries like Greece. In the same vein, the very small weights of the new member countries (see Table 2) imply that monetary aggregates derived for the group of EA-12 countries should be very close to the full EA-19 measure. Yet, the European experience in the aftermath of the great recession and the following debt crisis clearly demonstrated that developments in small countries could be extremely important, even if their share in euro area wide aggregates seems to be negligible.

The four largest countries (France, Germany, Italy, and Spain) account for more than 80% of the monetary unions total expenditure for monetary assets. The pre-dominant role of

the big countries for the monetary developments of the whole euro area is reflected in their dominant impact on the growth rates of the EA-12 Divisia aggregate. Figure 6 displays the annual growth rate of the Divisia EA-12 aggregate together with the growth contributions of France, Germany, Italy, and Spain. Apparently, the dynamics of the Divisia EA-12 aggregate can be attributed mostly to developments in these four countries. The contributions are mostly positive indicating that the amount of liquidity has typically increased. The notable exception is Spain where liquidity decreased in 2012, probably as a result of the European debt crisis. Note that the contributions of the four countries to the EA-12 Divisia aggregate have been very similar before the financial crisis. Since then, however, the monetary developments in Germany became more important for the EA-12 Divisia aggregate while the contribution of Spain has declined.

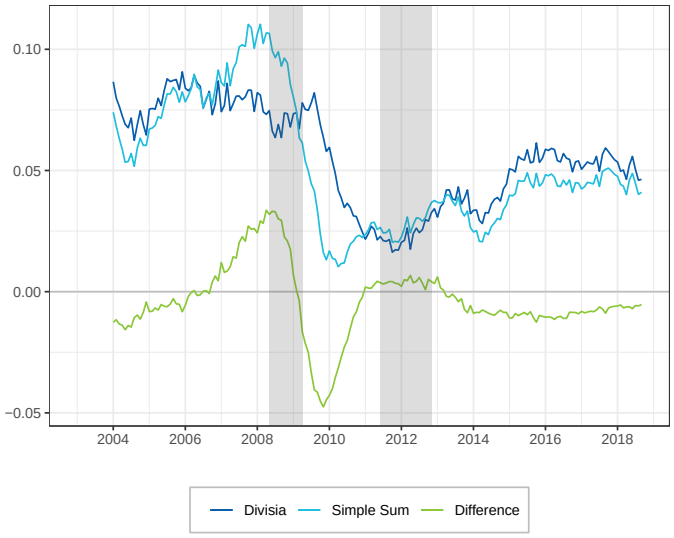
Let us now compare the EA-12 M2 Divisia aggregate with its simple sum counterpart. In accordance with the monetary developments in bulk of the EA-12 countries, Figure 7 shows that the growth rates of the simple sum and the Divisia aggregate differ particularly around the great recession. In line with the analysis of individual countries, the simple sum

Figure 6 Annual Contribution of the Largest Countries




Notes: The Figure shows the annual growth rate of the euro area M2 Divisia aggregate and how the four largest member countries Germany (DE), Spain (ES), France (FR), and Italy (IT) contribute to it.

Figure 7 Divisia and simple sum aggregates for the EA 12




Notes: The Figure shows the annual growth rates of the euro area M2 Divisia aggregate, its simple sum counterpart and the difference between the two growth rates. The shaded areas indicate recession periods.

aggregate of the EA-12 area overestimates the change in liquidity services in the run-up to the crisis when monetary assets shifted from overnight to longer-term deposits. By contrast, the amount of liquidity is clearly underestimated by the simple sum aggregate from about 2009 until 2011 when these sifts in money holdings were reversed. The recession in the euro area around 2012 was much weaker than the great recession, particularly for the big countries. This may explain why the difference between the growth rates of euro area simple sum and Divisia aggregates is less pronounced in that period.

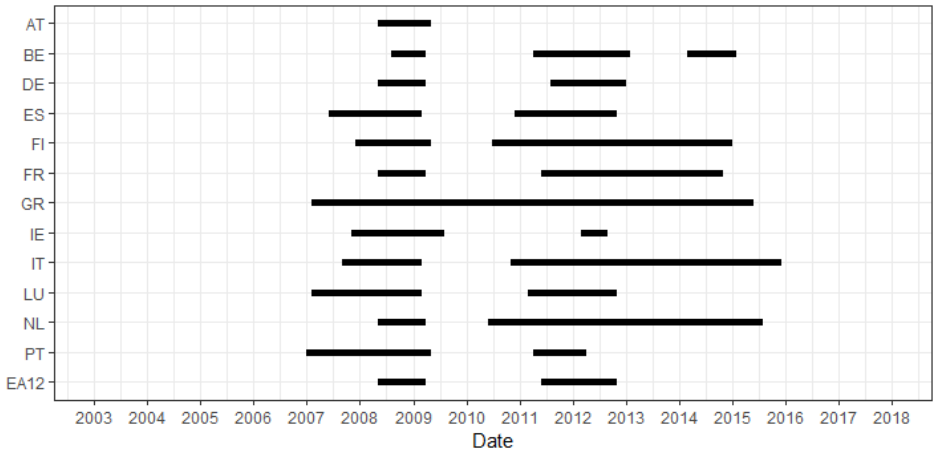
## 6 Divisia Aggregates and Recessions in Euro Area Countries

In accordance with the observation of Barnett and Chauvet (2011) for the U.S., our analysis suggested that the divergence between the Divisia aggregate and its simple sum counterpart could be a useful predictor of recessions for the EA-12 countries. In this section, we aim to investigate the predictive content of the divergence for recessions more closely.

The CEPR euro area Business Cycle Dating Committee publishes only a common Euro-

pean economic cycle. While there might be a convergence of business cycles in the long-run, recession periods in the EA-12 countries might not fully coincide in our sample period. Following e.g. Artis et al. (1997), we define a country-specific recession indicator based on the country's index of industrial production provided by Eurostat.[6] Figure 8 confirms that the timing and the length of recession periods differ significantly between EA-12 countries, particularly in the aftermath of the financial crisis.

Figure 8 Recessions in the EA-12 countries




Notes: The Figure shows for each EA-12 country the monthly recession indicator based on the country's index of industrial production. For further explanation, see e.g. Artis et al. (1997) and Footnote 6.

In the tradition of Estrella and Mishkin (1998), we use a probit model to estimate the predictive power of the Divisia aggregates with respect to future recessions. Following e.g. Borio et al. (2018), we employ a pooled panel probit model in order to exploit the panel dimension of our data set. The variable being predicted is the country-specific recession indicator $Y_{i,t}$ that equals one if country $i$ is in a recession in period $t$ and zero otherwise. The model is defined in reference to a theoretical linear relationship of the form

$$y^*_{i,t} = \beta x_{i,t-h} + \varepsilon_{i,t} \tag{6}$$

where the unobservable $y^*$ determines the occurrence of a recession, $h$ is the length of the

[6]Specifically, we define recession periods using a 7-month moving average of industrial production while peaks and troughs of the business cycles are identified as the absolute highest or lowest points within 24 months. Note that our results are robust with respect to alternative methods of defining recession dates.

forecast horizon, $\varepsilon$ is a normally distributed error term, $\beta$ is a vector of coefficients, and $x$ is a set of predictors, including a constant. The observable recession indicator $Y_{i,t}$ is assumed to be one if $y^*_{i,t} > 0$ and zero otherwise. In a probit model, the estimated equation is

$$Pr(Y_{i,t} = 1) = \Phi(\beta x_{i,t-h} + \varepsilon_{i,t})$$

where $\Phi$ denotes the cumulative normal distribution function.

To begin with, we estimate a benchmark probit model that ignores monetary aggregates and only includes information from long- and short-term interest rates, i.e. for each EA-12 country the 10 year government bond rate ($R^L_{i,t-h}$) and the three-month money market rate ($R^S_{i,t-h}$) provided by the OECD. Recently, the well-established predictive content of the spread between long- and short-term interest rates ($SP = R^L - R^S$) has been reconfirmed by Goodhart et al. (2019) for the UK. Following Goodhart et al. (2019), the benchmark model additionally controls for the level of the long term interest rate. The upper part of Table 4 summarizes the main results obtained for the benchmark model. In accordance with the empirical literature, the presented t-statistics show that the predictive content of the spread for recessions is significant and plausibly signed for all forecasting horizons.

Let us now test whether there is an additional predictive content of Divisia monetary aggregates. To that aim, the set of predictors ($x_{i,t-h}$) of the probit model is augmented by the divergence between the growth rates of the M2 Divisia aggregate and its simple sum counterpart ($DIV_{i,t-h}$). The results strongly suggest that Divisia monetary aggregates contain useful information for the prediction of recessions, see the lower part of Table 4. Particularly for shorter-term forecast horizons up to 9 months, the impact of the Divisia divergence is statistically significant and also plausibly signed. According to the pseudo $R^2$s, the predictive content of the divergence variable for recessions might not only be statistically but also economically relevant.

Table 4 Predicting recessions in the euro area: Results from a panel probit analysis

$Pr(Y_{i,t} = 1) = \Phi(\alpha_{0,h} + \alpha_{1,h} SP_{i,t-h} + \alpha_{2,h} R^L_{i,t-h})$

| | h = month ahead | | | | |
|---|---|---|---|---|---|
| | 3 | 6 | 9 | 12 | 18 |
| pseudo $R^2$ | 0.096 | 0.106 | 0.099 | 0.085 | 0.054 |
| $t-$stat $\alpha_1$ | $-6.78^{***}$ | $-6.75^{***}$ | $-5.14^{***}$ | $-3.60^{***}$ | $-1.77^{*}$ |
| $t-$stat $\alpha_2$ | $12.01^{***}$ | $10.84^{***}$ | $8.70^{***}$ | $6.43^{***}$ | $3.67^{***}$ |

$Pr(Y_{i,t} = 1) = \Phi(\alpha_{0,h} + \alpha_{1,h} SP_{i,t-h} + \alpha_{2,h} R^L_{i,t-h} + \gamma_h DIV_{i,t-h})$

| | h = month ahead | | | | |
|---|---|---|---|---|---|
| | 3 | 6 | 9 | 12 | 18 |
| pseudo $R^2$ | 0.151 | 0.140 | 0.114 | 0.088 | 0.055 |
| $t-$stat $\alpha_1$ | $-0.28$ | $-2.33^{**}$ | $-2.81^{***}$ | $-2.63^{***}$ | $-2.02^{**}$ |
| $t-$stat $\alpha_2$ | $5.64^{***}$ | $6.31^{***}$ | $6.10^{***}$ | $5.27^{***}$ | $3.81^{***}$ |
| $t-$stat $\gamma$ | $7.79^{***}$ | $4.73^{***}$ | $3.06^{***}$ | $1.70^{*}$ | $-0.85$ |

Notes: The Table shows measures of fit and t-statistics for pooled panel probit models with and without the divergence between the growth rate of M2 Divisia and its simple sum counterpart. $SP$ denotes the spread between the long- and short-term interest rate, $R^L$ is the 10 year government bond rate and $DIV$ is the divergence between the growth rate of two monetary aggregates. t-statistics are based on Newey-West standard errors with a autocorrelation length of $h - 1$. $^{***}/^{**}/^{*}$ indicate significance at the 1-/5-/10- percent level.

## 7 Conclusions

This paper introduced a new Divisia monetary aggregate for the EA-12 countries. Advancing on earlier contributions, the new Divisia data takes into account the heterogeneity of the euro area. We show that user cost and the composition of monetary assets have differed remarkably across euro area countries, particularly since the run-up to the financial crisis. Our findings confirm the important role of country-specific data for the analysis of monetary developments in the euro area. A panel probit analysis demonstrates the usefulness of country-specific Divisia aggregates for predicting recessions in the EA-12 countries.

# Appendix

Table A.1 Datasource

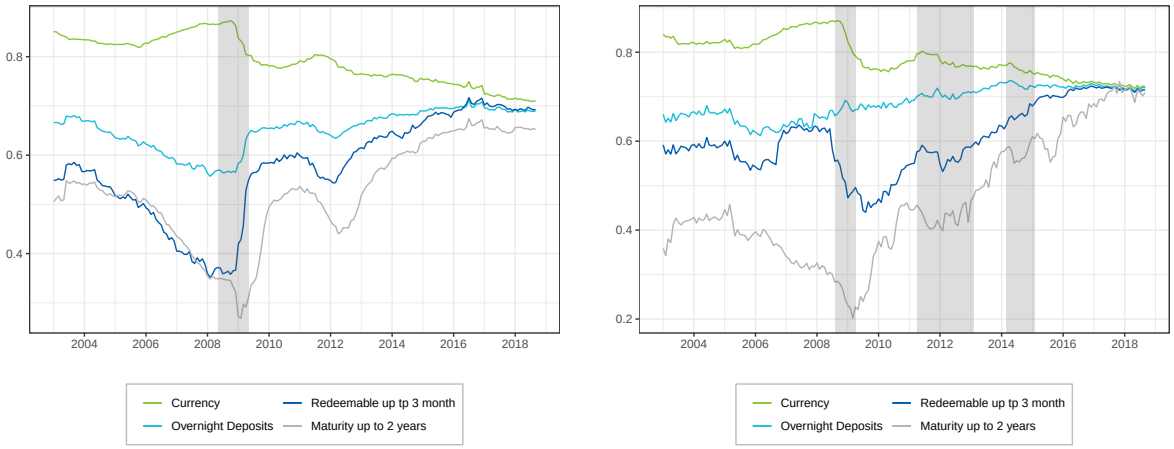
| Component | Type | Key | Description/Notes |
|---|---|---|---|
| Currency in Circulation | Level | BSI.M."CID".N.N.L10.X.1.Z5.0000.Z01.E<br>BSI.M.U2.N.C.L10.X.1.Z5.0000.Z01.E | |
| | Rate | N.A. | 0% |
| | Transaction | BSI.M.U2.N.C.L10.X.4.Z5.0000.Z01 | only available for U2 |
| Overnight Deposit | Level | BSI.M."CID".N.A.L21.A.1.U2.2300.Z01.E | |
| | Rate | MIR.M.CID.B.L21.A.R.A.2230.EUR.N | Annualised agreed rate, new buisness coverage |
| | Transaction | BSI.M."CID".N.A.L21.A.4.U2.2300.Z01.E | |
| Deposits with agreed maturities of up to 2 years | Level | BSI.M."CID".N.A.L22.L.1.U2.2300.Z01.E | |
| | Rate | MIR.M."CID".B.L22.L.R.A.2230.EUR.O | Annualised agreed rate, outstanding amount business coverage |
| | Transaction | BSI.M."CID".N.A.L22.L.4.U2.2300.Z01.E | |
| Deposits redeemable at notice of up to 3 month | Level | BSI.M."CID".N.A.L23.D.1.U2.2300.Z01.E | |
| | Rate | MIR.M."CID".B.L23.D.R.A.2250.EUR.N | Annualised agreed rate, new buisness coverage |
| | Transaction | BSI.M."CID".N.A.L23.D.4.U2.2300.Z01.E | |
| Benchmark | Rate | MIR.M."CID".B.A20.F.R.A.2240.EUR.O | - not available for Belgium for the entire period<br>- Bank interest rates, loans to corporations with an original maturity of up to one year (outstanding amounts) |
| | | MIR.M.BE.B.A2I.AM.R.A.2240.EUR.N | Cost of borrowing for corporations - Belgium |
| Population | Level | ENA.A.N."CID".W0.S1.S1._ Z.POP._ Z._ Z._-Z.PS._ Z.N | |

*Notes:*All Data were taken from the ECB Statistical Data Warehouse. Country ID ("CID"): Austria (AT), Belgium (BE), Finland (FI), France (FR), Germany (DE), Greece (GR), Ireland (IE), Italy (IT), Luxembourg (LU), Netherlands (NL), Portugal (PT), Spain (ES), Euro Area (changing composition) (U2).

N.A. (not available)

Figure A.1 The user costs for each EA-12 country

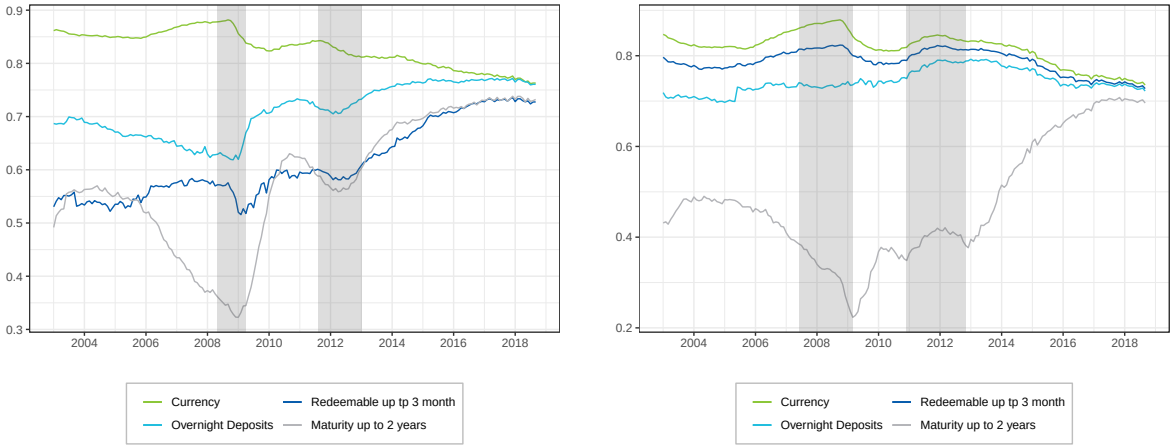
(a) Austria




(b) Belgium



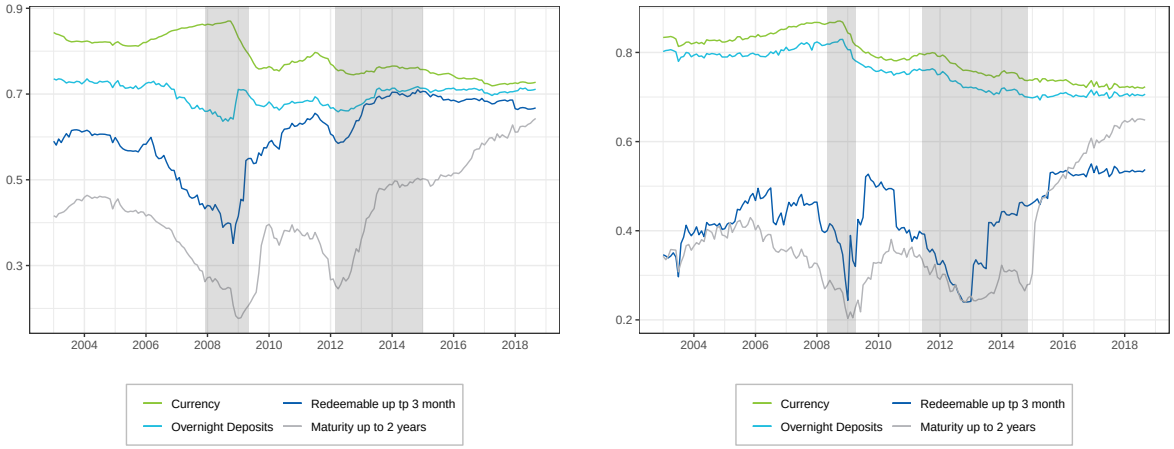

(c) Germany




(d) Spain



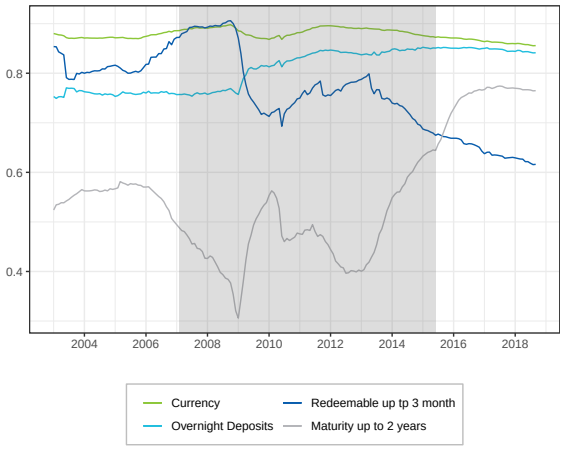

(e) Finland




(f) France

Figure A.1 The user cost for each EA-12 country

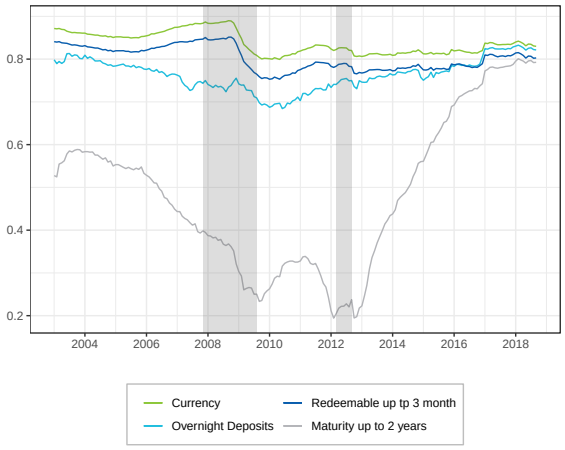
(g) Greece







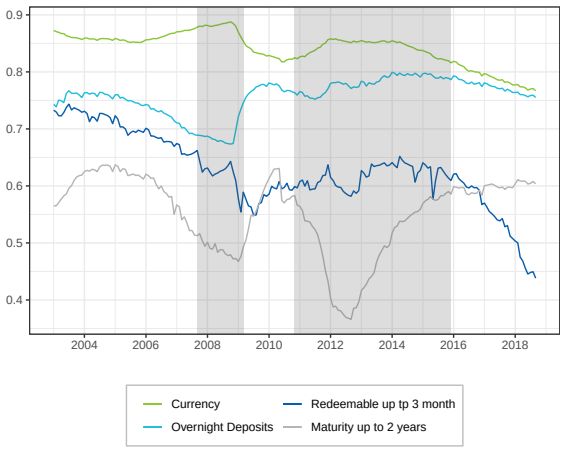
(h) Ireland







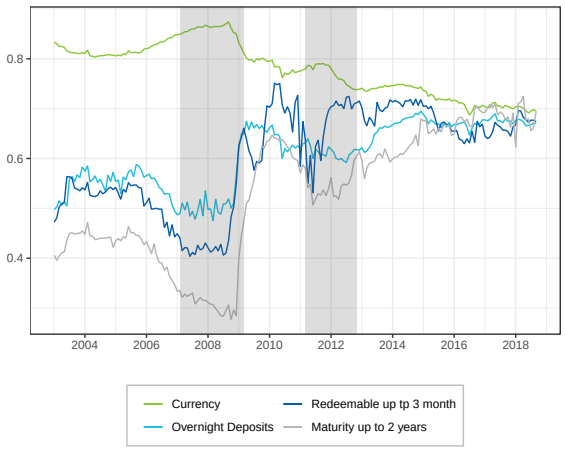
(i) Italy







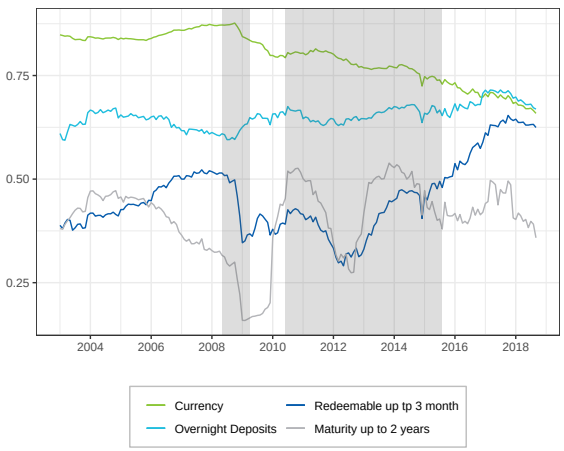
(j) Luxembourg







(k) The Netherlands







(l) Portugal

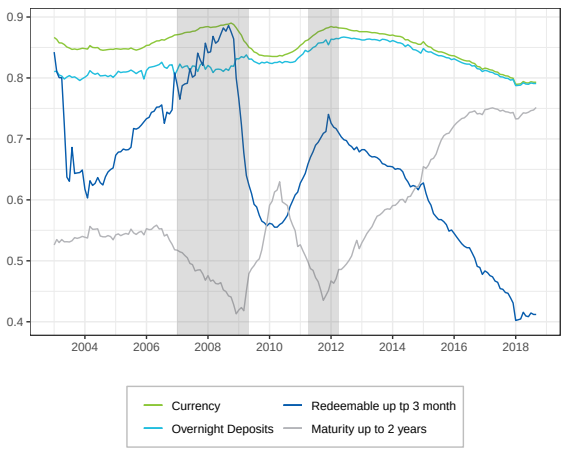
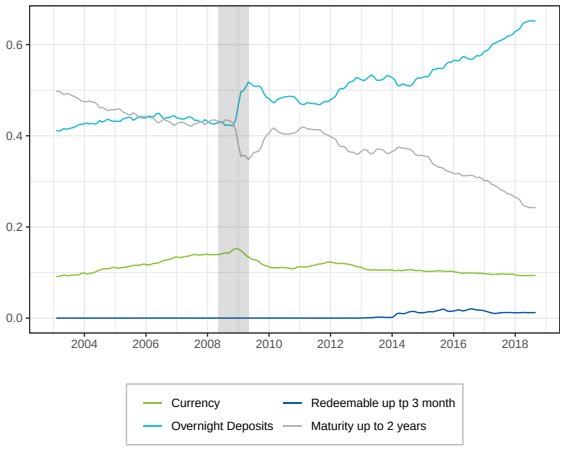


Figure A.2 The expenditure shares for each EA-12 country

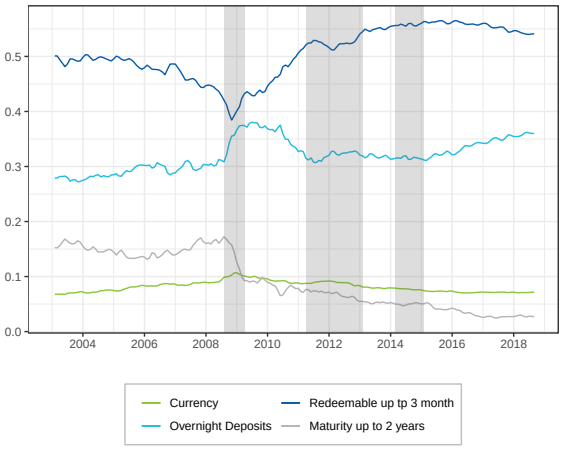
(a) Austria







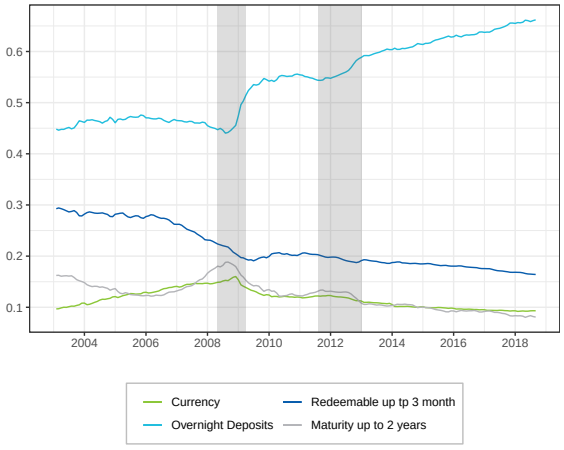
(b) Belgium







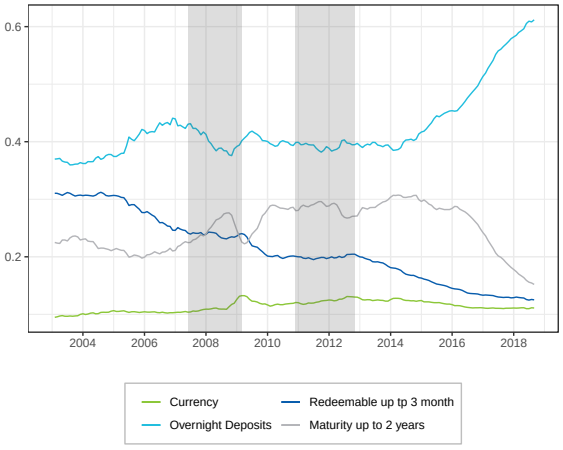
(c) Germany







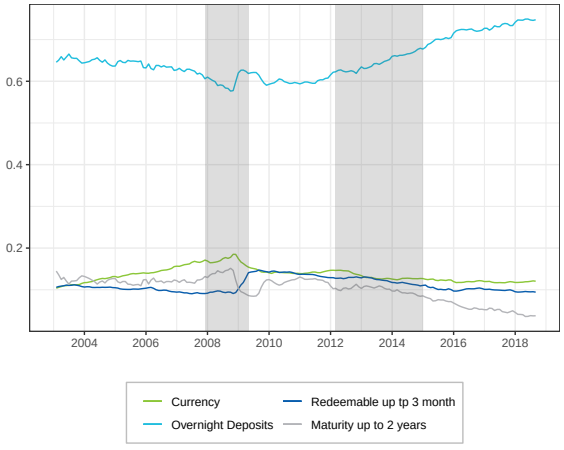
(d) Spain







(e) Finland







(f) France

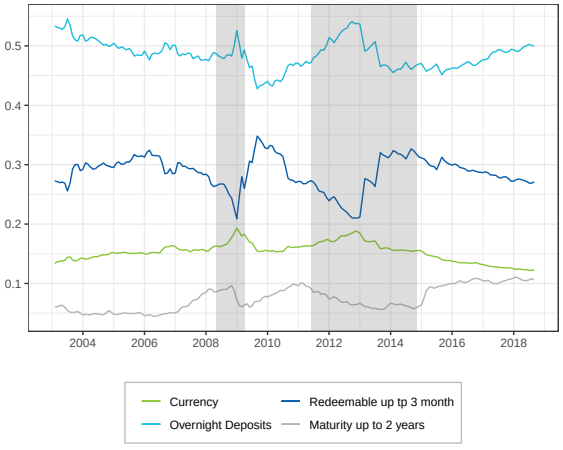
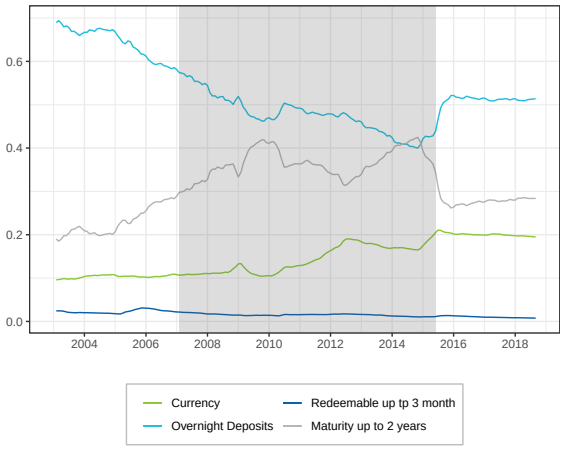


Figure A.2 The expenditure shares for each EA-12 country

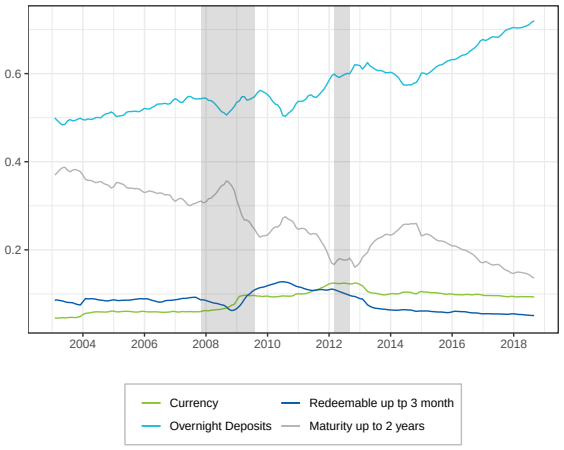
(g) Greece







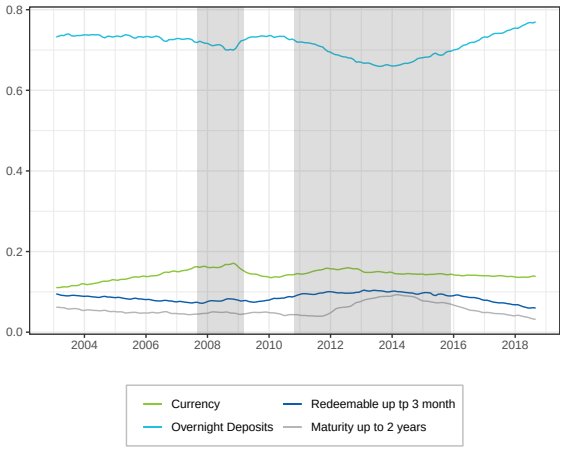
(h) Ireland







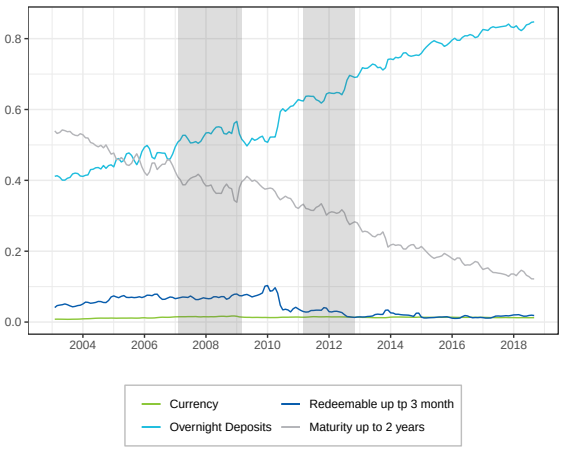
(i) Italy







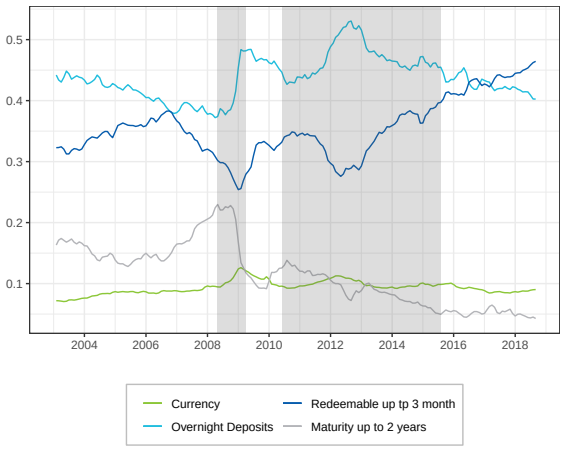
(j) Luxembourg







(k) The Netherlands







(l) Portugal

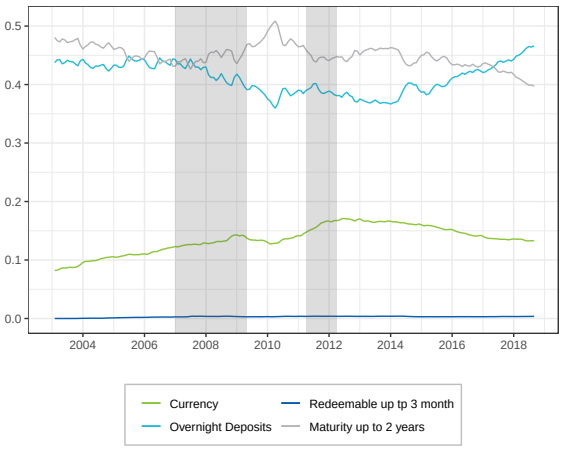
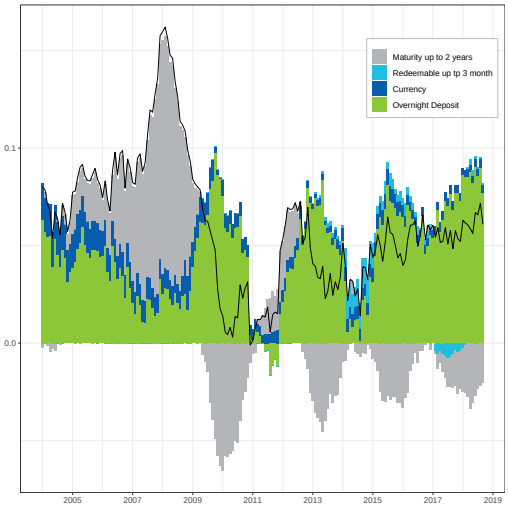
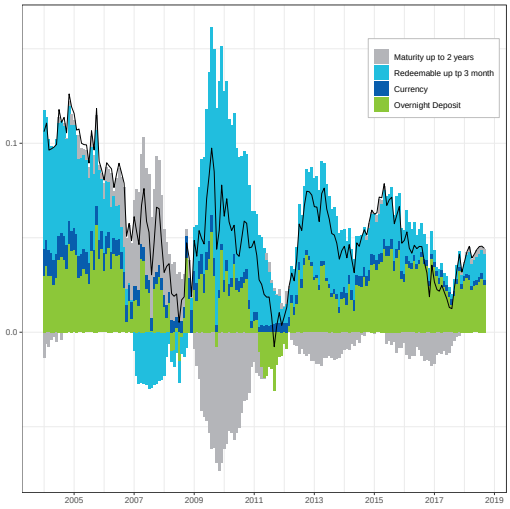


Figure A.3 The contributions of the monetary assets to the Divisia aggregate growth for each EA-12 country

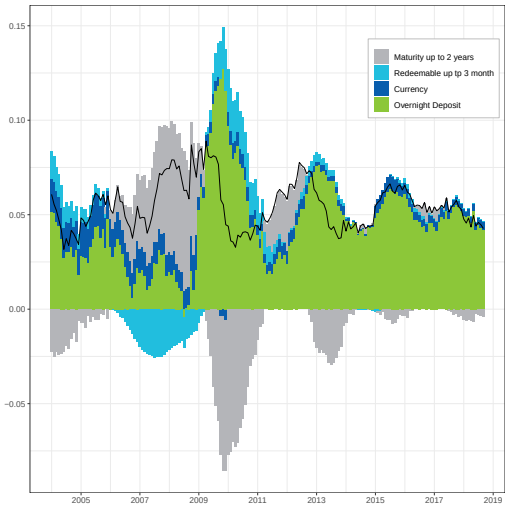
(a) Austria



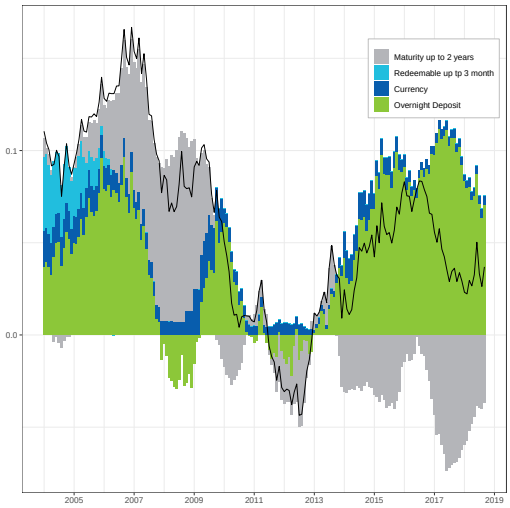

(b) Belgium



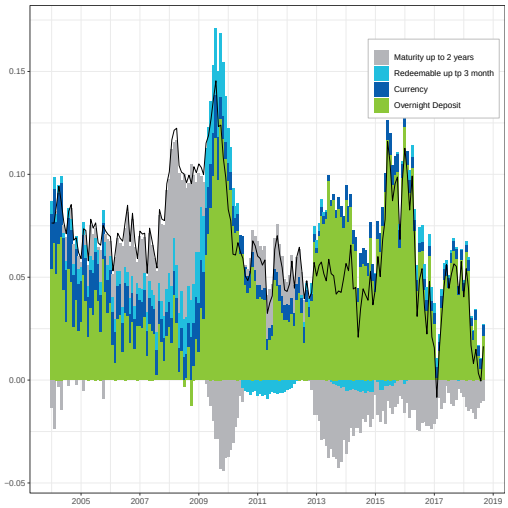

(c) Germany




(d) Spain




(e) Finland




(f) France

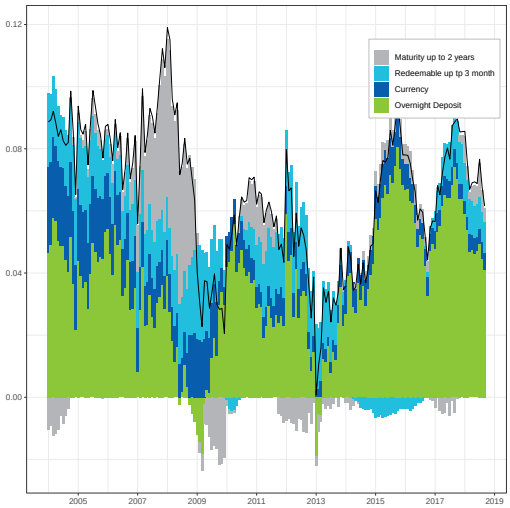
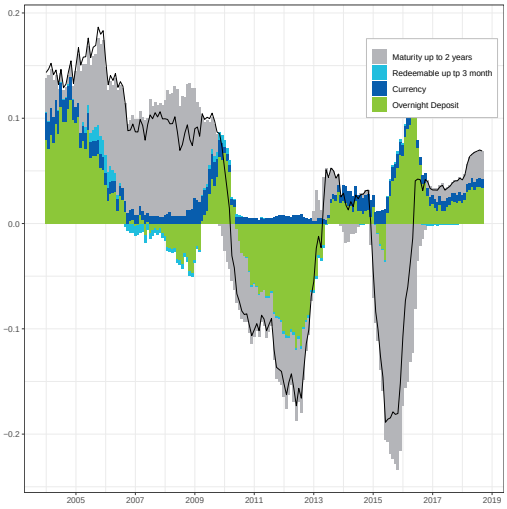
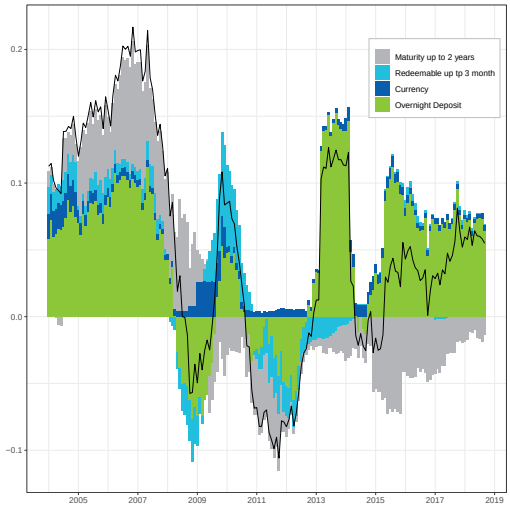


Figure A.3 The contributions of the monetary assets to the Divisia aggregate growth for each EA-12 country

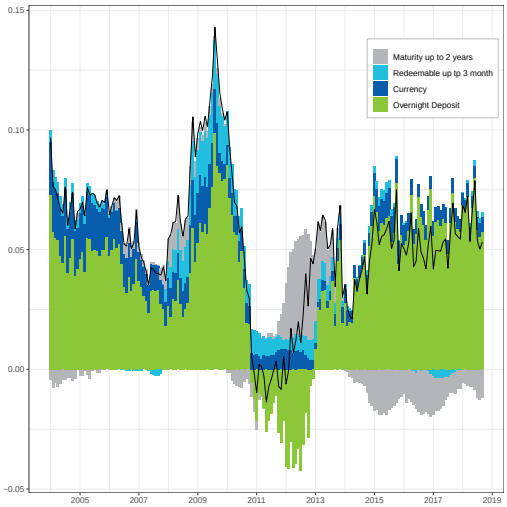
(g) Greece



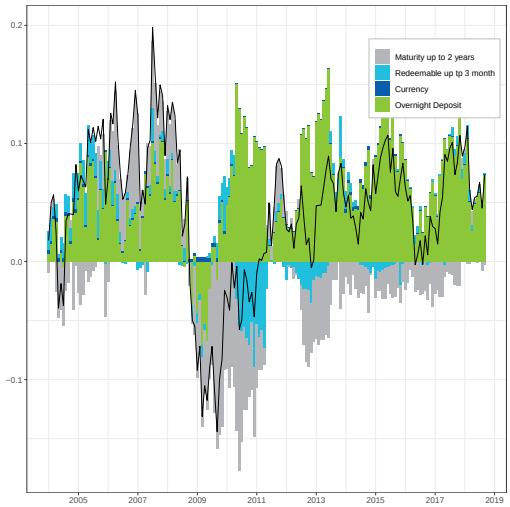

(h) Ireland



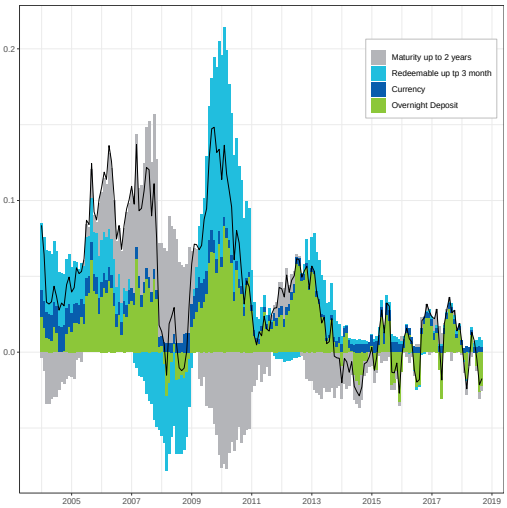

(i) Italy




(j) Luxembourg




(k) The Netherlands




(l) Portugal

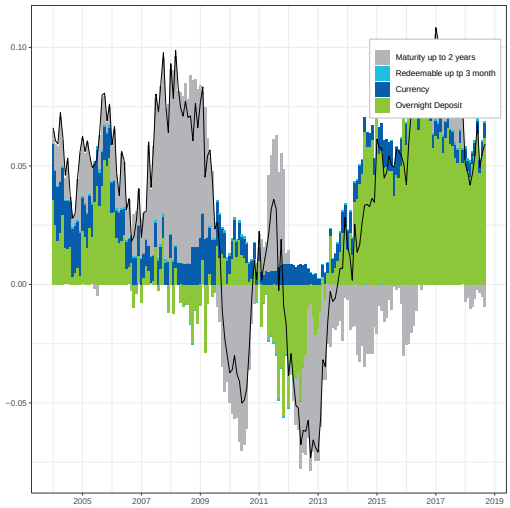
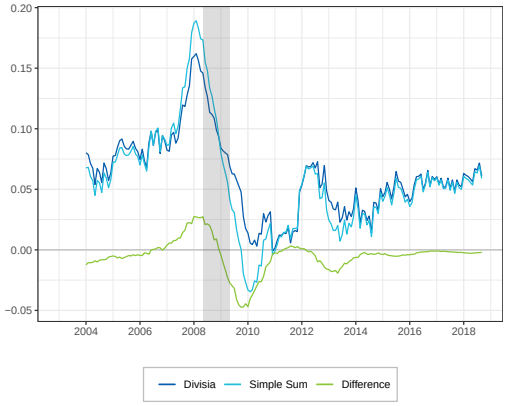
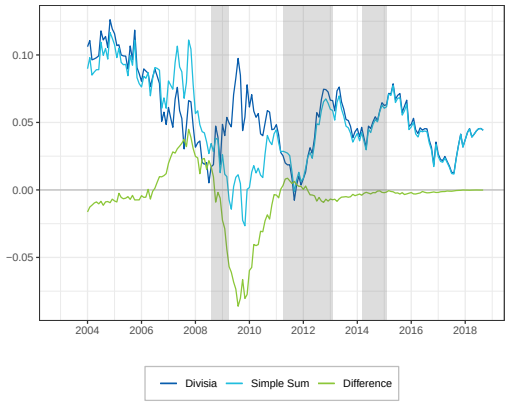


Figure A.4 The growth rates of Divisia and simple sum monetary aggregate for each EA-12 country

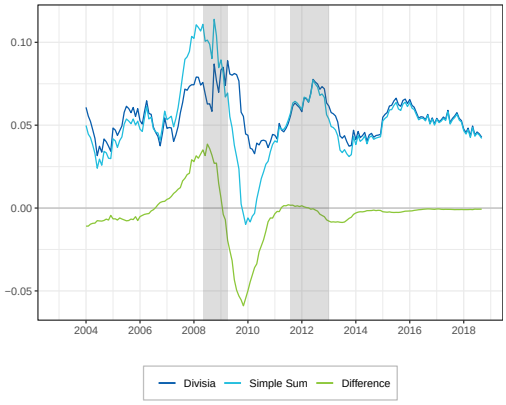
(a) Austria



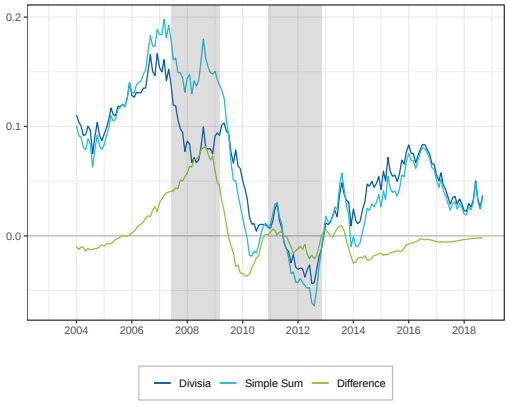

(b) Belgium




(c) Germany




(d) Spain




(e) Finland


(f) France

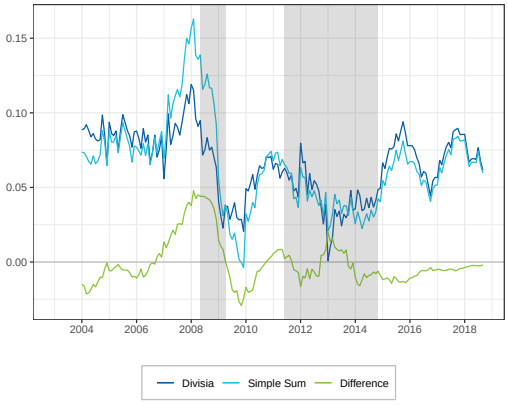
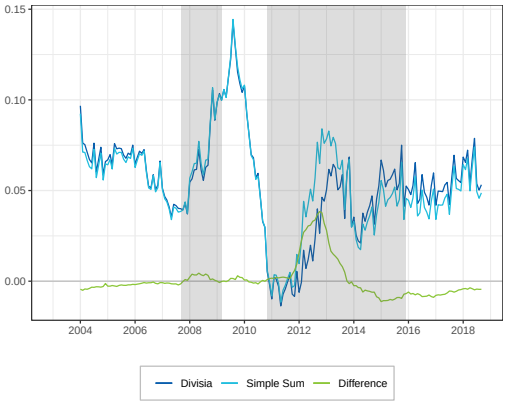


Figure A.4 EA-12 Countries growth

(g) Greece

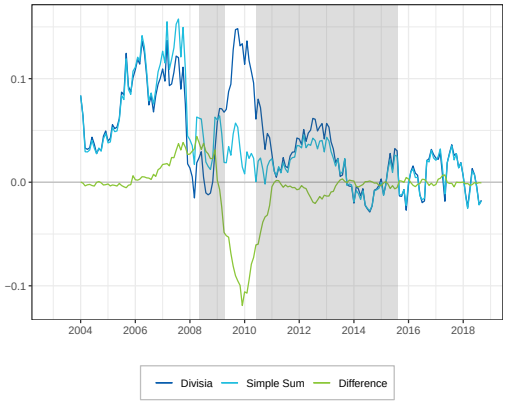

(h) Ireland


(i) Italy

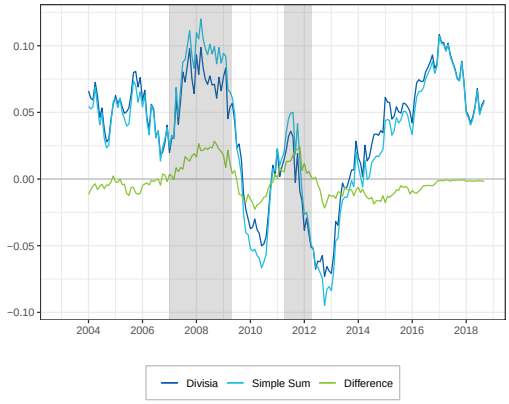



(j) Luxembourg


(k) The Netherlands




(l) Portugal



## References


Anderson, R. G. and Jones, E. B. (2011). A comprehensive revision of the US monetary services (divisia) indexes. *Federal Reserve Bank of St. Louis Review*, 93(5):325–359.

Artis, M. J., Kontolemis, Z. G., and Osborn, D. R. (1997). Business cycles for G7 and european countries. (Group of 7). *The Journal of Business*, 70(2).

Barnett, W. A. (1978). The user cost of money. *Economics Letters*, 1(2):145–149.

Barnett, W. A. (1980). Economic monetary aggregates an application of index number and aggregation theory. *Journal of Econometrics*, (11):11–48.

Barnett, W. A. (2007). Multilateral aggregation-theoretic monetary aggregation over heterogeneous countries. *Journal of Econometrics*, 136(2):457–482.

Barnett, W. A. and Chauvet, M. (2011). How better monetary statistics could have signaled the financial crisis. *Journal of Econometrics*, 161:6–23.

Barnett, W. A. and Gaekwad, N. (2018). The demand for money for EMU: a flexible functional form approach. *Open Economies Review*, 29(2):353–371.

Barnett, W. A., Liu, J., Mattson, R., and Noort, J. (2013). The new CFS divisia monetary aggregates: Design, construction, and data sources. *Open Economies Review*, 24(1):101–124.

Belongia, M. T. and Ireland, P. N. (2014). The Barnett critique after three decades : a New Keynesian analysis. *Journal of Econometrics*, 183(1).

Belongia, M. T. and Ireland, P. N. (2015). A working solution to the question of nominal gdp targeting. *Macroeconomic Dynamics*, 19(3).

Borio, C. E., Drehmann, M., and Xia, F. D. (2018). The financial cycle and recession risk. *BIS Quarterly Review December 2018*.

Chen, W. and Nautz, D. (2015). The information content of monetary statistics for the Great Recession : evidence from Germany. *SFB 649 discussion paper 2015-027*, Humboldt University Berlin, Germany.

Constâncio, V. (2018). Past and future of the ECB monetary policy. Speech by Vítor Constâncio, Vice-President of the ECB, at the Conference on Central Banks in Historical Perspective: What Changed After the Financial Crisis?, organised by the Central Bank of Malta, Valletta, 4 May 2018.

Darvas, Z. (2015). Does money matter in the euro area? evidence from a new divisia index. *Economics Letters*, 133(C):123–126.

De Nederlandsche Bank (2018). Combined balance sheet of DNB and Dutch-based MFI, adjusted for breaks (month).

Estrella, A. and Mishkin, F. S. (1998). Predicting U.S. recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1):45–61.

European Central Bank (2003). Editorial, Monthly Bulletin, May 2003. pages 5 – 8.

European Central Bank (2012). Manual on MFI balance sheet statistics

European Central Bank (2017). Manual on MFI interest rate statistics

Fisher, P., Hudson, S., and Pradhan, M. (1993). *Bank of England Quarterly Bulletin*, pages 240–255.

Goodhart, C. A. E., Mills, T. C., and Capie, F. (2019). The slope of the term structure and recessions: evidence from the UK, 1822-2016. *CEPR Discussion Paper*, (DP 13519).

Hancock, M. (2005). *Bank of England Quarterly Bulletin*, pages 39–46.

Offenbacher, E. A. and Shemesh, S. (2011). Divisia monetary aggregates for Israel: background note and metadata

Schunk, D. (2001). The relative forecasting performance of the divisia and simple sum monetary aggregates. *Journal of Money, Credit, and Banking*, 16:272–283.

Stracca, L. (2004). Does liquidity matter? Properties of a divisia monetary aggregate in the Euro area. *Oxford Bulletin of Economics and Statistics*, 66(3):309–331.

# Stochastic volatility model's predictive relevance for Equity Markets

**by**

**Per Bjarte Solibakke**[a]

**Abstract**

**This paper builds and implements multifactor stochastic volatility models. The main objective is volatility prediction and its relevance for equity markets. The paper outlines stylised facts from volatility literature showing density tails, persistence, mean reversion, asymmetry and long memory, all contributing to systematic dependencies. Applying long simulations from stochastic volatility (SV) models and filter volatility using a form of nonlinear Kalman filtering, the unobservables of the nonlinear latent variables can be forecasted with associated fit characteristics. The paper uses European equity data from United Kingdom (Ftse100) and Norway (Equinor) for relevance arguments and illustrational prediction purposes. Multifactor SV models seem to enrich volatility predictions empowering equity market relevance.**

**Classification:** C11, C63, G17, G32

**Keywords:** Stochastic Volatility, Bayesian Estimators, Metropolis-Hastings Algorithm, Markov Chain Monte Carlo (MCMC) Simulations, Nonlinear Kalman filter

[a]Per Bjarte Solibakke is Professor at Norwegian University of Science and Technology (NTNU).

E-mail: per.b.solibakke@ntnu.no

Telephone: +4790035606

**Corresponding author:**
Per Bjarte Solibakke, e-mail: per.b.solibakke@ntnu.no
Postal Adress:

Norwegian School of Science and Technology
Larsgårdsvn. 2,
6025 Ålesund
Norway

# Productivity and Real Exchange Rate: Investigating the Validity of the Balassa-Samuelson Effect in Five African Countries

**Joel Hinaunye Eita**[1]

*School of Economics, University of Johannesburg, South Africa, e-mail: hinaeita@yahoo.co.uk or jeita@uj.ac.za*

**Zitsile Zamantungwa Khumalo,**

*Department of Economics, North-West University, Mmabatho, South Africa, e-mail: zitsilek@gmail.com*

**Ireen Choga**

*Department of Economics, North-West University, Mmabatho, South Africa, e-mail: ireen.choga@nwu.ac.za*

[1] Corresponding author

# Productivity and Real Exchange Rate: Investigating the Validity of the Balassa-Samuelson Effect in Five African Countries


## ABSTRACT

Productivity in any economy is important because it increases economic performance and promotes economic growth. Given the importance of productivity, the study investigated the validity of the Balassa-Samuelson Effect in five African countries. The focal point of the Balassa-Samuelson (BS) effect is the relationship between productivity growth and real exchange rate appreciation. The study estimated the equilibrium real exchange with total factor productivity as an explanatory variable. Further, real exchange rate misalignment was derived and its effects on economic performance tested. The study employed more than one measure of economic performance in assessing the effects of real exchange rate misalignment on economic performance. The results revealed a valid Balassa-Samuelson effect in the selected countries and negative coefficients for real exchange rate misalignment. Recommendations emanating from this study include; countries pursuing policies that contain misalignment of the real exchange rate because sustained misalignment hampers economic growth, performance and competitiveness.



**Keywords:** Real Exchange Rate, Real Exchange Rate Misalignment, Balassa-Samuelson

## 1. INTRODUCTION

The Balassa-Samuelson hypothesis is a result of an augmentation of the Purchasing Power Parity (PPP). Balassa and Samuelson argue that labour productivity differentials between tradable and non-tradable sectors translate to fluctuations in real costs and relative prices and leads to divergences in the exchange rate adjusted national prices (Asea and Mendoza, 1994). Moreover, the hypothesis describes real exchange rate volatility using the differential productivity of the tradable and non-tradable sectors (Romanov, 2003).

The Balassa-Samuelson hypothesis relies on the differential productivity levels between a country and its trading partners; it also assumes that productivity growth is biased as it favours the traded goods sector. Countries with higher productivity levels than its trading partners are inclined to have greater productivity differentials in traded goods sectors than non-traded goods sectors. Higher productivity in traded goods permits the sector to move labour from the non-traded goods sector thereby increasing costs in the non-traded goods sector. In turn, this requires a higher relative price of non-traded goods to maintain profitability in the sector (Montiel, 2007).

The tradable goods sector constitutes manufacturing and agriculture while the non-tradable goods sector constitutes services. In level form, the Balassa effect forecasts that countries with relatively lower productivity in tradable goods than in non-tradable goods (synonymous with developing countries); have lower price levels than in other countries (Coudert, 2004). Higher productivity in the tradable sector of wealthy countries leads to an increment in the general level of prices and the real exchange rates; while low productivity in the tradable sector of poor countries is normally inclined to maintain or reduce the general level of prices and more devaluated/depreciated exchange rates under the Balassa-Samuelson effect (Martinez-Hernandez, 2017).

Several studies in literature have tested the validity of the Balassa-Samuelson hypothesis for both developed and developing countries (Drine and Rault (2002); Gubler and Sax (2011)). Other studies further explored incidences of real exchange rate misalignment (Kakkar and Yan (2012)) and other studies examined the influence of real exchange rate misalignment on economic performance (Sallenave (2010);

Vieira and MacDonald (2012)). Previous studies investigated the Balassa-Samuelson effect in different contexts with the following gap identified. These previous studies use inappropriate proxy variables (such as real GDP or relative GDP) for productivity or technology. This study computes total factor productivity by using the Cobb-Douglas production function. According to Tintin (2009) total factor productivity (TFP) is a better representation of productivity or technology. Like previous studies, this study attempts to test the validity of the Balassa-Samuelson hypothesis for a selection of African countries.

The study is organised as follows: Section 2 presents the literature review. Sections 3 and 4 present the methodology and the empirical models estimated. Sections 5 to 6 present and explain the empirical results, while the conclusion and recommendations are presented in Section 7.

## 2. LITERATURE REVIEW

Ito, Isard and Symansky (1999) investigated the Balassa-Samuelson hypothesis in high-growth Asian countries where a generally pronounced Balassa-Samuelson effect was observed in Japan, Korea, and Taiwan. Giacomelli (1998) found results in support of the Balassa-Samuelson effect in twenty-four developing countries and fourteen OECD economies. While Faria and León-Ledesma (2003) found no support of the Balassa-Samuelson effect in the long-run between two countries (the UK and US, German and Japan and Japan and the US).

Gubler and Sax (2011) investigated the robustness of the Balassa-Samuelson hypothesis for panel of OECD countries for the period of 1970 to 2008. No evidence of the Balassa-Samuelsson hypothesis was found. Omojimite and Oriavwote (2012) examined the relationship between the Naira real exchange rate and macroeconomic performance and the Balassa-Samuelson hypothesis in Nigeria. The results implied a valid hypothesis in Nigeria.

Kakkar and Yan (2012) examined the Balassa-Samuelson effect for six Asian economies and further examined real exchange rate misalignment with findings that most real exchange rates were overvalued. Sallenave (2010) created a model adjusted for the Balassa-Samuelson effect in a study about the growth effects of real effective exchange rate misalignments for the G20 countries. Similarly, Vieira and

MacDonald (2012) studied the impact of real exchange rate misalignment on long-run growth for a set of ninety countries with adjustments for the Balassa-Samuelson effect. They found that exchange rate misalignment impacted economic growth.

Based on the empirical inconclusiveness established in previous studies, this study investigated the Balassa-Samuelson effect in five African countries. The reviewed studies investigated the Balassa-Samuelson effect in different contexts with the following gaps identified:

i) Other studies use proxy variables for total factor productivity whereas this study computes productivity by using the Cobb-Douglas production function.
ii) Only a few studies further investigated real exchange rate misalignment such as Kakkar and Yan (2012).

However, other studies extended their analysis and examined the influence of real exchange rate misalignment on economic performance (Sallenave (2010); Vieira and MacDonald (2012). The study creates an equilibrium real exchange rate by substituting permanent values of the explanatory variables into an estimated cointegration relationship. The estimated coefficients are imposed on the permanent values of the explanatory variables using the Hodrick Prescott (HP) filter.

Previous studies used only the gross domestic product as a measure of economic performance. The gross domestic product indicates the amount of output produced in a country. This study employs the unit labor cost (ULC) which indicates the economic competitiveness of a country as an additional measure of economic performance. The ULC forms part of the best complementary indicators of an economy. It is regularly applied to evaluate economic development in numerous countries, both individual and grouped. It gives a holistic view of the quality of economic growth by placing into context the overall production output of an economy (GDP), labour productivity, wage and other costs connected with the workforce and price development (Lipská, Vlnková and Macková, 2005). Moreover, the use of multiple measures of economic performance offers improved robust results considered authentic or dependable through triangulation (Kuorikoski, Lehtinen and Marchionni, 2007).

## 3. METHODOLOGY

### 3.1. Model Specification

All variables are expressed in logarithms to reduce data variability and the empirical model is expressed as follows.

$$LRER_{i,t} = \alpha_0 + \beta_1 LPROD_{i,t} + \beta_2 LTOT_{i,t} + \beta_3 NFA_{i,t} + \varepsilon_t$$

[1]

The study used the weighted average of a country's currency in relation to an index or basket of other major currencies, that is, the real effective exchange rate to represent the real exchange rate (LRER). An increase in LRER is appreciation and a decrease is depreciation. (LPROD) represents total factor productivity, an increase in (LPROD) leads to real exchange rate appreciation. LPROD captures the Balassa-Samuelson effect which hypothesises that rapid economic growth is associated with real exchange rate appreciation because of differential productivity growth between tradable and non-tradable sectors.

The productivity, LPROD, is computed by using the Cobb-Douglas production function where: $Y = AL^{\beta}K^{\alpha}$ , Y is total production, A is total factor productivity, L is the labour input, K is the capital input and α and β are the output elasticities of capital and labour.

To obtain total factor productivity, the formula becomes:

$$\frac{Y}{K^{\alpha}L^{\beta}} = A \qquad [2]$$

Tintin (2009) put forth the argument in literature, that of total factor productivity (TFP) being a better representation of productivity. However, the impediment with TFP is the difficulty in computing and the unavailability of relevant data.

LTOT denotes Terms of Trade; LTOT presents an ambiguous impact on real exchange rate due to income and substitution effects. A rise in capital inflows permits an expansion of absorption and consequently an appreciation of the real

exchange rate. LNFA denotes net foreign assets; net foreign assets are cumulative current account of net capital transfers adjusted for the effects of capital gains and losses on inward and outward FDI as well as on portfolio equity holdings (Lane and Milesi-Ferretti, 2000). LNFA has a positive relationship in long-run equilibrium with the real exchange rate (Bleaney and Tian, 2014).

### 3.2. Real Exchange Rate Misalignment

Real exchange rate misalignment is the deviation of the real exchange rate from its desired equilibrium. Misalignment is calculated by subtracting the equilibrium real exchange rate from the actual exchange rate.

$$Misalignment = RER_t - ERER_{t-1}$$

[3]

Where misalignment is the real exchange rate misalignment, $RER$ is the actual real exchange rate, and $ERER$ is the equilibrium real exchange rate. A positive value implies an overvalued real exchange rate while a negative value implies an undervalued real exchange rate. Both occurrences have implications on the economic performance of a country.

#### 3.2.1. Real Exchange Rate and Economic Performance

Like studies by Sallenave (2010) and Vieira and MacDonald (2012), this study examined the Balassa-Samuelson and further examined real exchange rate misalignment and its implications for economic performance.

This study departs from previous studies by using difference measures of economic performance. Two models were estimated, one with GDP as an indicator of economic performance and the other indicator being unit labour costs. Unit labour costs was computed as remuneration of employees divided by total output of the Namibian economy as in Eita and Jordaan (2013). This study follows the same progression for each country. Eita and Jordaan (2013) mentioned that real exchange rate misalignment could upsurge unit labour costs thereby weakening the competitiveness of a country. Lipská, Vlnková and Macková (2005) acknowledged the unit labour cost (ULC) indicator as a paramount indicator of an economy.

The models are expressed as follows:

i) **Model 1** follows the form of Barro's Simple Model of Endogeneous Growth (1990):

Barro's original model was expressed as:

$$Y = AK^{\alpha}G_{Y}^{\beta}$$

[5]

Where Y = Real output; A = Productivity index; K = Private capital; G = Public investment.

- The model for the study is expressed as follows:

$$Y = \alpha_0 + \alpha_2 LPROD_{i,t} + \alpha_3 LPINV_{i,t} + \alpha_4 LPOP_{i,t} + \alpha_5 MISA_{i,t} + \varepsilon_{i,t}$$

[6]

Where LPROD is total factor productivity, LPINV is public investment proxied by government spending, LPOP denotes population growth and MISA is real exchange rate misalignment.

Increased productivity propels growth. Greater productivity enables firms to produce more output with the same input resources thus higher revenues and ultimately a higher gross domestic product. Public investment may affect growth positively or negatively. Primarily, public investment causes increased production which increases output and the employment level (Rabnawaz and Jafar, 2015). Population growth increases causes a decline in growth whilst real exchange rate misalignment impacts negatively on economic growth; an increase in misalignment leads to a reduction in economic growth.

ii) **Model 2** is expressed as:

$$LULC_{i,t} = \alpha_0 + \alpha_1 LINF_{i,t} + \alpha_2 LFDI_{i,t} + \alpha_3 LEXP_{i,t} + \alpha_4 MISA_{i,t} + \varepsilon_{i,t}$$

[7]

Where LULC is the unit labour cost computed by remuneration of employees divided by total output; LINF denotes inflation and LFDI denotes total investment. Higher inflation leads to a rise in unit labour costs thus there is a native relationship between inflation and unit labour costs. A negative relationship exists between foreign direct investment and unit labour costs; a fall in the unit labor costs encourages LFDI. LEXP denotes exports of goods and services. There is an inverse relationship

between export developments and developments in unit labour costs. Real exchange rate misalignment negatively affects unit labour costs as well.

## 4. DATA DESCRIPTION

The data for the period 1991 to 2016 was for five African countries which are the Democratic Republic of Congo (DRC), Mauritius (MAU), Morocco (MOR), South Africa (SA) and Tunisia (TUN) obtained from Quantec. The sample period and the countries under investigation were selected on the premise of data availability.

### 4.1. Estimation Technique

The pooled mean group estimator (PMG) is employed. The PMG is consistent and efficient in the estimation of parameters' averages and long-run estimators for large sample sizes (Pesaran and Smith, 1995). Parameters are independent across groups and potential homogeneity between groups is not considered. Short-run dynamic specifications differ from country to country and long-run coefficients are controlled to be similar. The PMG contains pertains maximum likelihood estimation of an ARDL model which can be written as an error correction model (ECM); it is a panel version of the ARDL (Saxegaard, Roudet and Tsangarides, 2007).

The study conducted the Levin, Lin and Chu test (LLC Test), Im, Pearson and Shin test (IPS) and the Kao test for cointegration**:**

#### **4.1.1.** *The Levin, Lin and Chu Test (LLC Test) and Im, Pesaran and Shin Test (IPS)*

The LLC panel unit root and Im-Pesaran-Shin (IPS) tests were utilised. The LLC was developed to cover the shortfall of individual unit root tests. Individual unit root tests are disadvantaged because of insufficient power against alternative hypotheses with rapid constant deviations from equilibrium (Baltagi, 2008). The Im-Pesaran-Shin (IPS) test is more flexible than the Levin-Lin-Chu test because it permits heterogeneous coefficients (Kunst, Nell and Zimmermann, 2011).

#### **4.1.2.** *The Kao Cointegration Test*

According to Chaiboonsri et al (2010), the Kao test begins with a panel regression model where X and Y are presumed to be nonstationary:

$$Y_{it} = X_{it}\beta_{it} + Z_{it}\gamma_0 + \varepsilon_{it}$$

and

$e^{\lambda}_{it} = \rho^{\lambda}_{it} + v_{it}$ are residuals from the estimated equation.

The null hypothesis and alternative hypothesis are expressed as:

- Null, no cointegration: $H_0 : \rho = 1$
- Alternative, cointegration: $H_1 : \rho < 0$

The Kao test used DF-Type test statistics and ADF test statistics to test for cointegration.

The existence of a cointegration amongst variables is followed by an estimation of the real exchange rate model. Based on its properties of providing optimal estimates of cointegrating regressions and accounting for serial correlation and endogeneity of regressors as proposed by Phillips and Hansen (1990), the fully modified least squares (FMOLS) was applied.

#### **4.1.3.** *The Fully Modified OLS Model*

Initially, the fully modified estimator was created to directly estimate cointegrating relationships by altering the traditional ordinary least squares. The traditional OLS is corrected for endogeneity and serial correlation. Previous simulation experience and empirical research has proven the good performance of FMOLS in relation to other methods estimating cointegrating relations as cited in Cappucio and Lubian (1992) and Hagreaves (1993) (Phillips, 1995).

## 5. ESTIMATION RESULTS

### 5.1. Unit Root (Stationarity) Tests

The variables were subjected to the LLC and the IPS stationarity tests. The results are presented in Table 1:

**Table1: Unit Root Test Results**

| Variable | LLC Test | | IPS Test | |
|---|---|---|---|---|
| | Levels | | Levels | |
| | Constant | Constant and Trend | Constant | Constant and Trend |
| **LRER** | -0.07054 (0.4719) | -1.14415 (0.1263) | 0.79557 (0.7869) | -1.07910 (0.1403) |
| **LPROD** | -1.15240 (0.1246) | -0.92410 (0.1777) | 1.50259 (0.9335) | -0.05751 (0.4771) |
| **LTOT** | -0.68398 (0.2470) | -0.32210 (0.3737) | 0.37277 (0.6453) | -1.37595 (0.0844)* |
| **LNFA** | 0.01093 (0.5044) | 0.79499 (0.7867) | 1.92327 (0.9728) | 3.01484 (0.9987) |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LRER** | -8.16733 (0.0000)* | -6.60331 (0.0000)* | -7.04103 (0.0000)* | -5.65658 (0.0000)* |
| **LPROD** | -9.43767 (0.0000)* | -7.70112 (0.0000)* | -8.46954 (0.0000)* | -6.76304 (0.0000)* |
| **LTOT** | -8.27779 (0.0000)* | -7.25819 (0.0000)* | -8.37758 (0.0000)* | -6.92036 (0.0000)* |
| **LNFA** | -5.00458 (0.0000)* | -4.11256 (0.0000)* | -3.19374 (0.0007)* | -2.31718 (0.0102)* |

*p-values are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 1%, 5% and 10% levels of significance

Table1 depicts the LLC and the IPS panel unit root test results at levels and first difference. At levels, only LTOT is stationary at 10% level of significance. LRER, LPROD, LNFA become stationary at first difference, they are integrated of order one I(1) while LTOT is I(0). The conclusion drawn is that variables are stationary therefore the null hypothesis of the presence of a unit root is rejected.

### 5.2. Estimation of the Real Exchange Rate Cointegration Results

Table 2 presents the Kao panel cointegration test results. The decision rule of this test is rejecting the null hypothesis of no cointegration when the p value is less than 5%. The results in this study are consistent with this rule therefore there is cointegration amongst the variables.

**Table 2: Kao Cointegration Test Results**

| Kao Test | t-statistic | Probability |
|---|---|---|
| | -4.050948 | 0.0000* |

**NB:** The ADF is the residual-based ADF statistic. The null hypothesis is no cointegration. * Indicates that the estimated parameters are significant at the 5% level

### 5.3. Long-run coefficient – Fully Modified OLS Estimates (FMOLS)

The results exhibit the presence of a cointegration relationship amongst the variables therefore the fully modified OLS approach was employed to estimate the long run RER model and the results are presented in Table 3.

**Table 3: FMOLS long run - estimation results. Dependent variable: LRER**

| Sample Period | 1991-2016 |
|---|---|
| **Explanatory Variables** | **Coefficients** |
| **LPROD** | 0.138157<br>(0.0943)* |
| **LTOT** | -0.665678<br>(0.0015)* |
| **LNFA** | -0.001194<br>(0.5424) |
| **R-squared**<br>**S.E. of regression** | 0.920705<br>0.200896 |

*p-values are in parentheses ()

*10 % statistically significant.

**5 % statistically significant.

***1 % statistically significant.

Table 3 presents the long-run coefficients results of the FMOLS estimator. The results reveal that LPROD is statistically significant and consistent with economic

theory. LTOT is statistically significant and consistent with economic theory. LNFA is not statistically significant and is in defiance of economic theory.

A 1% increase in LPROD will appreciate the real exchange rate by 0.1% thereby indicating a positive relationship between the two variables as stipulated by economic theory. This implies that the Balassa-Samuelson is valid and relevant to the selected African economies. In these countries, the Balassa-Samuelson theory holds. A 1% increase in LTOT will depreciate the real exchange rate by 0.7%. The relationship between terms of trade and the real exchange rate is negative and statistically significant.

## 6. COMPUTED REAL EXCHANGE RATE MISALIGNMENT

The RER misalignment is seen in figure 1. Generally, there were more periods of real exchange rate undervaluation than overvaluation. Real exchange rate overvaluation is not an ideal state as it impacts negatively on economic growth negatively, therefore economic policies adopted in these countries should circumvent such occurrences. Gylfason (2002) stated that sustained currency overvaluation deteriorates the trade balance, speculative attacks, increased foreign debt, decline in investment.

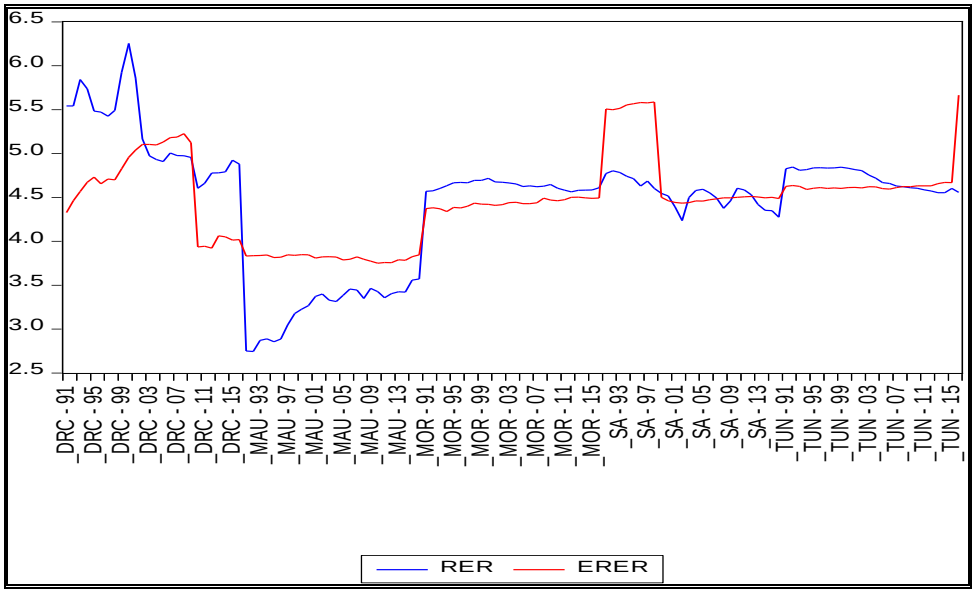



**FIGURE 1: ACTUAL AND EQUILIBRIUM EXCHANGE RATE**

*__DRC__-DEMOCRATIC REPUBLIC OF CONGO, __MAU__-MAURITIUS, __MOR__-MOROCCO, __SA__-SOUTH AFRICA, __TUN__-TUNISIA

*ERER is the equilibrium real exchange rate and RER is the actual real exchange rate

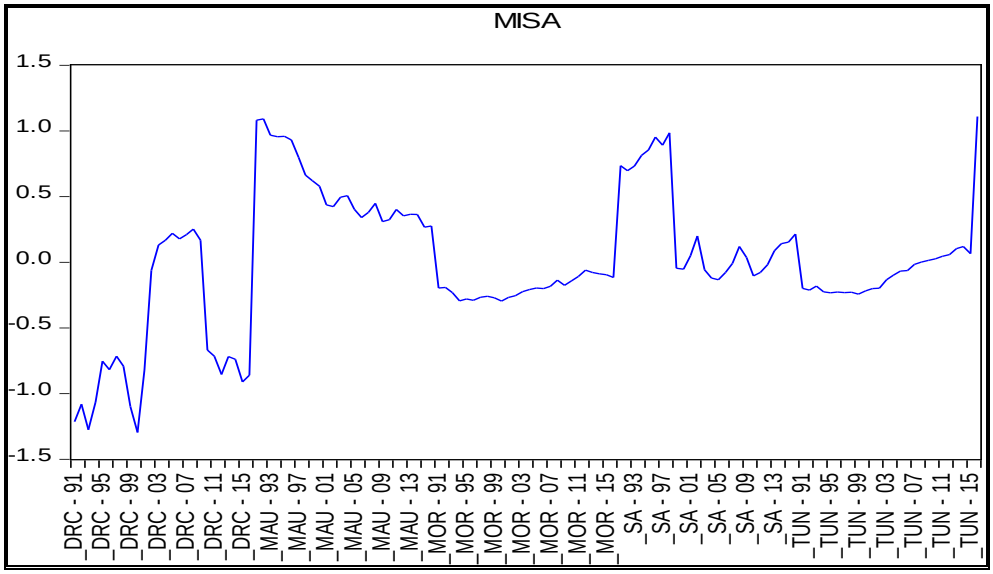



**FIGURE 2: REAL EXCHANGE RATE MISALIGNMENT**

*__DRC__-DEMOCRATIC REPUBLIC OF CONGO, __MAU__-MAURITIUS, __MOR__-MOROCCO, __SA__-SOUTH AFRICA, __TUN__-TUNISIA

*MISA denotes real exchange rate misalignment

## 6.1. REAL EXCHANGE RATE MISALIGNMENT AND MACROECONOMIC PERFORMANCE

### 6.1.1. Test for Stationarity

The results of the unit roots test are presented in Tables 4 and 5:

**Table 4: Unit Root Test - Model 1**

| Variable | LLC Test | | IPS Test | |
|---|---|---|---|---|
| | Levels | | Levels | |
| | Constant | Constant and Trend | Constant | Constant and Trend |
| **LGDP** | -3.82959<br>(0.0001)* | 0.01853<br>(0.5074) | -0.34338<br>(0.3657) | 1.91790<br>(0.9724) |
| **LPROD** | -1.15240<br>(0.1246) | -0.92410<br>(0.1777) | 1.50259<br>(0.9335) | -0.05751<br>(0.4771) |
| **LPINV** | -1.11885<br>(0.1316) | -0.62953<br>(0.2645) | -0.02338<br>(0.4907) | -0.89388<br>(0.1857) |
| **LPOP** | -1.27066<br>(0.1019) | -0.46339<br>(0.3215) | -0.68724<br>(0.2460) | -1.37608<br>(0.0844)* |
| **MISA** | -0.24806<br>(0.4020) | 0.46208<br>(0.6780) | 1.72121<br>(0.9574) | 1.22687<br>(0.8901) |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LGDP** | -7.44804<br>(0.0000)* | -2.18594<br>(0.0144) | -9.07668<br>(0.0000)* | -3.37360<br>0.0004 |
| **LPROD** | -9.43767<br>(0.0000)* | -7.70112<br>(0.0000)* | -8.46954<br>(0.0000)* | -6.76304<br>(0.0000)* |
| **LPINV** | -12.9409<br>(0.0000)* | -11.0543<br>(0.0000)* | -13.0238<br>(0.0000)* | -11.2803<br>(0.0000)* |
| **LPOP** | -2.20397<br>(0.0138) | 0.58811<br>(0.7218) | -4.61573<br>(0.0000)* | -2.81867<br>(0.0024) |
| **MISA** | -5.49298<br>(0.0000)* | -4.24400<br>(0.0000)* | -6.63116<br>(0.0000)* | -6.34736<br>(0.0000)* |

*p-values are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 1%, 5% and 10% levels of significance

**Table 5: Unit Root Test - Model 2**

| Variable | LLC Test | | IPS Test | |
|---|---|---|---|---|
| | Levels | | Levels | |
| | Constant | Constant and Trend | Constant | Constant and Trend |
| **LULC** | -0.85864<br>0.1953 | -0.05175<br>0.4794 | -1.89422<br>0.0291 | 0.22062<br>0.5873 |
| **LINF** | -4.78906<br>(0.0000)* | -4.58762<br>(0.0000)* | -4.51500<br>(0.0000)* | -4.91289<br>(0.0000)* |
| **LFDI** | -3.85214<br>0.0001 | -4.71136<br>(0.0000 | -4.09967<br>0.0000 | -5.71165<br>0.0000 |
| **LEXP** | -1.05856<br>(0.1449) | -1.09570<br>(0.1366) | 1.51865<br>(0.9356) | -1.55802<br>(0.0596)* |
| **MISA** | -0.24806<br>(0.4020) | 0.46208<br>(0.6780) | 1.72121<br>(0.9574) | 1.22687<br>(0.8901) |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LULC** | -7.79949<br>(0.0000)* | -6.64457<br>(0.0000)* | -8.18060<br>(0.0000)* | -7.11526<br>(0.0000)* |
| **LINF** | -14.7073<br>(0.0000)* | -12.9614<br>(0.0000)* | -14.1363<br>(0.0000)* | -13.2194<br>(0.0000)* |
| **LFDI** | -12.9409<br>(0.0000)* | -11.0543<br>(0.0000)* | -13.0238<br>(0.0000)* | -11.2803<br>(0.0000)* |
| **LEXP** | -8.98711<br>(0.0000)* | -8.21762<br>(0.0000)* | -8.11363<br>(0.0000)* | -7.05212<br>(0.0000)* |
| **MISA** | -5.49298<br>(0.0000)* | -4.24400<br>(0.0000)* | -6.63116<br>(0.0000)* | -6.34736<br>(0.0000)* |

*p-values are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 1%, 5% and 10% levels of significance

For model 1, population growth (LPOP) and LGDP are stationary at levels while public investment (LPINV), productivity (LPROD) and misalignment become

stationary at first difference. For model 2 variables inflation (LINF) is I(0) while LULC, LFDI, LEXP and misalignment are I(1). These results therefore necessitate the rejection of the null hypothesis of a panel unit root.

### 6.1.2. Real Exchange Rate Misalignment and Macroeconomic Performance Cointegration Results

The study found evidence of cointegration from Kao's panel cointegration tests, which rejected the null hypothesis of no cointegration because the p value is less than 5%, therefore there is a cointegration relationship amongst the variables.

**Table 6: Kao Cointegration Test Results for Models 1 and Model 2**

| Model 1 | t-statistic | Probability |
|---|---|---|
| **ADF** | -3.071796 | 0.0011* |
| **Model 2** | **t-statistic** | **Probability** |
| **ADF** | -2.445284 | 0.0072* |

**NB:** The ADF is the residual-based ADF statistic. The null hypothesis is no cointegration. * Indicates that the estimated parameters are significant at the 5% level

Based on the results, the study concluded that a panel long-run equilibrium relationship among the variables exists.

### 6.2. PMG Estimation Results

This section presents the PMG estimation results for the two models using different measures of economic performance:

**Table 7: PMG Results – Model 1**

**Dependent Variable:** LGDP

| Long-run Coefficients | |
|---|---|
| **LPROD** | 0.370866<br>(0.0000)* |
| **LPINV** | 0.723908<br>(0.0000)* |
| **LPOP** | -0.129164<br>(0.0001)* |
| **MISA** | -0.370738<br>(0.0003)* |
| **Short-run Coefficients** | |
| **ΔLPROD** | -0.027976<br>(0.7073) |
| **ΔLPINV** | 0.045155<br>(0.5857) |
| **ΔLPOP** | -0.155733<br>(0.1479) |
| **ΔMISA** | 0.048788<br>(0.1815) |
| **Error Correction Coefficient** | -0.018867<br>(0.7722) |

**Table 8: PMG Results – Model 2**

**Dependent Variable:** LULC

| Long-run Coefficients | |
|---|---|
| **LINF** | -0.063194<br>(0.2420) |
| **LFDI** | 0.048952<br>(0.7481) |
| **LEXP** | -0.241559<br>(0.0000)* |
| **MISA** | -0.111244<br>(0.4856) |
| **Short-run Coefficients** | |
| **ΔLINF** | -0.022329<br>(0.0619)* |
| **ΔLFDI** | -0.030851<br>(0.1684) |
| **ΔLEXP** | 0.093189<br>(0.6290) |
| **ΔMISA** | -0.158692<br>(0.1794) |
| **Error Correction Coefficient** | -0.176672<br>(0.2677) |

Tables 7 and 8 display the PMG estimation results for both models. Model 1 employed the GDP as a measure of economic performance while model 2 employed unit labour costs. In model 2, the long-run, the relationship between LULC and real exchange rate misalignment is negative as in model 1. However, misalignment is not statistically significant in the long-run in model 2.

## 7. CONCLUSION

The study investigated estimates of the Balassa-Samuelson effect on a selection of African countries given the importance of productivity for the growth of an economy. Additionally, real exchange rate misalignment was derived and its effects on economic performance were tested. Instead of using one measure of economic performance, the study used both the gross domestic product and unit labour costs.

The Balassa-Samuelson effect appeared to be valid for the selected African countries which are the Democratic Republic of Congo, Mauritius, Morocco, South Africa and Tunisia. The relationship between total factor productivity and the real exchange rate conformed to economic theory thereby confirming the validity of this theory. Both models revealed an undervalued real exchange rate in the long-run test for real exchange rate misalignment.

An undervalued real exchange rate is an ideal condition enhancing for economic growth and development in the Democratic Republic of Congo, Mauritius, Morocco, South Africa and Tunisia. Conversely, these countries need to monitor misalignment and reduce or control its impediment on economic growth and competitiveness. Furthermore, the study suggests that the Democratic Republic of Congo, Mauritius, Morocco, South Africa and Tunisia should pursue economic policies and strategies that contain real exchange rate misalignment to promote economic growth and competitiveness.

**References**


Asea, P.K., & Mendoza, E.G. (1994). The Balassa-Samuelson Model: A general-equilibrium appraisal. *Review of International Economics*, *2*(3), 244-267.

Balassa, B. (1964). The purchasing power parity doctrine: A reappraisal. *Journal of Political Economy*, *72*(6), 584 - 596.

Baltagi, B. (2008). *Econometric analysis of panel data*. San Francisco: John Wiley & Sons.

Barro, R.J. (1990). Government spending in a simple model of endogenous growth. *Journal of Political Economy*, *98*(5), 103-125.

Bergin, P.R., Reuven, G., & Taylor, A.M. (2004). *Productivity, tradability and the long-run price puzzle* (NBER Working Paper series, No. 10569). Cambridge, MA: National Bureau of Economic Research.

Bleaney, M., & Tian, M. (2014). *Classifying exchange rate regimes by regression methods* (Discussion Papers in Economics, No. 14/02). Nottingham: University of Nottingham School of Economics.

Chaiboonsri, C., Sriboonjit, J., Sriwichailamphan, T., Chaitip, P., & Sriboonchitta, S. (2010). A panel cointegration analysis: an application to international tourism demand of Thailand. *Annals of the University of Petrosani Economics*, *10*(3), 69 - 86.

Coudert, V. (2004). Measuring the Balassa-Samuelson effect for the countries of Central and Eastern Europe?. *Banque de France Bulletin Digest*, 122, 23-43.

Drine, I. & Rault, C. (2004). Does the Balassa-Samuelson Hypothesis hold for Asian countries? An Empirical Analysis using Panel Data Cointegration Tests. *Applied Econometrics and International Development, 4*(4), 59-84.

Eita, J. H., & Jordaan, A. C. (2013). Real exchange rate misalignment and economic performance in Namibia. Corporate Ownership & Control, I(3-4), 440-455.

Faria, J.R., & Leon-Ledesma, M. (2003). Testing the Balassa–Samuelson effect: Implications for growth and the PPP. *Journal of Macroeconomics*, *25*(2), 241-253.

Genius, M., & Tzouvelekas, V. (2008). The Balassa-Samuelson productivity bias hypothesis: Further evidence using panel data. *Agricultural Economics Review*, *9*(2), 31-41.

Giacomelli, D.S. (1998). *Essays on Consumption and the Real Exchange Rate* (Doctoral Dissertation), Massachusetts Institute of Technology: Massachusetts.

Gubler, M., & Sax, C. (2011). *The Balassa-Samuelson effect reversed: New evidence from OECD Countries* (WWZ Discussion Paper, No. 2011/09). Basel: University of Basel.

Gylfason, T. (2002). The real exchange rate always floats. *Australian Economic Papers*, 41(4), 369-381.

Ito, T., Isard, P., & Symansky, S. (1999). Economic growth and real exchange rate: an overview of the Balassa-Samuelson hypothesis in Asia. In O. Krueger and T. Ito (Eds), Changes in exchange rates in rapidly developing countries: theory, practice, and policy issues (109 -132). Chicago: University of Chicago Press.

Kakkar, V., & Yan, I. (2012). Real exchange rates and productivity: Evidence from Asia", Journal of Money, Credit and Banking, 44(2-3), 301-322.

Kunst, R., Nell, C., & Zimmermann, S. (2011). Summary based on Chapter 12 of Baltagi: Panel Unit Root Tests. Lecture Notes", Department of Economics, the University of Vienna.

Kuorikoski, J., Lehtinen, A. & Marchionni, C. (2007): "Economics as Robustness Analysis", http://philsci-archive.pitt.edu/3550/.

Lane, M.P.R., & Milesi-Ferretti, M.G.M. (2000). *External capital structure: Theory and evidence* (IMF Working Paper, No WP/00/152). Washington: International Monetary Fund.

Lipská, E., Vlnková, M., & Macková, I. (2005). Unit labour costs. *BIATEC*, *X111*(2005).

Martinez-Hernandez, F.A.M. (2017). The political economy of real exchange rate behavior: theory and empirical evidence for developed and developing countries, 1960-2010. *Review of Political Economy*, 29(4), 566-596.

Montiel, P.J. (2007). *Equilibrium real exchange rates, misalignment and competitiveness in the southern cone* (Vol. 62)", United Nations Publications.

Omojimite, B.U., & Oriavwote, V.E. (2012). Real exchange rate and macroeconomic performance: testing for the Balassa-Samuelson hypothesis in Nigeria. *International Journal of Economics and Finance*, 4(2), 127 -134.

Pesaran, M.H., & Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1), 79-113.

Phillips, P.C. (1995). Fully modified least squares and vector autoregression", *Econometrica*, *63*(5), 1023-1078.

Phillips, P.C., & Hansen, B.E. (1990). Estimation and inference in models of cointegration: A simulation study. *Advances in Econometrics*, *8*, 225-248.

Rabnawaz, A., & Sohail, J.R. (2015). Impact of public investment on economic growth. *South Asia Journal of Multidisciplinary Studies (SAJMS), 1*(8), 62-75.

Romanov, D. (2003). *The real exchange rate and the Balassa-Samuelson hypothesis: An appraisal of Israel's case since 1986 (*Bank of Israel Discussion Paper, No. 2003.09). Jerusalem: Banḳ Yiśra'el, Maḥlaḳat ha-meḥḳar.

Sallenave, A. (2010). Real exchange rate misalignments and economic performance for the G20 countries. *Economia Internazionale*, 1, 59-80.

Saxegaard, M., Roudet, S., & Tsangarides, C.G. (2007). *Estimation of equilibrium exchange rates in the WAMEU: A robust analysis* (IMF Working Paper, WP/07/94). Washington: International Monetary Fund.

Suleiman, H., & Muhammad, Z. (2011). *The real exchange rate of an oil exporting economy: Empirical evidence from Nigeria* (FIW Working Paper, No. 72). Dundee: Dundee Business School.

Tintin, C. (2009). Testing the Balassa-Samuelson hypothesis: Evidence from 10 OECD Countries (Master Thesis). Lund: University of Lund.

Vieira, F.V., & MacDonald, R. (2012). A panel data investigation of real exchange rate misalignment and growth. *Estudos Econômicos (São Paulo), 42*(3), 433-456.

# Estimating the Equilibrium Real Exchange Rate, Misalignment and Economic Performance in Selected African Countries

**Joel Hinaunye Eita**[1]

*School of Economics, University of Johannesburg, South Africa, e-mail: jeita@uj.ac.za or hinaeita@yahoo.co.uk*

**Zitsile Zamantungwa Khumalo,**

*Department of Economics, North-West University, Mmabatho, South Africa, e-mail: zitsilek@gmail.com*

**Ireen Choga**

*Department of Economics, North-West University, Mmabatho, South Africa, e-mail: ireen.choga@nwu.ac.za*

## Abstract

This study estimates the equilibrium real exchange rate and the resulting misalignment for selected African countries. It then tests the impact of real exchange rate misalignment on economic performance. The fundamental approach model was used to estimate, the equilibrium real exchange rate and resulting misalignment. The results revealed that the real exchange rate was misaligned. Additionally, both negative and positive coefficients for real exchange rate misalignment for the different models and samples are revealed, indicating periods of undervaluation and overvaluation of the real exchange rate.

**Keywords:** Real Exchange Rate, Real Exchange Rate Misalignment, Macroeconomic Fundamentals Economic Growth, Panel Data

[1] Corresponding author

## 1. Introduction

The real exchange rate (RER) has gained increasing attention over the years (Elbadawi and Soto 1997). Today, the real exchange rate is predominantly the focus of debates on economic development, growth strategies, structural adjustment and economic stabilisation. Economic research has further embarked on missions to uncover its empirical determination, the calculation of its equilibrium path, the assessment of its misalignment and the estimation of a set of fundamentals consistent with internal and external balances.

The real exchange rate is a key relative price in any economy, hence, its importance and emphasis on the maintenance of its stability. In the same vein, the real exchange rate is a popular real target in developing countries. Countries employ strategies to control the level of the real exchange rate allowing for domestic or external shocks to attain a different level which is normally depreciated (Reinhart and Vegh 1995).

Economists have thought exchange rate variations to be dictated by changes in one or more of the important economic variables proposed by the main theories advocated in the leading schools of economic thought. However, there has been no consensus about the fundamentals that should determine exchange rates. Moreover, it has been acknowledged that exchange rates could be disproportionate to fundamentals for substantial timeframes (Cencini 2005).

Thus, the determination of the real exchange rate through macroeconomic fundamentals has been an enduring debate in literature. The ability of macroeconomic models to explain exchange rates has been questioned since the early 1980s (Devereux 1997). Studies in international economics have struggled to establish the link between floating exchange rates to macroeconomic fundamentals such as money supplies, outputs, and interest rates (Engel and West 2005).

This puzzle is about the weak short-run relationship between the exchange rate and its macroeconomic fundamentals; for example, fundamentals such as interest rates, inflation rates and output do not elucidate the short-term volatility in exchange rates (Evrensel 2013). Despite this predicament, the largely unstable relationship between exchange rates and macroeconomic fundamentals is well documented in literature (Bacchetta, Van Wincoop and Beutler 2009, Engel and West 2005, Sarno and Schmeling 2013).

The standard models of exchange rates and macroeconomic fundamentals propose that exchange rates are determined by expected future fundamentals thereby suggesting that current exchange rates have predictive information about future fundamentals (Sarno and Schmeling 2013). They provide evidence that exchange rates forecast fundamentals, which infers that fundamentals are a crucial determinant of exchange rates (Sarno and Schmeling 2013). Exchange rate theories by Engel and West (2005) expressed that fundamental variables influenced the exchange rate but floating exchange rates between countries with generally comparable inflation rates were estimated as random walks. Engel and West (2005) envisaged that their findings would change the landscape of the exchange rate debate as they found an inverse link between fundamentals and the exchange rate. This implied that exchange rates helped forecast the fundamentals. They further concluded that exchange rates and fundamentals are linked in a way which is consistent with asset pricing models of the exchange rate.

However, empirical models applied in the late 1980s tended to neglect the likelihood of the presence of a long-run relationship between the fundamentals and the exchange rate. In the beginning of the 1990s, structural models were employed to test for this long-run relationship. An observation concerning the structural models which had their premise in cointegration relationships was made; they were seen to improve the evidence in favour of predictability in the long-run (MacDonald and Taylor 1993, 1994).

Other economic studies have documented the association between high volatility of the exchange rate and macroeconomic fundamentals. Bacchetta, van Wincoop and Beutler (2009) attributed this high volatility to large and recurrent changes in the relationship between the exchange rate and macro fundamentals; these occur when structural parameters in the economy are obscure and transform gradually. Bacchetta, Van Wincoop and Beutler (2009) concluded that the reduced form relationship between exchange rates and fundamentals was determined by expectations of parameters and not by structural parameters.

Where the real exchange rate is concerned, for a typical African country, their economies are dominated by unstable and uncompetitive exchange rates and equally unstable macroeconomic fundamentals. The imbalances in some instances may be exacerbated by changes in the macroeconomic fundamentals which may lead to real exchange rate misalignment. Real exchange rate misalignment in turn influences economic performance.

Given the pertinence of the real exchange rate, fundamentals and real exchange rate misalignment, various studies have been conducted (Miyajima 2007, Ozsoz and Akinkunmi 2012, Mkenda 2001). However, most of the research on real exchange rate behaviour generally overlooks the impact of real exchange rate misalignment on economic performance. Most research studies do not consider real exchange rate misalignment. The limited studies such as (Eita and Sichei 2014, Ndlela 2012, Mkenda 2001) were based on single countries that cannot be generalised to Africa. This study fills this gap in literature by investigating real exchange rate misalignment and economic performance in a panel of selected African countries (Algeria, Cameroon, Central African Republic, Equatorial Guinea, Gabon, Gambia, Ghana, Lesotho, Morocco, Nigeria, South Africa, Sierra Leone, Togo, Tunisia, Uganda and Zambia). Moreover, the study extends the previous analysis by Ghura and Grennes (1993) and uses high frequency data, that is, annual data from 1990 to 2016.

The study is organised as follows. Section 2 presents the literature review. Sections 3 and 4 present the methodology and the empirical models estimated. Sections 5 to 7 present and explain the empirical results, while the conclusion and recommendations are presented in Section 8.

## 2. Literature Review

### 2.1. The Theory of Real Exchange Rates

Because of the failure of the Bretton Woods system, major currencies around the globe began to float against each other. During this episode, the monetary approach which assumed that the purchasing power parity exchange (PPP) rate held constantly was the main method of determining exchange rates (Taylor and Taylor 2004).

In 1978, Frenkel concurred that the PPP was constant and further promoted the idea of the PPP being considered as a theory of exchange rate determination. The notion that the PPP was useful in providing a guide to the general trend of exchange rates was brought forward. But the mid-1980s brought a wave of doubt about the PPP as researchers reached a conclusion opposite to the original notion. This cast a shadow of doubt on the role of PPP as a rule-of-thumb predictive model and its position as equilibrium condition. The PPP had supposedly collapsed (Lothian and Taylor 1997).

In view of this theory, numerous empirical studies were conducted to establish the validity of the purchasing power parity; moreover, investigations into the monetary approach and its impact on the exchange rate were undertaken with encouraging results. These results were attributed to the stability of the US dollar in the early days of the floating system. Thereafter, the US dollar became increasingly volatile and this exposed the inability of the PPP to be constant thus the monetary approach was rejected. The collapse of the PPP was identified easily through the examination of the real exchange rate. However, the PPP still presented a certain measure of the real exchange rate relating to PPP, this as well as the changes in the real exchange rate still need to reflect deviations from PPP (Taylor and Taylor 2004).

The PPP real exchange rate $({}^{e}ppp)$ was defined as equal to the nominal exchange rate $(E)$ corrected (that is, multiplied by the ration of the foreign price level $(P^{*})$ to the domestic price level: ${}^{e}ppp = EP^{*}/P$ depending on whether $_{P}$ and $P^{*}$ are consumer price indexes or producers price indexes; ${}^{e}ppp$ thus amounts to the relative price of foreign to domestic consumption of production of baskets. This definition of real exchange rates was employed by some policymakers due to the challenges experienced in explaining the relative price of tradables to non-tradables (Edwards 1988). Studies like Abuaf and Jorion (1990) suggested that long-run PPP might hold and further called into question the notion that real exchange rates followed a random walk. Other studies like Isard (1978) had cast doubt on the ability of the PPP as a theory to present the correct predictions of exchange-rate behaviour in the short run.

Ricci (2005) further discredited the PPP theory by stating that indications in literature were that the PPP was an inappropriate model for ascertaining equilibrium exchange rates; this was largely due to the slow pace at which real exchange rates returned to a constant level (which is the long-run equilibrium as implied by the PPP assumption). Literature has largely focused on the equilibrium relationship between the real exchange rate and various economic fundamentals and has moved away from PPP-based measures of the equilibrium exchange rate (Ricci 2005).

Some of the economic fundamentals identified for developing countries include commodity price movements (or the terms of trade), productivity and real interest rate differentials, measures of openness of the trade and exchange system, the size of the fiscal balance or of government

spending, and net foreign assets. These variables are employed based on a simple neoclassical theoretical framework. This framework is of the view that prices of tradable goods are equalised across countries and aims to depict the reflection of changes in the real exchange rate in relation to the relative price of non-tradables across countries. In most instances, the PPP neglects the evolution of fundamentals thus rendering it inaccurate. The PPP must then be substituted by the natural real exchange rate produced by the fundamentals (Stein 1994).

### 2.2. The Real Exchange Rate and Macroeconomic Fundamentals

The fundamental determinants of the real exchange rate are the real variables that have an influence in determining a country's internal and external equilibrium. These variables and the real exchange rate mutually determine the internal and external equilibrium position of a country. In reality, there are an extensive number of such factors, but analytical and policy discussion only focuses on the most vital. Real exchange rate fundamentals have been separated into two classes: external fundamentals which encompass international prices and world interest rates amongst other factors and domestic fundamentals which encompass import tariffs, government expenditure amongst other factors (Edwards 1987).

Edwards (1988) developed a model for real exchange rate determination where both real and nominal factors played a role in the short run. The long-run only employs real factors or fundamentals which impact on the real exchange rate. The model contained developing economy macroeconomic features such as exchange controls, trade barriers and a freely determined parallel market for foreign financial transactions.

Three goods are considered in the model: exportables, importables and non-tradables in a small open economy setting. The long-run equilibrium real exchange rate is defined as a function of the fundamentals and changes in these fundamentals result in changes in the equilibrium RER; some of these changes include increment in tariffs and terms of trade disturbance.

The model is presented as follows:

Portfolio Decisions:

$$A = M + \delta F$$

(1)

$$a = m + \rho F, where\ a = A/E; M = M/E; \rho = \delta/E$$

(2)

$$m = \sigma(\dot{\delta}/\delta)\rho F;$$

(3)

$$\dot{F} = 0$$

(4)

Demand Side:

$$p_M = EP^* + r; \quad (5)$$

$$C_M = C_M(e_M, a) \quad (6)$$

$$C_N = C_N(e_M, a) \quad (7)$$

Supply Side:

$$Q_X = Q_X(e_X); \quad (8)$$

$$Q_N = Q_N(e_X); \quad (9)$$

Government Sector:

$$G = P_N G_N + EP_M^* G_M \quad (10)$$

$$\frac{EP_M^* G_M}{G} = \lambda \quad (11)$$

$$G = t + \dot{D} \quad (12)$$

External Sector:

$$CA = Q_X(e_X) - P_M^* C_M(e_M{}'a) - P_M^* G_M \quad (13)$$

$$\dot{R} = CA \quad (14)$$

$$\dot{M} = \dot{D} + E\dot{R} \quad (15)$$

$$e = \alpha e_M^* + (1-\alpha)e_X = \frac{E[\alpha_M^{P*} + (1-\alpha)P_X^*]}{P_N} \quad (16)$$

Equation 1 defines the total assets $A$ in domestic currency as a sum of foreign money and domestic money. Equation 2 defines real assets in terms of the exportable good. Equation 3 is the

portfolio composition equation. Equation 4 determines that capital mobility is inexistent and that no commercial transactions are subject to the financial state $\delta$ . Equation 5 to 9 summarises the demand and supply notion, $e_M$ and $e_X$ are the domestic prices of importables and exportables with respect to non-tradables. Equation 10 and 11 summarises the government sector, where $G_N$ and $G_M$ are consumption of $M$ and $N$ . Equation 12 is the government budget constraint. Equation 13 to 16 summarises the external sector. The attainment of a sustainable long-run equilibrium occurs when the non-tradable good market and the external sector (current account and balance of payments) are concurrently in equilibrium.

### 2.3. Empirical Studies

#### 2.3.1. Real Exchange Rates and Macroeconomic Fundamentals

Ricci, Milesi-Ferretti and Lee (2008) estimated a panel cointegrating relationship between real exchange rates and macroeconomic fundamentals for forty-eight industrial countries and emerging markets. Improved measures for productivity differentials, external imbalances, and commodity terms of trade were employed and the results showed a robust positive association between the CPI-based real exchange rate and commodity terms of trade. Productivity growth differentials between traded and non-traded goods was small but statistically significant. The study placed emphasis on the significance of employing productivity data for both tradables and non-tradables with respect to trading partners to substitute for the Balassa-Samuelson effect.

Zhang (2001) estimated the behavioural equilibrium exchange rate and the resultant misalignment in China based on the theory of equilibrium real exchange rate. The Johansen cointegration method was employed in the estimation of the equilibrium real exchange rate and the resulting misalignment. Results revealed that a rise in investment and openness of the economy led to the depreciation of the real exchange rate depreciation. Whereas a rise in government expenditure and export led to the appreciation of the real exchange rate. Chronic overvaluation in China's central planning period is evident in the study; however economic reforms brought the real exchange rate closer to equilibrium. These results provided some evidence of China's proactive exchange rate policy which sought to employ the nominal exchange rate as a policy tool, as a means of attaining targets in the real sector or a real exchange rate target.

Mathisen (2003) studied the equilibrium real exchange rate in Malawi for the period of 1980 to 2002. The real exchange rate was presented as a function of fundamentals drawn from economic theory. Edward's theoretical model was adopted to model the relationship between real exchange rates and fundamentals in Malawi. Fundamentals cited included: government consumption, investment, terms of trade and technological progress amongst others. The results favoured the equilibrium approach to real exchange rate determination.

Miyajima (2007) evaluated the competitiveness in Namibia. The findings revealed a real effective exchange rate in equilibrium. Additionally, suggestions on improving competitiveness

of the country were stated in the study. Exchange rate misalignment was derived, and findings showed that Namibia experienced great misalignments in 1990 which weakened in the 1990s and increased in the 2000s.

Ozsoz and Akinkunmi (2012) assessed the determinants of real exchange rates for the Nigerian Naira. The study proposed that oil prices, broad money supply, level of foreign reserves and interest rate differentials with trading partners were possible predictors of the long-run Naira equilibrium real exchange rate. Furthermore, the study employed the behavioural equilibrium exchange rate approach as means of identifying misalignments in the real Naira rate. Findings of the study revealed an undervaluation of the Naira at the end of 2010. Mkenda (2001) investigated the main determinants of the real exchange rate in Zambia by employing cointegration analysis. The tested fundamentals (terms of trade, trade taxes etc.) were found to be influencers of the real exchange rate for exports in the long-run and the internal real exchange rate was influenced by terms of trade, investment share, and the rate of growth of real GDP in the long-run. Additionally, the study derived exchange rate misalignment and found that in some periods the exchange rates were overvalued.

Eita and Sichei (2014) studied the equilibrium real exchange rate for Namibia for the period 1998 to 2012 by means of quarterly data using the Vector Error Correction Model (VECM). The study found an increment in the ratio of investment to GDP and resource balance linked with a subsequent appreciation of the real exchange rate. The terms of trade resulted in the real exchange rate depreciation implying that the substitution effect dominated the income effect. The study further revealed periods of undervaluation and overvaluation of the real exchange meaning that real exchange rate misalignment occurred in some periods.

With the real exchange rate as a key policy variable in the South African open economy, Aron, Elbadawi and Kahn (1997) presented possibly the first formal definition and estimation of the fundamental (long-run) and short-run influences in a model for the real exchange rate in South Africa. They employed a single equation cointegration model to investigate the short-run and long-run equilibrium determinants of the quarterly real exchange rate in 1970:1 to 1995:1. The macroeconomic balance approach was used to define the equilibrium real exchange rate with focus on the concurrent realisation of internal and external balance for given sustainable values

of variables such as taxes, terms of trade, trade policy, capital flows and technology. The model employed in the study revealed that, over time, the real exchange rate is not constant but it too changes in an array of fundamentals and shocks to the economy.

### 2.3.2. Real Exchange Rate Misalignment and Economic Performance

Rodrik (2008) tested RER misalignment and growth in one hundred and eighty-four countries from 1950 to 2004 using an index as a measure of the degree of RER undervaluation. The Balassa-Samuelson effect was taken into consideration through the employment of real per capita GDP. The study found that overvaluation impedes growth while undervaluation promotes it. Abida (2011) explored real exchange rate misalignment and growth in Tunisia, Algeria and Morocco by using the Fundamental Equilibrium Exchange Rate (FEER) approach to derive misalignment. Findings of the study showed negative misalignment. Similarly, Elbadawi, Kaltani and Soto (2012) found that misalignment negatively affected growth in a study about aid, real exchange rate misalignment, and economic growth in Sub-Saharan Africa. Sallenave (2010) obtained similar findings in a study about the growth effects of real effective exchange rate misalignments for the G20 countries.

Vieira and MacDonald (2012) studied the effect of real exchange rate misalignment on long-run growth from 1980 to 2004 by approximating a panel data model for a set of ninety-nine countries. Measures of real exchange rate misalignment were created by using approximations of the equilibrium real exchange rate. The results revealed positive coefficients for real exchange rate misalignment meaning that a depreciated (appreciated) real exchange rate aided (impaired) long-run growth.

Ghura and Grennes (1993) studied the incidences of real exchange rate misalignment and the resultant impact on economic performance for thirty-three Sub-Saharan African countries over the time span of 1972 to 1987. The study found an inverse relationship between the real exchange rate (RER) misalignment and economic performance. Additionally, the study indicated that macroeconomic instability contributed to slower growth whilst high misalignment was similarly associated with high macroeconomic instability levels. On the one hand, low levels of RER misalignment and instability translated to improved economic performance.

Ndlela (2012) investigated the relationship between real gross domestic product growth and real exchange rate misalignment for Zimbabwe and the results concretised the hypothesis of real exchange rate overvaluation being a factor in contracting economic growth in Zimbabwe. Tsen Wong (2013) explored real exchange rate misalignment and economic growth in Malaysia and found that a rise in real exchange rate misalignment resulted in a fall in economic growth. Naseem and Hamizah (2013) also investigated real exchange rate misalignment and economic growth in Malaysia with the results indicating the presence of a positive and significant relationship between RER misalignment and economic growth.

### 2.3.3. Limitations of Reviewed Studies and Contribution to Literature

The reviewed studies revealed the pertinence of the real exchange rate in various economies around the globe. Numerous studies explored various fundamentals that potentially influence the behaviour of the real exchange rate, with the concern of the effects it may pose on economic performance. Some of these reviewed studies further derived real exchange rate misalignment which impacts the on economic performance of a country (Miyajima 2007, Ozsoz and Akinkunmi 2012, Mkenda 2001). However, many of these studies do not test the impact of misalignment on measures of economic performance. The limited studies by Eita and Sichei 2014, Ndlela 2012 and Mkenda 2001 were based on single countries and cannot be generalised to Africa. However, Ghura and Grennes (1993) did conduct a study on Sub-Sahara African countries where the previously popular three measures of real exchange rate misalignment were employed. These measures included a measure based on the Purchasing Power Parity (PPP) (also used by Balassa (1990) and Cottani *et al* 1990; a measure based on the official nominal exchange rates (also used by Edwards (1989) and Cottani *et al* 1990); and a black market nominal exchange rates measure (also used by Edwards 1989, 1990).

Some of these measures have limitations, for instance, the PPP theory is insufficient in explaining the equilibrium exchange rate because real exchange rates depart for long periods from their PPP levels (MacDonald and Ricci 2002 and Siregar 2011). Hossfeld (2010) affirmed the shortcoming of the PPP as a determinant of equilibrium exchange rate by stating that the PPP was unable to capture the role of capital flows and other fundamental determinants of real exchange rates.

Therefore, this study makes a contribution by using the most recent data to estimate the equilibrium real exchange rate and further tests the effects of real exchange rate misalignment on economic performance. The study constructs an equilibrium real exchange rate by substituting permanent values of fundamentals into the estimated co-integrating relationship. The estimated coefficients are imposed on the permanent values of the fundamentals. The permanent values of the fundamentals were constructed using the Hodrick Prescott (HP) filter.

In macroeconomics, time series are generally considered as the sum of transitory and permanent components. The HP filter helps capture the smooth path of the trend component by maximising the sum of the squares of its second difference (Choudhary, Hanif and Iqbal 2014). The HP is also advantageous because it is insensitive towards periods therefore, there is little arbitrariness; long-term trends fluctuate over time under the HP (Anaya 1999). The study then proceeds to compute real exchange rate misalignment as the percentage difference between the actual real exchange rate and the equilibrium RER as in Hinkle and Montiel (1999) who interpreted misalignment as the gap between the actual real exchange rate (e) and the equilibrium real exchange rate (e*) following Edwards (1989) and Hinkle and Montiel (1999).

The use of recent data helps capture current developments in the African region as most of the economies have undergone structural changes and exchange rate reforms. In the process, this will improve the robust estimation of the relationship between real exchange rate misalignment and economic performance. This will enable the assessment of the African economy in terms of growth, that is, whether growth is progressive or regressive to inform the formulation of suitable exchange rate policies that promote growth in the African region.

## 3. Methodology

This section presents the analytical tools employed to investigate the impact of macroeconomic fundamentals on the real exchange rate, and the resulting RER misalignment on economic performance in a panel of African countries (Algeria, Cameroon, Central African Republic, Equatorial Guinea, Gabon, Gambia, Ghana, Lesotho, Morocco, Nigeria, South Africa, Sierra Leone, Togo, Tunisia, Uganda and Zambia). The section is organised as follows: the first section outlines the estimation of the equilibrium real exchange rate and the resulting real exchange rate misalignment and economic performance.

### 3.1. Model Specification

The study employed an empirical model adapted from Edwards' (1988) fundamental approach to real exchange rate determination and variables considered as determinants of the exchange rate. In the model, Edwards expressed the equilibrium exchange rate as a function of certain fundamental factors.

Edwards specified his model as follows:

$$\log e_t^* = \beta_0 + \beta_1 \log(TOT)_t + \beta_2 \log(NGCGDP)_t + \beta_3 \log(TARIFFS)_t + \beta_4 \log(TECHPRO)_t + \beta_5 (KAPFLO)_t + \beta_6 \log(OTHER)_t + \mu_t$$

(17)

Where: *TOT* represents the *external terms of trade*, *NGCGDP* represents the *ratio of government consumption on non-tradables* to *GDP*, *TARIFFS* represent the proxy for the level of *import tariffs*, *TECHPRO* represents a *measure of technological progress*, *KAPFLO* represents *capital inflows* and outflows depending on whether the value is positive or negative, OTHER represents variables such as the *investment/GDP ratio* and $\mu$ , the *error term*.

Following the approach employed by Edwards (1988), the study adopted the following variations of Edward's model, due to data constraints, the study tests two variations of the model. The model is specified as follows:

1995-2016:

$$LRER_{i,t} = \alpha_0 + \alpha_1 LINFL_{i,t} + \alpha_2 LGOVEXP_{i,t} + \alpha_3 LTARIFFS_{i,t} + LINV_{i,t} + \varepsilon_{i,t}$$

(18)

Where the real exchange rate, $LRER$ is explained by *import tariffs* ($LTARIFFS$), *government expenditure* ($LGOVEXP$) and the *rate of inflation*, consumer prices as an annual percentage ($LINFL$), *total investment as a share of GDP* ($LINV$) and $\varepsilon$ is the *error term*. All variables are expressed in logarithms to reduce data variability. Cointegration was used to determine the short and long-run determinants of the equilibrium RER.

1990-2016:

$$LRER_{i,t} = \alpha_0 + \alpha_1 LGOVEXP_{i,t} + \alpha_2 LFDI_{i,t} + \alpha_3 LINFL_{i,t} + LTOT_{i,t} + \varepsilon_{i,t}$$

(19)

Where the real exchange rate, $LRER$ is explained by *terms of trade* ($LTOT$), *government expenditure* ($LGOVEXP$) and the *rate of inflation*, consumer prices as an annual percentage ($LINFL$), *foreign direct investment* ($LFDI$), and $\varepsilon$ is the *error term*. All variables are expressed in logarithms to reduce data variability. Cointegration was used to determine the short and long-run determinants of the equilibrium *RER.*

The Edwards' (1988) model is a good representation for African countries because it is a general equilibrium model for developing countries. It involves nominal and real factors in the short run and fundamentals that influence the equilibrium real exchange rate in the long-run. In addition, determinants of the equilibrium real exchange such as changes in tariffs, terms of trade, capital account liberalisation and government consumption are specified. Due to data constraints, the models in this study employed additional variables other than those specified in Edward (1988).

For the real exchange rate variable, the study used the real effective exchange rate (*REER*). It is the weighted average of a country's currency in relation to an index or basket of other major currencies. The *REER* takes into account the influences of inflation. The *REER* is calculated using a geometric average formula: $REER_t = \frac{NEER_t \times CPI_t}{CPI_t^{foreign}}$ where *NEER* is the nominal effective exchange rate, *CPI* is the weighted average of *CPI* indices of trading partners. An increase in *REER* is appreciation and a decrease is depreciation.

Terms of trade possibly has two effects on the real exchange rate, that is, the income and substitution effects. The income effect occurs when there is an increase in export prices or a decline in import prices which in turn increases the income in an economy and the demand for non-tradables. This decreases the relative prices of tradables to non-tradables and appreciates the Real Exchange Rate (*RER*). On the substitution effect, an improvement in *TOT* due to an increment in export prices results in a depreciation of *RER* for certain levels of nominal exchange rate and non-tradable prices.

Tariffs refer to import tariffs which are defined as the tax levied on imported goods and services. An import tariff leads to an improvement of the current account and an appreciation of the real

exchange rate (Edwards 1987). Ravn, Schmitt-Grohe and Uribe (2012) recorded that a rise in government purchases increased output and private consumption, weakened the trade balance and depreciated the real exchange rate.

An increase in inflation results in the depreciation of the real exchange rate while a reduction in inflation normally appreciates the real exchange rate. The relationship between *FDI* and the *RER* is ambiguous because the effect on the real exchange rate is dependent on the import content of the *FDI* (Rochester 2013). A rise in ratio of investment to *GDP* leads to increased spending and a decline in the current account therefore leading to the depreciation of the real exchange rate (Eita 2007).

### 3.2. Real Exchange Rate Misalignment

As discussed in previous sections, the study tests for real exchange rate misalignment and its subsequent impact on economic performance. Real exchange rate misalignment is the deviation of the actual real exchange rate from its long-run equilibrium value (Hinkle and Montiel 1999). After establishing the short and long-run determinants of *RER*, the *RER* misalignment is computed by subtracting the equilibrium real exchange rate from the actual real exchange rate.

$$Misalignment = RER_t - EREER_{t-1} \quad (20)$$

Where misalignment is the real exchange rate misalignment, $RER$ is the actual real exchange rate, and $EREER$ is the equilibrium real exchange rate. Positive results imply real exchange rate undervaluation while negative results imply real exchange rate overvaluation.

#### 3.2.1. Real Exchange Rate Misalignment and Economic Performance

After obtaining the RER misalignment indicator, the impact of misaligned RER on economic performance was assessed. Studies such as Ndlela 2012, Tsen Wong 2013, Naseem and Hamizah 2013 also applied the same technique. However, these previous studies did not specify the type of model used, while this study employed the Cobb Douglas function to test the effect of real exchange rate misalignment on economic performance.

The equations are expressed with the variables affecting economic performance drawn from the Cobb Douglas Function for both time periods (1995 to 2016) and (1990 to 2016). The model is specified as follows:

**Cobb-Douglas Function** with two factors, *capital* (*K*) and *labour* (*L*)

$$Y_t = A_t K_t^{\alpha} L_t^{\beta} \tag{21}$$

$L_t$ denotes the labour input, $K_t$ is the capital input, *A* is total factor productivity and *Y* is the gross domestic product

**Model 1:**

$$Y = \alpha_0 + \alpha_1 K_{i,t} + \alpha_2 L_{i,t} + \alpha_3 MISA_{i,t} + \varepsilon_{i,t} \tag{22}$$

Y denotes the *gross domestic product* (*GDP*); a measure of economic performance, *K* is *capital input* proxied by gross capital formation; *Labour* (*L*) denotes the *total labour force,* MISA denotes real exchange rate misalignment and ε is the error term.

Capital influences the gross domestic product positively; the more capital invested the higher the gross domestic product. There is a positive relationship between labour and economic growth as this implies greater productivity therefore a higher gross domestic product. Real exchange rate misalignment impacts negatively on economic growth; an increase in misalignment leads to a reduction in economic growth.

**Model 2:**

$$Y = \alpha_0 + \alpha_1 K_{i,t} + \alpha_2 L_{i,t} + \alpha_3 TOT_{i,t} + \alpha_4 MISA_{i,t} + \varepsilon_{i,t} \tag{23}$$

*Y* represents the *gross domestic product* (*GDP*); a measure of economic performance, *K* is *capital input* proxied by *gross capital formation*, Labour (*L*) denotes the *total labour force*, *TOT* represents terms of trade, *MISA* represents *real exchange rate misalignment* and ε is the *error term*.

There is a positive relationship between capital and the gross domestic product; labour and economic growth; terms of trade and economic growth. A percentage increase in any of these variables results in higher economic growth. Whilst real exchange rate misalignment impacts

negatively on economic growth; an increase in misalignment leads to a reduction in economic growth.

## 4. Data Description

Two models and two sample periods with annual data (1995 to 2016 and1990 to 2016) were employed due to data constrains for some of the variables in the models. The African countries are Algeria, Cameroon, Central African Republic, Equatorial Guinea, Gabon, Gambia, Ghana, Lesotho, Morocco, Nigeria, South Africa, Sierra Leone, Togo, Tunisia, Uganda and Zambia and they were selected based on data availability. Data was sourced from Quantec which gives access to organised and updated economic data. Quantec is a consultancy providing economic and financial data, country intelligence and quantitative analytical software with data from sources such as the International Monetary Fund, the World Bank and Central Banks of individual countries. The variables are transformed into logarithms to reduce their variability.

### 4.1. Estimation Technique

Panel estimation techniques were used because of its advantages over cross-sectional and time series data for a large data set. It can control heterogeneity among individual countries, minimise collinearity and allows for more degrees of freedom thus eliminating any biasness from aggregation.

To capture long-run effects of variables with homogeneous coefficients, the pooled mean group estimator (PMG) is applied. Traditional models such as ordinary least squares, fixed effects, random effects fail to capture this relationship well, hence the PMG estimator. Asteriou and Hall (2007) asserts the appropriateness of the PMG as means of avoiding spurious regressions resulting from the trends and unit roots of present in most macroeconomic data.

The pooled mean group estimator (PMG) accounts for both pooling and averaging. Intercepts, short-run coefficients, and error variances fluctuate across groups and maintain the long-run coefficients (Pesaran, Shin and Smith 1999).This technique produces consistent and asymptotically normal estimates of the long-run coefficients regardless of the order of integration of underlying regressors, that is, whether I(1) or I(0) (Pesaran, Shin and Smith 1999).It is also advantageous because it permits heterogeneous short-run dynamics per cross section and they are country specific (Iheonu, Ihedimma and Omenihu 2017).

The study also employed the Kao test to test for cointegration to determine the presence of a long-run relationship between the equilibrium real exchange rate and the fundamentals. Furthermore, the Dynamic Ordinary Least Squares estimator (DOLS) proposed by Stock and Watson (1993) was applied.

### 4.1.1. Panel Unit Root Tests

The data was first subjected to the unit root test to test for the stationarity of the data. There are numerous panel unit root techniques available for such an assessment. The study employed the Levin, Lin and Chu Test (LLC Test) and the Im, Pesaran and Shin test.

### 4.1.2. Levin, Lin and Chu Test (LLC Test)

The LLC test was developed due to the power restrictions of singular unit root tests against alternative hypotheses containing continual shifts from equilibrium. This was more evident in small samples thus the creation of the LLC which proposed an improved panel unit root test which did not test individual unit root tests per cross-section. The null hypothesis is that each individual time series contains a unit root and the alternative hypothesis is each time series is stationary (Baltagi 2008). The hypothesis is presented as follows:

$$\Delta y_{it} = \rho\, y_{i,t-1} + \sum_{L=1}^{\rho'} \theta_{iL}\, \Delta y_{it-L} + \alpha_{mi} d_{mt} + \varepsilon_{it} \quad m = 1,2,3 \tag{24}$$

The LLC test proceeds in the following manner:

**Step 1:** Performance of individual augmented Dickey-Fuller (ADF) regressions per cross-section:

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{\rho'} \theta_{iL}\, \Delta y_{i,t-L} + \alpha_{mi} d_{mt} + \varepsilon_{it} \quad m = 1,2,3$$

(25)

The lag order $p_i$ is permitted to vary across individuals.

**Step 2:** Estimation of the ratio of long-run to short-run standard deviations under the null hypothesis**:**

$$\hat{\sigma}_{yi}^2 = \frac{1}{T-1}\sum_{t=2}^{t}\Delta y_{it}^2 + 2\sum_{L=1}^{\tilde{K}}\omega_{\tilde{K}L}\left|\frac{1}{T-1}\sum_{t=2+L}^{T}\Delta_{y_{it}}\Delta_{y_{i,t-L}}\right|$$

(26)

**Where** $\tilde{K}$ is a truncation lag that can be data dependent.

**Step 3:** Computation of the panel test statistics and running of the pooled regression:

$$\tilde{e}_{it} = \rho\tilde{v}_{i,t-1} + \tilde{\varepsilon}_{it}$$

(27)

Based on $N\tilde{T}$ observations where $\tilde{T} = T - \bar{p} - 1$. $\tilde{T}$ represents the number of observations per individual in the panel with $\bar{p} = \sum_{I=1}^{N} p_i / N$. $\bar{p}$ is the average lag order of individual ADF regressions (Baltagi 2008).

### 4.1.3. Im, Pearson and Shin Test (IPS)

The Im, Pearson and Shin Test (IPS) improved on the LLC test because it permitted a heterogeneous coefficient of $y_{it-1}$ and suggested a different testing technique centred on averaging individual unit root test statistics. It differs from the LLC which requires $\rho$ to be homogeneous across $i$.

IPS proposed that an average of the ADF tests when $u_{it}$ is serially correlated with different serial correlation properties across cross-sectional units. The null hypothesis is that each series in the panel contains a unit root, i.e. $H_0 : \rho_i = 0$ for all $i$ and the alternative hypothesis permits some of the individual series to have unit roots (Baltagi 2008).

$$\begin{aligned} H_1 : &\{\rho_i < 0 \ for \quad i = 1,2,...N_1 \\ &\{\rho_i = 0 \ for \quad i = N_1 + 1,..., N \end{aligned} \tag{28}$$

### 4.2. Test for Cointegration

After determining the stationarity of the variables, Kao's (1999) cointegration test was undertaken to determine the presence of a long-run equilibrium relationship among the variables. This method was chosen because it accounts for spurious regression of panel data and employs two types of panel cointegration tests. It makes use of the Dickey-Fuller (DF) and augmented Dickey-Fuller (ADF) tests. Moreover, the sequential limit theory of Phillips and Moon (1999) which argued for sequential limits being essential in obtaining asymptotic distributions is used.

**4.2.1. Kao Test for Cointegration**

Kao (1999) showed DF and ADF type tests for cointegration in panel data. For a model $Y_{it} = a_i + \beta X_{it} + \hat{u}_{it}$.

(29)

The residual based cointegration is used in the equation $\hat{u}_{it} = e\hat{u}_{it-1} + v_{it}$.

(30)

Under the Kao test, the following ADF test regression is run:

$$u_{i,t} = \rho u_{i,t-1} + \sum_{j=1}^{n} \phi_{j\Delta u_{i,t-j} + v_{it}}$$

(31)

The ADF statistic of the null hypothesis of no cointegration is expressed calculated by the following formula:

$$ADF = \frac{t_{ADF} + \sqrt{6N\hat{\sigma}_V} / (2\hat{\sigma}_{0v})}{\sqrt{\hat{\sigma}_{0v}^2 / (2\hat{\sigma}_{"v}^2) + 3\hat{\sigma}_v^2 / (10\hat{\sigma}_{0v}^2}}$$

(32)

The ADF test statistic is represented by $t_{ADF}$ and it is found in the equation above (Asteriou and Hall 2016).

Once cointegration amongst variables was established, the dynamic OLS approach was employed to estimate the long-run RER model. The dynamic OLS (DOLS) was used based on its

performance on bias reduction in finite sample, homogenous and heterogeneous (Kao and Chiang 2000).

### 4.3. The Dynamic OLS approach

The Dynamic OLS approach developed by Stock and Watson (1993) considers past, present and future values of the change in $X_t$:

$$C_t = B^{'} X_t + \_ \sum_{j=-J}^{j=J} \eta_j \Delta P_{t-j} + \sum_{j=-K}^{j=K} \lambda_j \Delta Y_{t-j} + \varsigma_t$$

(33)

Fundamentally, the DOLS procedure regresses 1(1) variables on other I(1) variables, I(0) variables and leads and lags of the first differences of I(1) variables (Masih and Masih 1996).This method takes into consideration efficient estimators of co-integrating vectors including deterministic components. In addition, different orders of integration and potential concurrence between variables is considered. Leads and lags of different variables in the equation containing a co-integrating vector eradicates the bias of concurrence and the small sample bias (Irffi *et al* 2008).

## 5. Estimation Results

### 5.1. Stationarity Tests

Stationarity test results are presented in Tables 1 and 2:

**Table 1**: Unit Root Test - Model 1 (1995 to 2016)

| **Variable** | **LLC Test** | | **IPS Test** | |
|---|---|---|---|---|
| | **Levels** | | **Levels** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LRER** | -1.42897<br>(0.0765) | -0.63679<br>(0.2621) | -0.49669<br>(0.3097) | -0.58972<br>(0.2777) |
| **LINFL** | -11.8844<br>(0.0000)* | -10.7254<br>(0.0000)* | -12.4162<br>(0.0000)* | -11.0833<br>(0.0000)* |

| | | | | |
|---|---|---|---|---|
| **LGOVEXP** | -0.64036<br>(0.2610) | -4.03337<br>(0.0000)* | -0.57794<br>(0.2817) | -2.51781<br>(0.0059)* |
| **LTARIFFS** | -0.98765<br>(0.1617) | 2.41675<br>(0.9922) | 2.72052<br>(0.9967) | 1.52334<br>(0.9362) |
| **LINV** | -11.8844<br>(0.0000)* | -10.7254<br>(0.0000)* | -12.4162<br>(0.0000)* | -11.0833<br>(0.0000)* |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LRER** | -13.4907<br>(0.0000)* | -10.3966<br>(0.0000)* | -11.1666<br>(0.0000)* | -7.95154<br>(0.0000)* |
| **LINFL** | -20.0685<br>(0.0000)* | -16.7294<br>(0.0000)* | -22.0215<br>(0.0000)* | -19.1432<br>(0.0000)* |
| **LGOVEXP** | -11.3570<br>(0.0000)* | -9.06609<br>(0.0000)* | -12.7766<br>(0.0000)* | -10.3254<br>(0.0000)* |
| **LTARIFFS** | -12.7728<br>(0.0000)* | -9.18791<br>(0.0000)* | -11.3018<br>(0.0000)* | -8.03328<br>(0.0000)* |
| **LINV** | -20.0685<br>(0.0000)* | -16.7294<br>(0.0000)* | -22.0215<br>(0.0000)* | -19.1432<br>(0.0000)* |

**p-values* are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 5% level of significance

**Table 2:** Unit Root Test - Model 2 (1990 to 2016)

| **Variable** | **LLC Test** | | **IPS Test** | |
|---|---|---|---|---|
| | **Levels** | | **Levels** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LRER** | -3.54621<br>(0.0002)* | -4.32355<br>(0.0000)* | -3.47528<br>(0.0003)* | -3.87624<br>(0.0001)* |
| **LGOVEXP** | -5.51397<br>(0.0000)* | -4.18105<br>(0.0000)* | -2.83844<br>(0.0023)* | -2.07515<br>(0.0190)* |
| **LFDI** | -3.35397<br>(0.0004)* | -3.96965<br>(0.0000)* | -2.62995<br>(0.0043)* | -7.39800<br>(0.0000)* |
| **LINFL** | -6.92330<br>(0.0000)* | 2.47077<br>(0.9933) | -1.02864<br>(0.1518) | -1.18551<br>(0.1179) |
| **LTOT** | -1.25129 | -0.06803 | -0.38659 | -1.25918 |

| | | | | |
|---|---|---|---|---|
| | (0.1054) | (0.6316) | (0.3495) | (0.1040) |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LRER** | -18.5659 (0.0000)* | -16.4115 (0.0000)* | -17.1066 (0.0000)* | -14.8175 (0.0000)* |
| **LGOVEXP** | -11.6682 (0.0000)* | -9.76634 (0.0000)* | -12.9793 (0.0000)* | -11.4472 (0.0000)* |
| **LFDI** | -19.4604 (0.0000)* | -15.4665 (0.0000)* | -22.2376 (0.0000)* | -20.0447 (0.0000) |
| **LINFL** | -9.19896 (0.0000)* | -6.26249 (0.0000)* | -8.80218 (0.0000)* | -7.87933 (0.0000)* |
| **LTOT** | -15.0798 (0.0000)* | -12.2396 (0.0000)* | -15.0580 (0.0000)* | -12.6015 (0.0000)* |

**p-values* are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 5% level of significance

Tables 1 and 2 are a display of the LLC and the IPS panel unit root test results at levels and first difference for models 1 and 2.

The LLC and IPS test unit root tests coincide and they reveal that in model 1, variables *LINFL*, *LGOVEXP* and *LINV* are stationary at levels, they are integrated of order zero I(0). While *LRER* and *LTARIFFS* become stationary at first difference, therefore they are integrated of order one I(1).

In model 2, all variables except (*LTOT*) are stationary at levels, they are integrated of order zero I(0). *LTOT* becomes stationary at first difference, therefore it is integrated of order one I(1).

### 5.2. Estimation of the Real Exchange Rate Cointegration Results

Table 3 reports the results of Kao's residual panel cointegration tests, which rejected the null hypothesis of no cointegration because the *p-value* is less than 5%, therefore, there is a cointegration relationship amongst the variables.

**Table 3:** Kao Cointegration Test Results for Model 1 and Model 2

| Model 1 | t-statistic | Probability |
|---|---|---|
| **ADF** | -2.556528 | 0.0053* |

| Model 2 | t-statistic | Probability |
|---|---|---|
| **ADF** | -3.585961 | 0.0002* |

**NB:** The ADF is the residual-based ADF statistic. The null hypothesis is no cointegration.

* Indicates that the estimated parameters are significant at the 5% level

The conclusion therefore, is that there is a panel long-run equilibrium relationship among the real exchange rate, terms of trade, inflation, import tariffs and government expenditure in the long-run.

### 5.3. Long-run coefficient - Dynamic OLS Estimates (DOLS)

The results exhibit the presence of a cointegration relationship amongst the variables therefore the dynamic OLS approach is employed to estimate the long-run *RER* model and the results are presented in Table 4.

**Table 4:** DOLS long-run estimation results. Dependent variable: LRER

| Sample Period | 1995-2016 | 1990-2016 |
|---|---|---|
| **Explanatory Variables** | **Model 1 Coefficients** | **Model 2 coefficients** |
| **LINFL** | -0.040711<br>(0.2051) | -0.140248<br>(0.0665)* |
| **LGOVEXP** | -0.153702<br>(0.0013)** | -0.013619<br>(0.7523) |
| **LTARIFFS** | 0.219374<br>(0.0001)** | - |
| **LINV** | -0.283052<br>(0.0002)** | - |
| **LFDI** | - | -0.023019<br>(0.3110) |
| **LTOT** | - | 0.149011<br>(0.0110)** |
| **R-squared**<br>**S.E. of regression** | 0.773453<br>0.161295 | 0.811598<br>0.150437 |

**p-values* are in parentheses ()

*10 % statistically significant.

**5 % statistically significant.
***1 % statistically significant.

Table 4 presents the long-run coefficients results of the DOLS estimator. The results reveal that inflation is statistically significant and consistent with economic theory in model 2. Inflation complies with theory in model 1, however, the coefficient is not statistically significant. Government expenditure is statistically significant in model 1, while investment as a share of GDP, tariffs and terms of trade from model 1 are consistent with economic theory and are statistically significant.

A 1% increase in inflation would depreciate the real exchange rate by 0.14% thereby indicating a negative relationship between the two variables as stipulated by economic theory. 1% increase in government expenditure would depreciate the real exchange rate by 0.15% while a 1% increase in tariffs would lead to an appreciation of the real exchange rate by 0.21%. The relationship between terms of trade and the real exchange rate is positive and statistically significant. 1% increase in the terms of trade appreciates the real exchange rate by 0.14%.

Each of the models have an $R^2$ greater than 70%. This implies that the models are generally a good fit as more than 70% of the variations of the dependent variable is explained by the independent variables.

After establishing the long-run relationship amongst the fundamentals, real exchange rate misalignment was computed.

### 5.4. Computed Real Exchange Rate Misalignment

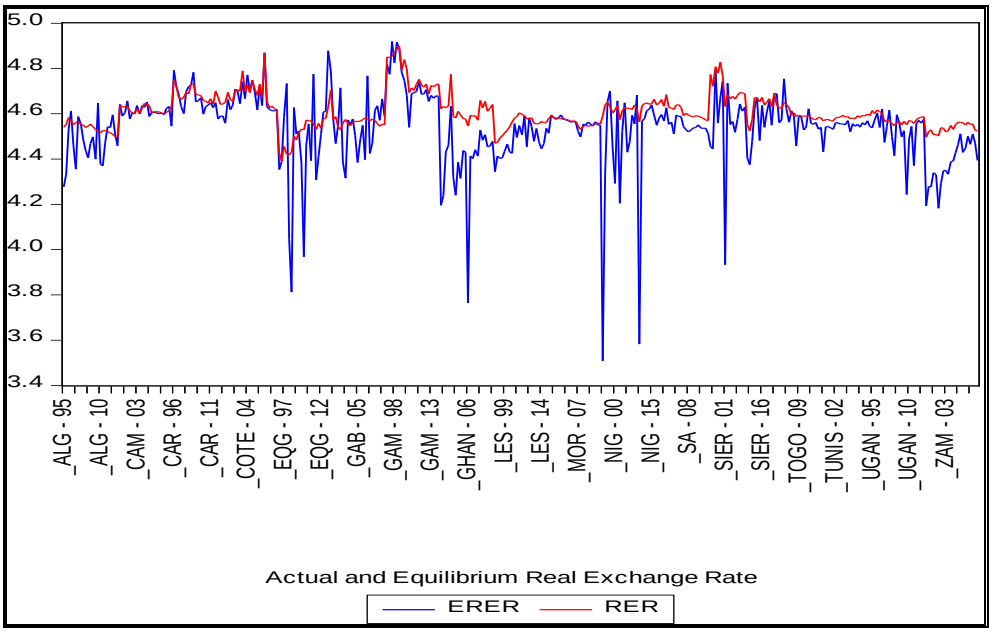



Figure 1: Actual and Equilibrium Real Exchange Rate (Model 1: 1995-2016)

*ERER is the equilibrium real exchange rate and RER is the actual real exchange rate

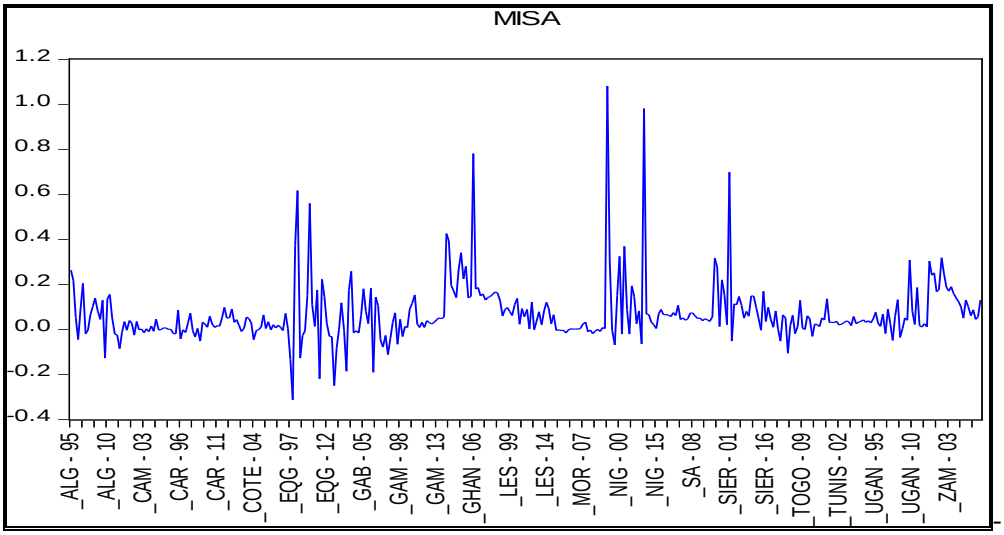



Figure 2 Real Exchange Rate Misalignment - Model 1(1995-2016)

*MISA denotes real exchange rate misalignment

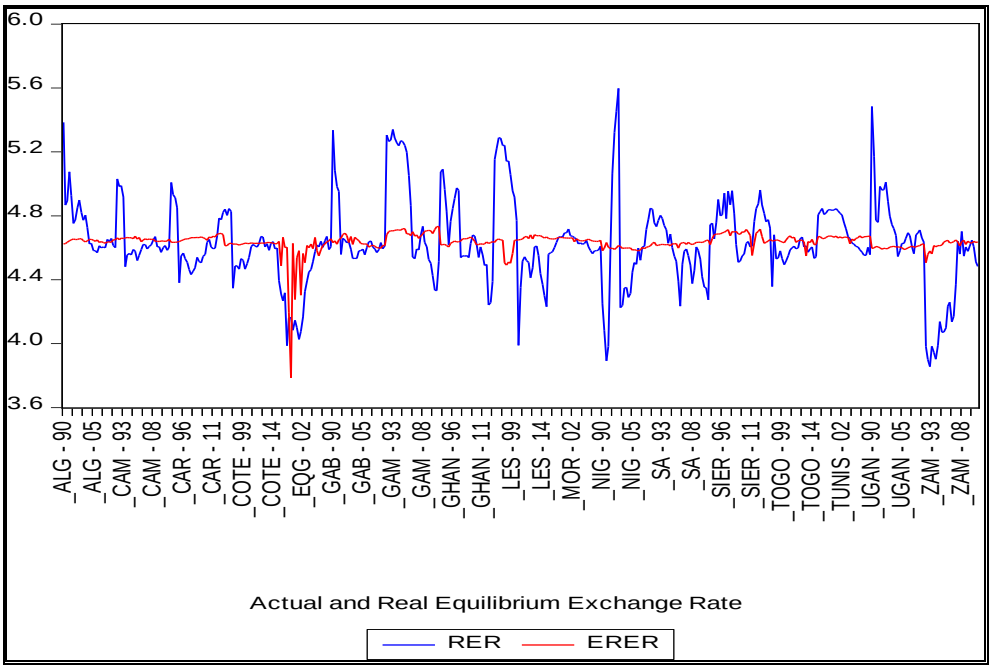



Figure 3: Actual and Real Exchange Rate (Model 2: 1990-2016)

*ERER is the equilibrium real exchange rate and RER is the actual real exchange rate

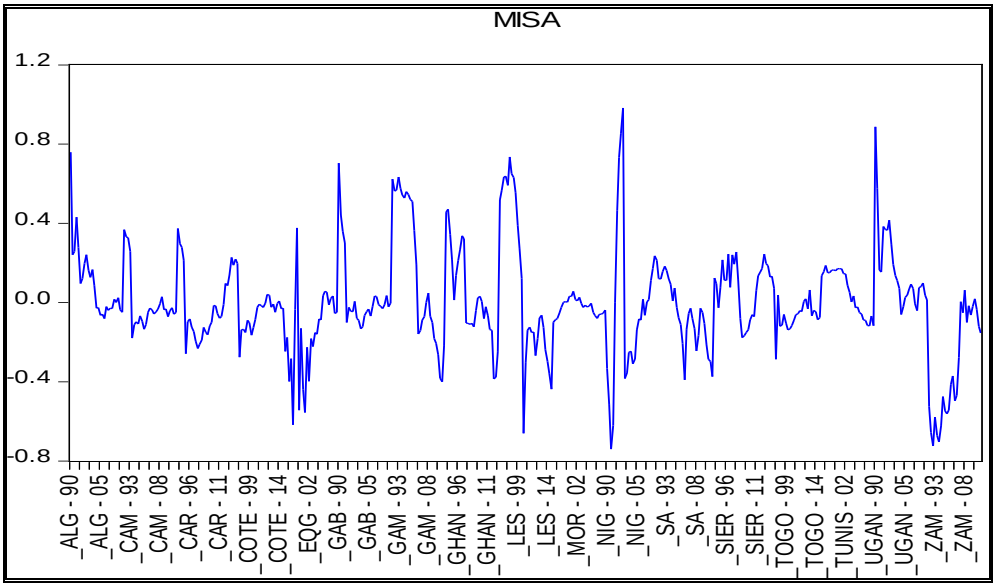



Figure 4 Real Exchange Rate Misalignment - Model 2 (1990-2016)

*MISA denotes real exchange rate misalignment

Figures 1, 2, 3 and 4 display the *RER* misalignment levels for models 1 and 2, time periods 1990 to 2016 and 1995 to 2016.Overall, there were more periods of overvaluation than undervaluation. These results are similar to Ali *et al* (2015) who found more periods of overvaluation than

undervaluation in Nigeria. They attributed these findings to the possible exchange rate policies that halted undervaluation episodes during the estimation period.

## 6. Real Exchange Rate Misalignment and Macroeconomic Performance

### 6.1. Test for Stationarity

The table presents results of the unit roots test conducted on the model specified in section 2.3.2.1. The variables were subjected to the LLC and the IPS stationarity tests. The results are presented in Tables 5 and 6 below:

**Table 5:** Unit Root Test - Model 1 (1995 to 2016)

| **Variable** | **LLC Test** | | **IPS Test** | |
|---|---|---|---|---|
| | **Levels** | | **Levels** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LGDP** | -0.55879<br>(0.2882) | 0.66099<br>(0.74570) | 4.09194<br>(1.0000) | 2.00461<br>(0.9775) |
| **LCAP** | -11.8844<br>(0.0000)* | -10.7254<br>(0.0000)* | -12.4162<br>(0.0000)* | -11.0833<br>(0.0000)* |
| **LLAB** | -2.23151<br>(0.0128)* | -3.26076<br>(0.0006)* | 3.49022<br>(0.9998) | -0.56796<br>(0.2850) |
| **MISA** | -5.58002<br>(0.0000)* | -4.52281<br>(0.0000)* | -7.76501<br>(0.0000)* | -6.60874<br>(0.0000)* |
| **Variable** | **First Difference** | | **First Difference** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |

| | | | | |
|---|---|---|---|---|
| **LGDP** | -4.94848<br>(0.0000)* | -7.83556<br>(0.0000)* | -5.28434<br>(0.0000)* | -5.47200<br>(0.0000)* |
| **LCAP** | -20.0685<br>(0.0000)* | -16.7294<br>(0.0000)* | -22.0215<br>(0.0000)* | -19.1432<br>(0.0000)* |
| **LLAB** | -3.26425<br>(0.0005)* | -0.08344<br>(0.4668) | -2.70014<br>(0.0035)* | 1.43395<br>(0.9242) |
| **MISA** | -15.9539<br>(0.0000)* | -13.0606<br>(0.0000)* | -17.7796<br>(0.0000)* | -15.1994<br>(0.0000)* |

**p-values* are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 5% level of significance

**Table 6**: Unit Root Test - Model 2 (1990 to 2016)

| **Variable** | **LLC Test** | | **IPS Test** | |
|---|---|---|---|---|
| | **Levels** | | **Levels** | |
| | **Constant** | **Constant and Trend** | **Constant** | **Constant and Trend** |
| **LGDP** | 0.24447<br>(0.5966) | -0.55213<br>(0.2904) | 5.79912<br>(1.0000) | -0.03212<br>(0.4872) |
| **LCAP** | -1.09888<br>(0.1359)* | -0.80533<br>(0.2103)* | -1.01012<br>(0.1562)* | -1.47663<br>(0.0699)* |
| **LLAB** | 0.84785<br>(0.8017)* | -6.49238<br>(0.0000)* | 3.57978<br>(0.9998) | -2.23595<br>(0.0127) |
| **LTOT** | -1.25129<br>(0.1054) | -0.06803<br>(0.6316) | -0.38659<br>(0.3495) | -1.25918<br>(0.1040) |
| **MISA** | -1.70119<br>(0.0445)* | -2.17400<br>(0.0149)* | -2.65832<br>(0.0039)* | -2.71383<br>(0.0033)* |
| **Variable** | **First Difference** | | **First Difference** | |

| | Constant | Constant and Trend | Constant | Constant and Trend |
|---|---|---|---|---|
| **LGDP** | -6.43737 (0.0000)* | -5.68072 (0.0000)* | -6.59037 (0.0000)* | -4.27152 (0.0000)* |
| **LCAP** | -10.5465 (0.0000)* | -8.36146 (0.0000)* | -13.1706 (0.0000)* | -11.0200 (0.0000)* |
| **LLAB** | -2.00272 (0.0226)* | 2.01594 (0.9781) | -1.29763 (0.0972)* | 1.02507 (0.8473) |
| **LTOT** | -15.0798 (0.0000)* | -12.2396 (0.0000)* | -15.0580 (0.0000)* | -12.6015 (0.0000)* |
| **MISA** | -12.4163 (0.0000)* | -11.0808 (0.0000)* | -12.3524 (0.0000)* | -10.6043 (0.0000)* |

**p-values* are in parentheses ()

* indicates rejection of the null hypothesis of unit root at 5% level of significance

Tables 5 and 6 depict the LLC and the IPS panel unit root test results at levels and first difference for models 1 and 2. For model 1 first difference, all variables (*LGDP*, *LLAB*, *LCAP* and *MISA*) are I(1) both at the constant trend of the panel unit root regression, while *LGDP* is I(0). These results therefore result in the rejection of the null hypothesis of a panel unit root.

### 6.2. Real Exchange Rate Misalignment and Macroeconomic Performance Cointegration Results

Table 7 reports the results of Kao's residual panel cointegration tests, which reject the null hypothesis of no cointegration because the *p-value* is less than 5%, therefore there is a cointegration relationship amongst the variables.

**Table 7:** Kao Cointegration Test Results for Model 1 and Model 2

| Model 1 | t-statistic | Probability |
|---|---|---|
| **ADF** | -3.452817 | 0.0003* |
| **Model 2** | **t-statistic** | **Probability** |
| **ADF** | -5.493989 | 0.0000* |

**NB:** The ADF is the residual-based ADF statistic. The null hypothesis is no cointegration. * Indicates that the estimated parameters are significant at the 5% level

The conclusion therefore, is that there is a panel long-run equilibrium relationship among the gross domestic product, capital, labour and misalignment in the long-run.

### 6.3. Pooled Mean Group (PMG) Estimates

**Table 8:** PMG Results – Model

**Dependent Variable:** GDP

| **Long-run Coefficients** | |
|---|---|
| **CAP** | 0.285056<br>(0.0012)* |
| **LAB** | 1.890046<br>(0.0000)* |
| **MISA** | -2.439693<br>(0.0001)* |
| **Short-run Coefficients** | |

| | |
|---|---|
| **ΔCAP** | 0.038257<br>(0.0514*) |
| **ΔLAB** | -3.134675<br>(0.1714) |
| **ΔMISA** | -0.017637<br>(0.7028) |
| **Constant** | 0.033032<br>(0.7825) |
| **Error Correction Coefficient** | -0.069643<br>(0.0017)* |

**Table 9:** PMG Results – Model 2

**Dependent Variable:** GDP

| **Long-run Coefficients** | |
|---|---|
| **CAP** | -0.022233<br>(0.5395) |
| **LAB** | 1.148974<br>(0.0000)* |
| **TOT** | 0.963529<br>(0.0000) |
| **MISA** | 0.096344<br>(0.1115) |

| Short-run Coefficients | |
|---|---|
| **ΔCAP** | 0.059628<br>(0.0031)* |
| **ΔLAB** | -4.575919<br>(0.0862) |
| **ΔTOT** | -0.014487<br>(0.6086) |
| **ΔMISA** | 0.004427<br>(0.8276) |
| **Constant** | 0.033032<br>(0.7825) |
| **Error Correction Coefficient** | -0.073185<br>(0.0357)* |

Table 8 displays the PMG estimation results for model 1, period 1995 to 2016. In the long-run, capital and labour influence growth positively. A one percent growth in capital leads to 0.29 percent increase in growth. This positive influence of capital also resonates in the short-run. One percent growth in labour leads to 1.9 percent increase in growth.

Like Abida (2011), model 1 experienced an undervalued currency which is a condition that encourages growth in an economy. A depreciated real exchange rate encourages economic growth. While an appreciated real exchange rate is not favourable as it tends to harm long-run growth; these results are like those of Vieira and MacDonald (2012). The RER misalignment coefficient for model 2 is positive implying an overvaluation which promotes economic growth.

## 7. Conclusion

The study aimed to estimate the equilibrium real exchange rate, construct measures of real exchange rate misalignment and test the effects of real exchange rate misalignment on economic performance for a selection of African countries. Two models with different time periods were estimated, from 1995 to 2016 and 1990 to 2016. The real exchange rate model and the real exchange rate misalignment and macroeconomic performance model were also estimated.

Cointegration tests were undertaken to determine the presence of a long-run equilibrium relationship among the variables. Upon establishing cointegration amongst the different variables, long-run RER model was performed using the DOLS. RER misalignment results revealed more periods of overvaluation than undervaluation for models 1 and 2, time periods 1995 to 2016 and 1990 to 2016. The findings of the study are similar to the study on Nigeria by

Ali *et al* (2015) who found more periods of overvaluation than undervaluation which were attributed to the possibility of exchange rate policies halting undervaluation episodes during the estimation period. According to literature, overvaluation of the real exchange rate affects economic growth negatively, especially for developing countries. It may result in a tightened monetary policy thereby leading to a recession, discrimination against exports and capital flight amongst other economic issues.

Based on the results, the study advocates for the selected countries, most of which are classified as developing countries to adopt economic policies that promote competitive real exchange rates; and avoid sustained real exchange rate appreciation where the actual real exchange rate (RER) differs significantly from its long-run equilibrium value as stated by Hinkle and Montiel (1999).

**References**

Abida, Z. (2011). Real exchange rate misalignment and economic growth: An empirical study for the Maghreb countries. *Zagreb International Review of Economics & Business*, 14(2), 87-105.

Abuaf, N., & Jorion, P. (1990). Purchasing power parity. *Journal of Finance*, 45(1), 157-74.

Ali, A.I., Ajibola, I.O., Omotosho, B.S., Adetoba, O.O., & Adeleke, A.O. (2015). Real exchange rate misalignment and economic growth in Nigeria. *CBN Journal of Applied Statistics*, 6(2), 103-131.

Anaya, J.A.G. (1999). *Labor market flexibility in 13 Latin American countries and the United States*. Retrieved from https://elibrary.worldbank.org

Aron, J., Elbadawi, I., & Kahn, B. (1997). Determinants of the real exchange rate in South Africa.

Retrieved from https://www.researchgate.net

Asteriou, D., & Hall, S. (2007). *Applied econometrics: A modern approach*. New York, NY: Palgrave Macmillan.

Asteriou, D., & Hall, S.G. (2016). *Applied econometrics*. London: Palgrave.

Bacchetta, P., Van Wincoop, E., & Beutler, T. (2009). Predictability. *Journal of International Money and Finance*, 28(3), 406-426.

Baltagi, B. (2008). *Econometric analysis of panel data*. New York, NY: John Wiley & Sons.

Cencini, A. (2005). *Macroeconomic foundations of macroeconomics*. London: Routledge.

Choudhary, M.A., Hanif, M.N., & Iqbal, J. (2014). On smoothing macroeconomic time series using the modified HP filter. *Applied Economics*, *46*(19), 2205-2214.

Cottani, J.A., Cavallo, D.F., & Khan, M.S. (1990). Real exchange rate behavior and economic performance in LDCs. *Economic Development and Cultural Change*, 39(1), 61-76.

Devereux, M.B. (1997). Real exchange rates and macroeconomics: Evidence and theory. *Canadian Journal of Economics*, 30(4a), 773-808.

Edwards, S. (1987). Tariffs, terms of trade, and the real exchange rate in an intertemporal optimizing model of the current account. *Economica*, 56(223), 343-358.

Edwards, S. (1988). Real and monetary determinants of real exchange rate behavior: Theory and evidence from developing countries. *Journal of Development Economics*, 29(3), 311-341.

Edwards, S., & Savastano, M.A. (1999). *Exchange rates in emerging economies: What do we know? What do we need to know?* (No. w7228). Retrieved from www.nber.org

Eita, J.H. (2007). *Estimating the equilibrium real exchange rate and misalignment for Namibia*.

(Doctoral dissertation, The University of Pretoria, Pretoria, South Africa). Retrieved from https://repository.up.ac.za

Eita, J.H., & Sichei, M.M. (2014). Estimating Namibia's equilibrium real exchange rate. *The International Business & Economics Research Journal*, 13(3), 561.

Elbadawi, I.A., & Soto, R. (1997). Real exchange rates and macroeconomic adjustments in Sub-Saharan Africa and other developing countries. *Journal of African Economies*, 6(3), 74-120.

Elbadawi, I.A., Kaltani, L., & Soto, R. (2012). Aid, real exchange rate misalignment, and economic growth in Sub-Saharan Africa. *World Development*, 40(4), 681-700.

Engel, C., & West, K.D. (2005). Exchange rates and fundamentals. *Journal of Political Economy*,
113(3), 485-517.

Evrensel, A. (2013). *International finance for dummies*. New Jersey: John Wiley & Sons.

Ghura, D., & Grennes, T.J. (1993). The real exchange rate and macroeconomic performance in Sub-Saharan Africa. *Journal of Development Economics*, 42(1), 155-174.

Hinkle, L.E., & Montiel, P.J. (1999). *Exchange rate misalignment: Concepts and measurement for*
*developing countries*. New York, NY: Oxford University Press.

Hossfeld, O. (2010). *Equilibrium real effective exchange rates and real exchange rate misalignments: Time series vs. panel estimates* (No. 65). Retrieved from https://ideas.repec.org

Iheonu, C.O., Ihedimma, G.I., & Omenihu, M.C. (2017). A pooled mean group estimation of capital inflow and growth in sub Saharan Africa. *Romanian Economic Journal*, 20(65).

Irffi, G., Castelar, I., Siqueira, M., & Linhares, F. (2008). *Dynamic OLS and regime switching models to forecast the demand for electricity in the northeast of Brazil*. Retrieved from http://epge.fgv.br/finrio/myreview/FILES/CR2/p44.Pdf

Isard, P. (1978). *Exchange-rate determination: a survey of popular views and recent models. international finance section*. Retrieved from https://www.princeton.edu

Kao, C., & Chiang, M.H. (2001). On the estimation and inference of a cointegrated regression in panel data. In B.H. Baltagi (Ed.), *Nonstationary Panels, Panel Cointegration, And Dynamic Panels* (pp. 179-222). Emerald Group Publishing Limited.

Kao, C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, 90(1), 1-44.

Lothian, J.R., & Taylor, M.P. (1997). Real exchange rate behavior. *Journal of International Money and Finance*, 16(6), 945-954.

MacDonald, R., & Ricci, L.A. (2002). *Purchasing power parity and new trade theory* (Vol. 2). Retrieved from https://papers.ssrn.com

MacDonald, R., & Taylor, M.P. (1993). The monetary approach to the exchange rate: rational expectations, long-run equilibrium, and forecasting. *Staff Papers*, 40(1), 89-107.

MacDonald, R., & Taylor, M.P. (1994). The monetary model of the exchange rate: long-run relationships, short-run dynamics and how to beat a random walk. *Journal of International Money and Finance*, 13(3), 276-290.

Masih, R., & Masih, A.M. (1996). Stock-Watson Dynamic OLS (DOLS) and error-correction modelling approaches to estimating long-and short-run elasticities. *Energy Economics*, *18*(4), 315-334.

Mathisen, M.J. (2003). *Estimation of the equilibrium real exchange rate for Malawi* (No. 3-104). Retrieved from htpps://www.imf.org

Miyajima, M.K. (2007). *What do we know about Namibia's competitiveness* (No. 7-191). Retrieved from https://www.imf.org

Mkenda, B.K. (2001). Long-run and short-run determinants of the real exchange rate in Zambia.

Retrieved from https://www.researchgate.net

Naseem, N.A., & Hamizah, M.S. (2013). Exchange rate misalignment and economic growth: recent evidence in Malaysia. *Pertanika Journal of Social Science and Humanities*, 21, 47-66.

Ndlela, T. (2012). Implications of real exchange rate misalignment in developing countries: theory, empirical evidence and application to growth performance in Zimbabwe. *South African Journal of Economics*, 80(3), 319 – 344.

Ozsoz, E., & Akinkunmi, M. (2012). Real exchange rate assessment for Nigeria: An evaluation of determinants, strategies for identification and correction of misalignments. *OPEC Energy Review*, 36(1), 104-123.

Pesaran, M.H., Shin, Y., & Smith, R.P. (1999). Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 94(446), 621-634.

Phillips, P.C., & Moon, H.R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, *67*(5), 1057-1111.

Ravn, M.O., Schmitt-Grohé, S., & Uribe, M. (2012). Consumption, government spending, and the real exchange rate. *Journal of Monetary Economics*, 59(3), 215-234.

Reinhart, C.M., & Végh, C.A. (1995). Nominal interest rates, consumption booms, and lack of credibility: A quantitative examination. *Journal of Development Economics*, 46(2), 357-378.

Ricci, L.A. (2005). *South Africa's real exchange rate performance. Post-apartheid South Africa:*

*The First Ten Years.* Retrieved from https://www.imf.org

Ricci, M.L.A., Milesi-Ferretti, M.G.M., & Lee, M.J. (2008). *Real exchange rates and fundamentals: a cross-country perspective* (No. 8-13). Retrieved from https://www.imf.org

Rochester, L. (2013). *Estimating Jamaica's fundamental equilibrium exchange rate*. Retrieved from boj.org.jm

Rodrik, D. (2008). The real exchange rate and economic growth. *Brookings Papers on Economic Activity*, 2, 365–412.

Sallenave, A. (2010). Real exchange rate misalignments and economic performance for the G20 countries. *Economie Internationale*, 1, 59-80.

Sarno, L., & Schmeling, M. (2013). Which fundamentals drive exchange rates? A cross-sectional perspective. *Journal of Money, Credit and Banking*, 46(2-3), 267-292.

Siregar, R. (2011). *The concepts of equilibrium exchange rate: A survey of literature*. Retrieved from https://mpra.ub.uni-muenchen.de

Stein, J. (1994). The natural real exchange rate of the US dollar and determinants of capital flows.

In J. Williamson (Ed.), *Estimating equilibrium exchange rate* (pp.23-46). Washington DC: Institute for International Economics.

Stock, J.H., & Watson, M. (1993). A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica*, 61,783-820.

Taylor, A.M., & Taylor, M.P. (2004). The purchasing power parity debate. *Journal of Economic Perspectives*, 18(4), 35-158.

Tsen Wong, H. (2013). Real exchange rate misalignment and economic growth in Malaysia. *Journal of Economic Studies*, 40(3), 298-313.

Vieira, F.V., & MacDonald, R. (2012). A panel data investigation of real exchange rate misalignment and growth. *Estudos Econômicos* (São Paulo), 42(3), 433-456.

Zhang, Z. (2001). Real exchange rate misalignment in China: An empirical investigation. *Journal of Comparative Economics*, 29(1), 80-94.

# Low frequency estimation of Lévy-driven moving averages*

Mikkel Slot Nielsen[0000−0001−8427−1382]

Aarhus University, Denmark
mikkel@math.au.dk

**Abstract.** In this paper we consider least squares estimation of the driving kernel of a moving average and argue that, under mild regularity conditions and a decay condition on the kernel, the suggested estimator is consistent and asymptotically normal. On one hand this result unifies scattered results of the literature on low frequency estimation of moving averages, and on the other hand it emphasizes the validity of inference also in cases where the moving average is not strongly mixing. We assess the performance of the estimator through a simulation study.

**Keywords:** Least squares estimation · Lévy-driven moving averages · Long memory processes.

## 1 Introduction

The class of continuous time Lévy-driven moving averages of the form

$$X_t = \int_{\mathbb{R}} \varphi(t-s)\, dL_s, \quad t \in \mathbb{R}, \tag{1}$$

where $(L_t)_{t\in\mathbb{R}}$ is a Lévy process with $\mathbb{E}L_1 = 0$ and $\mathbb{E}L_1^2 < \infty$ and $\varphi \in L^2$, is large and has received much attention in earlier literature. Part of the reason for this popularity might be explained by the celebrated discrete time counterpart (in particular, ARMA processes) as well as the Wold-Karhunen decomposition. The latter states that, up to a drift term, essentially any centered and square integrable stationary process may be written in the form (1) with $(L_t)_{t\in\mathbb{R}}$ replaced by a process with second order stationary and orthogonal increments ([2, 15]). While $\varphi$ may be specified directly, one often characterizes it in the spectral domain in terms of its Fourier transform,

$$\widehat{\varphi}(y) = \int_0^\infty e^{-iyt}\varphi(t)\, dt, \quad y \in \mathbb{R}.$$

One class in the framework of (1) is the continuous time ARMA (CARMA) processes, where $\widehat{\varphi}(y) = Q(iy)/P(iy)$ for some monic polynomials $P, Q : \mathbb{C} \to \mathbb{C}$

* Supported by the Danish Council for Independent Research (Grant DFF - 4002–00003).

with real coefficients, $p := \deg(P) > \deg(Q) =: q$, and $P(z) \neq 0$ for all $z \in \mathbb{C}$ with $\mathrm{Re}(z) \geq 0$. One may regard a CARMA process as the solution to the formal equation

$$P(D)X_t = Q(D)DL_t, \quad t \in \mathbb{R}, \tag{2}$$

where $D$ denotes the derivative with respect to time. Indeed, by heuristically applying the Fourier transform to (2) and rearranging terms one reaches the conclusion that $X$ is the convolution between $\varphi$ and $DL$. The simplest CARMA process, which has been particularly popular, is the Ornstein-Uhlenbeck process which corresponds to $p = 1$ and $q = 0$. CARMA processes have been used as models for various quantities including stochastic volatility, electricity spot prices and temperature dynamics ([4, 11, 24]), and there exists a vast amount of literature on their existence, uniqueness and representations as well as generalizations to the multivariate and fractional noise setting ([5, 18, 19]). Another class consists of affine stochastic delay differential equations (SDDEs) of the form

$$dX_t = \int_{[0,\infty)} X_{t-s}\, \eta(ds)\, dt + dL_t, \quad t \in \mathbb{R}. \tag{3}$$

Here $\eta$ is a suitable finite signed measure satisfying $z - \int_{[0,\infty)} e^{-zt}\, \eta(dt) \neq 0$ for all $z \in \mathbb{C}$ with $\mathrm{Re}(z) \geq 0$. In this case, the solution of (3) is a moving average and the kernel $\varphi$ is determined by the relation

$$\widehat{\varphi}(y) = \left( iy - \int_{[0,\infty)} e^{-iyt}\, \eta(dt) \right)^{-1}, \quad y \in \mathbb{R}. \tag{4}$$

The choice $\eta = -\lambda\delta_0$, $\lambda > 0$, results in the Ornstein-Uhlenbeck process; a related example is considered in Example 3. (We use the notation $\delta_x$ for the Dirac measure at $x$.) Some relevant references on SDDEs are [3, 13].

Estimation of $P$ and $Q$, given a sample $X_{n:1} = [X_{n\Delta}, X_{(n-1)\Delta}, \ldots, X_{\Delta}]^\top$ of equidistant observations of a CARMA process sampled at some frequency $\Delta > 0$, has received some attention. For instance, Brockwell et al. [7] show that a sampled CARMA process $(X_{t\Delta})_{t\in\mathbb{Z}}$ is a weak ARMA process. By combining this with the fact that CARMA processes are strongly mixing ([19, Proposition 3.34]), they can rely on general results of Francq and Zakoïan [10] to prove strong consistency and asymptotic normality for an estimator of least squares type. Other papers dealing with low frequency estimation of CARMA processes are [9, 22]. Küchler and Sørensen [17] studied low frequency parametric estimation of the measure $\eta$ in (3) in case the support of the measure is known to be contained in some compact set and $(L_t)_{t\in\mathbb{R}}$ is a Brownian motion. They used results about strong mixing properties of Gaussian processes to obtain consistency and asymptotic normality of a maximum pseudo likelihood estimator. Generally, these results for CARMA processes and solutions to SDDEs cannot be extended to other parametric classes of $\varphi$ in (1), since they use specific properties of the subclass in question. Indeed, strong mixing conditions may be difficult to verify and there exist several non-trivial examples of processes which are not strongly

mixing (see the discussion and the corresponding examples in [21]). There exist results on strong mixing properties for discrete time moving averages, such as [12], but to the best of our knowledge, no version for the continuous time counterpart (1) has been proven (not even when it is sampled on a discrete grid).

In this paper we provide a result (Theorem 2) concerning consistency and asymptotic normality of an estimator of least squares type when parametrically estimating $\varphi$ in (1) from a sample of low frequency observations $X_{n:1}$. To be more concrete, let $\Theta$ be a compact subset of $\mathbb{R}^d$, let $\varphi_\theta \in L^2$ for $\theta \in \Theta$, and suppose that $(X_t)_{t\in\mathbb{R}}$ follows the model (1) with $\varphi = \varphi_{\theta_0}$ for some unknown parameter $\theta_0 \in \Theta$. Then we will be interested in the estimator $\hat{\theta}_n$ obtained as a point, which minimizes

$$\sum_{t=k+1}^{n} (X_{t\Delta} - \pi_k(X_{t\Delta};\theta))^2, \quad \theta \in \Theta, \tag{5}$$

where $\pi_k(X_{t\Delta};\theta)$ denotes the projection of $X_{t\Delta}$ onto the linear $L^2(\mathbb{P})$ subspace spanned by $X_{(t-1)\Delta}, \ldots, X_{(t-k)\Delta}$ under the model (1) with $\varphi = \varphi_\theta$. Besides the usual identifiability and smoothness conditions, the conditions given here to ensure asymptotic normality of the estimator concern the decay of the kernel. This ensures that we can apply our result in situations where the process is not, or cannot be verified to be, strongly mixing. In cases where $\varphi_\theta$ can be specified directly, e.g., when it belongs to the class of CARMA processes or fractional noise processes, it is a straightforward task to check the decay condition, but even when the kernel is not explicitly known (e.g., when it can only be specified through its Fourier transform as in the SDDE case) one can sometimes still assess its decay properties. In Example 1 we consider some situations where the imposed decay condition is satisfied. Section 3 demonstrates the properties of the estimator through a simulation study.

## 2 Estimators of interest and asymptotic results

Let $(L_t)_{t\in\mathbb{R}}$ be a centered Lévy process with $\mathbb{E}L_1 = 0$ and $\mathbb{E}L_1^4 < \infty$, and suppose that $\mathbb{E}L_1^2 = 1$. Moreover, let $\Theta$ be a compact subset of $\mathbb{R}^d$ and, for each $\theta \in \Theta$, suppose that $\varphi_\theta \in L^2$ and define the corresponding stationary process $(X_t^\theta)_{t\in\mathbb{R}}$ by

$$X_t^\theta = \int_{\mathbb{R}} \varphi_\theta(t-s)\, dL_s, \quad t \in \mathbb{R}. \tag{6}$$

To avoid trivial cases we assume that $\{t : \varphi_\theta(t) \neq 0\}$ is not a Lebesgue null set. Let $\gamma_\theta$ be the autocovariance function of $(X_t^\theta)_{t\in\mathbb{R}}$, that is,

$$\gamma_\theta(h) := \mathbb{E}\big[X_h^\theta X_0^\theta\big] = \int_{\mathbb{R}} \varphi_\theta(h+t)\varphi_\theta(t)\, dt, \quad h \in \mathbb{R}. \tag{7}$$

It will be assumed throughout that $\theta \mapsto \gamma_\theta(h)$ is twice continuously differentiable for all $h$. Recall that, for fixed $\Delta > 0$ and any $t \in \mathbb{Z}$, the projection of

$X^\theta_{t\Delta}$ onto the linear span of $X^\theta_{(t-1)\Delta}, \dots, X^\theta_{(t-k)\Delta}$ is given by $\alpha_k(\theta)^\top X^\theta_{t-1:t-k}$ where $\alpha_k(\theta) = \Gamma_k(\theta)^{-1}\gamma_k(\theta)$, $\Gamma_k(\theta) = [\gamma_\theta((i-j)\Delta)]_{i,j=1,\dots,k}$ is the covariance matrix of $X^\theta_{t-1:t-k}$, and $\gamma_k(\theta) = [\gamma_\theta(\Delta), \dots, \gamma_\theta(k\Delta)]^\top$. (Here we use the notation $Y_{t:s} = [Y_{t\Delta}, Y_{(t-1)\Delta}, \dots, Y_{s\Delta}]^\top$ for $s, t \in \mathbb{Z}$ with $s < t$.) Note that by [6, Proposition 5.1.1], $\Gamma_k(\theta)$ is always invertible. Now suppose that $X_t = X^{\theta_0}_t$ for all $t \in \mathbb{R}$ and some unknown parameter $\theta_0$ belonging to the interior of $\Theta$, and consider $n$ equidistant observations $X_{n:1} = [X_{n\Delta}, \dots, X_\Delta]^\top$. We will estimate $\theta_0$ by the least squares estimator $\hat{\theta}_n$, which is chosen to minimize (5). Thus, with the introduced notation,

$$\hat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \sum_{t=k+1}^{n} (X_{t\Delta} - \alpha_k(\theta)^\top X_{(t-1):(t-k)})^2. \tag{8}$$

The estimator (8) can be seen as a truncated version of

$$\tilde{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \sum_{t=2}^{n} (X_{t\Delta} - \alpha_{t-1}(\theta)^\top X_{(t-1):1})^2. \tag{9}$$

From an implementation point of view, while evaluation of the objective function in (9) will demand computing $\alpha_1(\theta), \dots, \alpha_{n-1}(\theta)$ (usually obtained recursively by the Durbin-Levinson algorithm, [6, Proposition 5.2.1]), one only needs to compute $\alpha_k(\theta)$ in order to evaluate the objective function in (8). As discussed in [17], in short-memory models where the projection coefficients are rapidly decaying, it is reasonable to use $\hat{\theta}_n$ with a suitably chosen depth $k$ as a proxy for (9).

To show strong consistency and asymptotic normality of $\hat{\theta}_n$ we impose the following set of conditions:

**Condition 1**

*(a)* $\gamma_\theta(j\Delta) = \gamma_{\theta_0}(j\Delta)$ *for* $j = 0, 1, \dots, k$ *if and only if* $\theta = \theta_0$.
*(b)* $\gamma_k'(\theta_0) - \Gamma_k'(\theta_0)\big[\alpha_k(\theta_0) \otimes I_d\big]$ *has full rank.*
*(c)* $\big(t \mapsto \sum_{s\in\mathbb{Z}} |\varphi_{\theta_0}(t + s\Delta)|^\beta\big) \in L^2([0, \Delta])$ *for* $\beta = 4/3, 2$.

*Remark 1.* Concerning Condition 1, (a)-(b) are standard assumptions ensuring that $\theta_0$ is identifiable from the auto-covariances and that the (suitably scaled version of the) second derivative of the objective function in (8) converges to an invertible deterministic matrix. The difference between Condition 1 and the typical set of conditions for proving asymptotic normality is that an assumption on the strong mixing coefficients of $(X_{t\Delta})_{t\in\mathbb{Z}}$ is replaced by (c), a rather explicit condition on the driving kernel. In fact, according to [20, Theorem 1.2], sufficient conditions for (c) to be satisfied are that

$$\varphi_{\theta_0} \in L^4 \quad \text{and} \quad \sup_{t\in\mathbb{R}} |t|^\beta |\varphi_{\theta_0}(t)| < \infty \tag{10}$$

for a suitable $\beta \in (3/4, 1)$.

*Example 1.* In view of Remark 1 the key condition to check is if we are in a subclass of moving average processes, where (1) (or, more generally, Condition 1(c)) holds true. In the following we consider a few popular classes of kernels $\varphi$.

(i) *CARMA and gamma:* It is clear that the gamma kernel $\varphi(t) \propto t_+^\beta e^{-\gamma t}$ meets (1) when $\beta \in (-1/4, \infty)$ and $\gamma \in (0, \infty)$. The CARMA kernel characterized in Section 1 can always be bounded by a sum of gamma kernels (see, e.g., [5, Equation (36)]), and hence (1) is satisfied for this choice as well.

(ii) *SDDE:* If the variation $|\eta|$ of $\eta$ satisfies $\int_{[0,\infty)} t^2 \, |\eta|(dt) < \infty$, it follows by [20, Example 3.10] that the kernel $\varphi$ associated to the solution of (3) meets (1).

(iii) *Fractional noise:* If $\varphi(t) \propto t_+^d - (t-\tau)_+^d$ for some $d \in (0, 1/4)$ and $\tau \in (0, \infty)$, then $\varphi$ is continuous on $\mathbb{R}$ and the mean value theorem implies that $\varphi(t)$ is asymptotically proportional to $t^{d-1}$ as $t \to \infty$. These properties establish the validity of (1). Note that the corresponding discretely sampled moving average $(X_{t\Delta})_{t\in\mathbb{Z}}$ is not strongly mixing in this setup (cf. [8, Theorem A.1]).

Before stating and proving consistency and asymptotic normality of $\hat{\theta}_n$ in (8) we introduce some notation. For a twice continuously differentiable function $f$, defined on some open subset of $\mathbb{R}^d$ and with values in $\mathbb{R}^m$, the gradient and Hessian of $f$ at $\theta$ are denoted by $f'(\theta)$ and $f''(\theta)$, respectively:

$$f'(\theta) = \left[ \frac{\partial f}{\partial \theta_1}(\theta), \ldots, \frac{\partial f}{\partial \theta_d}(\theta) \right] \in \mathbb{R}^{m \times d}, \quad f''(\theta) = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_d \partial \theta_d} \end{bmatrix} \in \mathbb{R}^{dm \times d}.$$

Moreover, with $v_1(\theta)^\top = [1, -\alpha_k(\theta)^\top]$, $v_2(\theta)^\top = [0, \alpha_k'(\theta)^\top]$ and $F_s(t;\theta) = [\varphi_\theta(t - (i-1)\Delta)\varphi_\theta(t - (j-s-1)\Delta)]_{i,j=1,\ldots,k+1}$, we define

$$V_s^{ij}(t;\theta) = v_i(\theta)^\top F_s(t;\theta) v_j(\theta), \quad i,j = 1,2, \quad s \in \mathbb{Z}. \tag{11}$$

Finally, we set $\sigma^2 = \mathbb{E}L_1^2$ and $\kappa_4 = \mathbb{E}L_1^4 - 3\sigma^4$.

**Theorem 2.** *Suppose that $\theta_0$ belongs to the interior of $\Theta$ and that Condition 1 is in force. Let $\hat{\theta}_n$ be the estimator given in* (8)*. Then $\hat{\theta}_n \to \theta_0$ almost surely and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, H^{-1}AH^{-1})$ as $n \to \infty$, where $H = 2\alpha_k'(\theta_0)^\top \Gamma_k(\theta_0)\alpha_k'(\theta_0)$ and*

$$\begin{aligned} A = \sum_{s\in\mathbb{Z}} \Big( & \kappa_4 \int_{\mathbb{R}} V_s^{11}(t;\theta_0) V_s^{22}(t;\theta_0) \, dt + \sigma^4 \int_{\mathbb{R}} V_s^{11}(t;\theta_0) \, dt \int_{\mathbb{R}} V_s^{22}(t;\theta_0) \, dt \\ & + \sigma^4 \int_{\mathbb{R}} V_s^{21}(t;\theta_0) \, dt \int_{\mathbb{R}} V_s^{12}(t;\theta_0) \, dt \Big). \end{aligned} \tag{12}$$

*Proof.* Set $\ell_n(\theta) = \sum_{t=k+1}^n (X_{t\Delta} - \alpha_k(\theta)^\top X_{(t-1):(t-k)}))^2$, and let $\ell_n'$ and $\ell_n''$ be the first and second order derivative of $\ell_n$, respectively. As usual, the consistency and part of the asymptotic normality rely on an application of a suitable

(uniform) ergodic theorem to ensure almost sure convergence of the sequences $(n^{-1}\ell_n)_{n\in\mathbb{N}}$ and $(n^{-1}\ell_n'')_{n\in\mathbb{N}}$. The difference lies in the proof of a central limit theorem for $(n^{-1/2}\ell_n'(\theta_0))_{n\in\mathbb{N}}$.

Consistency: Note that $\mathbb{E}\big[\sup_{\theta\in\Theta}(X_{k\Delta}-\alpha_k(\theta)^\top X_{(k-1):0})^2\big]<\infty$, since the vector of projection coefficients $\alpha_k(\theta)$ is bounded due to the continuity of $\theta\mapsto\gamma_\theta(h)$. Thus, we find by the ergodic theorem for Banach spaces ([23, Theorem 2.7]) that $n^{-1}\ell_n(\theta)\to\mathbb{E}\big[(X_{k\Delta}-\alpha_k(\theta)^\top X_{(k-1):0})^2\big]=:\ell^*(\theta)$ almost surely and uniformly in $\theta$ as $n\to\infty$. Thus, strong consistency follows immediately if $\ell^*$ is uniquely minimized at $\theta_0$. Since $\alpha_k(\theta_0)^\top X_{(k-1):0}$ is the projection of $X_{k\Delta}$ onto the linear span of $X_0,\dots,X_{(k-1)\Delta}$, it must be the case that $\ell^*(\theta_0)\le\ell^*(\theta)$ for all $\theta\in\Theta$. If $\theta\neq\theta_0$, Condition 1(a) implies that $\gamma_\theta(j\Delta)\neq\gamma_{\theta_0}(j\Delta)$ for at least one $j$, and hence $\ell^*(\theta_0)<\ell^*(\theta)$ by uniqueness of the projection coefficients.

Asymptotic normality: It suffices to show that (i) $n^{-1}\ell_n''(\theta)$ converges almost surely and uniformly in $\theta$ as $n\to\infty$ and $H:=\lim_{n\to\infty}n^{-1}\ell_n''(\theta_0)$ is a deterministic positive definite matrix, and (ii) $n^{-1/2}\ell_n'(\theta_0)$ converges in distribution to a Gaussian random variable. Concerning (i), note that

$$\ell_n''(\theta)=2\sum_{t=k+1}^{n}\Big[\alpha_k'(\theta)^\top X_{(t-1):(t-k)}X_{(t-1):(t-k)}^\top\alpha_k'(\theta)\\ -\big(X_{t\Delta}-\alpha_k(\theta)^\top X_{(t-1):(t-k)}\big)\big[X_{(t-1):(t-k)}^\top\otimes I_d\big]\,\alpha_k''(\theta)\Big],$$

where $I_d$ is the $d\times d$ identity matrix and the $j$-th row of $\alpha_k'$ (resp. the $j$-th $d\times d$ block of $\alpha_k''$) is the gradient (resp. Hessian) of the $j$-th entry of $\alpha_k$. Thus, it follows by [23, Theorem 2.7] that $n^{-1}\ell_n''(\theta)\to 2\alpha_k'(\theta)^\top\Gamma_k(\theta_0)\alpha_k'(\theta)=:H(\theta)$ almost surely and uniformly in $\theta$ as $n\to\infty$. Since $\Gamma_k(\theta_0)$ is positive definite and

$$\alpha_k'(\theta_0)=\Gamma_k(\theta_0)^{-1}\big(\gamma_k'(\theta_0)-\Gamma_k'(\theta_0)\big[\alpha_k(\theta_0)\otimes I_d\big]\big),$$

it follows from Condition 1(b) that $H=H(\theta_0)$ is positive definite. To show (ii), observe that $\ell_n'(\theta_0)$ takes the form

$$\ell_n'(\theta_0)=\sum_{t=k+1}^{n}\int_{\mathbb{R}}\psi_1(t\Delta-s)\,dL_s\int_{\mathbb{R}}\psi_2(t\Delta-s)\,dL_s$$

with $\psi_i(t)=v_i(\theta_0)^\top\varphi_{\theta_0,k}(t)$, using the notation

$$\varphi_{\theta_0,k}(t)=\big[\varphi_{\theta_0}(t),\varphi_{\theta_0}(t-\Delta),\dots,\varphi_{\theta_0}(t-k\Delta)\big]^\top.$$

Since the space of functions $f$ satisfying

$$\Big(t\mapsto\sum_{s\in\mathbb{Z}}|f(t+s\Delta)|^\beta\Big)\in L^2([0,\Delta])\quad\text{for}\quad\beta=4/3,2 \tag{13}$$

forms a vector space, and $\varphi_{\theta_0}$ satisfies (13) by Condition 1(c), $\psi_1$ and (each entry of) $\psi_2$ satisfy (13) as well. Moreover, as $\int_{\mathbb{R}}\psi_1(t\Delta-s)\,dL_s=X_{t\Delta}-$

$\alpha_k(\theta_0)^\top X_{(t-1):(t-k)}$ is orthogonal to $\int_{\mathbb{R}} \psi_2(t\Delta - s)\, dL_s = X_{(t-1):(t-k)}^\top \alpha_k'(\theta_0)$ in $L^2(\mathbb{P})$ (entrywise), we have that $\mathbb{E}\ell_n'(\theta_0) = 0$. Consequently, by [20, Theorem 1.2], $n^{-1/2}\ell_n'(\theta_0)$ converges in distribution to a centered Gaussian vector with covariance matrix given by

$$\sum_{s\in\mathbb{Z}}\Big(\kappa_4 \int_{\mathbb{R}} \psi_1(t)\psi_1(t+s\Delta)\psi_2(t)\psi_2(t+s\Delta)^\top\, dt + \sigma^4 \int_{\mathbb{R}} \psi_1(t)\psi_1(t+s\Delta)\, dt \\ \cdot \int_{\mathbb{R}} \psi_2(t)\psi_2(t+s\Delta)^\top\, dt + \sigma^4 \int_{\mathbb{R}} \psi_1(t)\psi_2(t+s\Delta)\, dt \int_{\mathbb{R}} \psi_2(t)^\top \psi_1(t+s\Delta)\, dt\Big),$$

which is equal to $A$ given in (12). This concludes the proof.

## 3 Examples

In this section we give two examples where Theorem 2 is applicable and accompany these by simulating the properties of the estimator $\hat{\theta}_n$. In both examples we fix the sample frequency $\Delta = 1$ as well as the depth $k = 10$. We have checked (by simulation) that the estimator is rather insensitive to the choice of $k$; this is supported by the fact that both models result in geometrically decaying projection coefficients.

*Example 2.* Suppose that $(L_t)_{t\in\mathbb{R}}$ is a standard Brownian motion and, for $\theta = (\nu, \lambda) \in (3/4, \infty) \times (0, \infty)$, set

$$\varphi_\theta(t) = \Gamma(\nu)^{-1} t^{\nu-1} e^{-\lambda t}, \quad t > 0. \tag{14}$$

The moving average model (6) with gamma kernel (14) has received some attention in the literature and has, e.g., been used to model the timewise behavior of the velocity in turbulent regimes (see [1] and references therein). Moreover, particular choices of $\nu$ result in special cases of well-known and widely studied models. To be concrete, if $\nu = 1$ then $(X_t^\theta)_{t\in\mathbb{R}}$ is an Ornstein-Uhlenbeck process with parameter $\lambda > 0$ and, more generally, if $\nu \in \mathbb{N}$ then $(X_t^\theta)_{t\in\mathbb{R}}$ is a CAR($\nu$) process with polynomial $P(z) = (z + \lambda)^\nu$. The autocovariance function $\gamma_\theta$ of $(X_t^\theta)_{t\in\mathbb{R}}$ under the model specification (6) and (14) takes the form

$$\gamma_\theta(h) = \begin{cases} \Gamma(2\nu - 1)(2\lambda)^{1-2\nu} & \text{for} \quad h = 0, \\ \Gamma(\nu)(2\pi^{-1})^{1/2} 2^{-\nu} (\lambda^{-1}|h|)^{\nu-1/2} K_{\nu-1/2}(\lambda|h|) & \text{for} \quad h \neq 0, \end{cases}$$

where $K_{\nu-1/2}$ denotes the modified Bessel function of the third kind of order $\nu - 1/2$ (cf. [1]). The corresponding autocorrelation function $\gamma_\theta/\gamma_\theta(0)$ is known as the Whittle-Matérn correlation function ([14]). In Figure 1 we have simulated $X_{400:1}$ and plotted the corresponding sample and theoretical autocorrelation function for $\theta_0 = (1.3, 1.1)$. To demonstrate the ability to infer the true parameter $\theta_0 = (\nu_0, \lambda_0)$ from $X_{n:1}$ using the least squares estimator (8) we simulate $X_{n:100}$ under the model corresponding to $\theta_0$ for $n = 400, 1600, 6400$, obtain the associated realizations of $\hat{\theta}_n = (\hat{\nu}_n, \hat{\lambda}_n)$ for truncation lag $k = 10$ and repeat the experiment

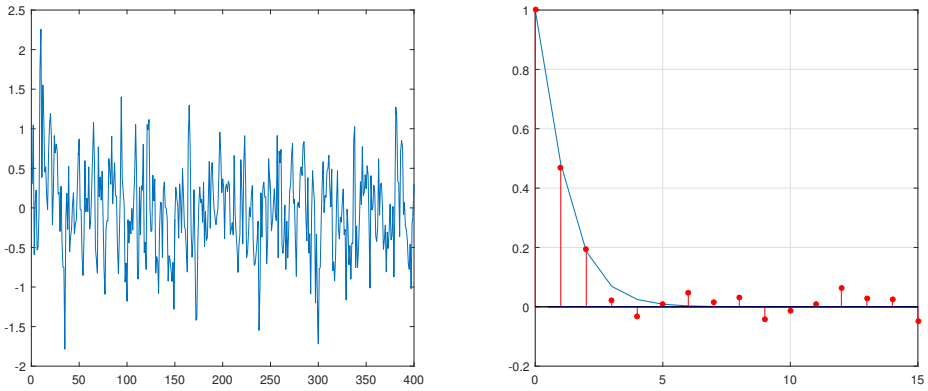



**Fig. 1.** Left: simulation of $X_{400:1}$ under the model specification (6) and (14) with $\theta_0 = (1.3, 1.1)$ when $(L_t)_{t\in\mathbb{R}}$ is a Brownian motion. Right: the corresponding sample autocorrelation function and its theoretical counterpart.

500 times. We perform this study for different choices of $\theta_0$. In Table 1 we have, for each $n$, summarized the sample mean, bias and variance for the realizations of the least squares estimator. To show the robustness regarding the choice of the underlying noise we did the same analysis in the case where $(L_t)_{t\in\mathbb{R}}$ is a centered gamma Lévy process with both shape and scale parameter equal to one. In other words, $(L_t)_{t\in\mathbb{R}}$ was chosen to be the unique Lévy process where $L_1$ has density $t \mapsto \mathbb{1}_{\{t\geq -1\}} e^{-t-1}$. The findings, which are reported in Table 2, are seen to be similar to those of Table 1. To illustrate the asymptotic normality of $\hat{\theta}_n$ we have plotted histograms based on the 500 realizations of $\hat{\nu}_n$ and $\hat{\lambda}_n$ when $n = 6,400$ in the situation where $(\nu_0, \lambda_0) = (1.3, 1.1)$ (see Figure 2).

*Example 3.* As in the last part of Example 2 let $(L_t)_{t\in\mathbb{R}}$ be a centered gamma Lévy process with both shape and scale parameter equal to one, and consider the model (3) where $\eta = \alpha\delta_0 + \beta\delta_1$ for some $\alpha, \beta \in \mathbb{R}$:

$$dX_t = (\alpha X_t + \beta X_{t-1})\, dt + dL_t, \quad t \in \mathbb{R}. \tag{15}$$

We will perform a simulation study similar to that of [17], except that they consider a Brownian motion as the underlying noise and use a certain pseudo (Gaussian) likelihood rather than the least squares estimator in (8). In [16] it is argued that a stationary solution to (15) exists if $\alpha < 1$ and

$$\beta \in \begin{cases} \left(-\frac{\alpha}{\cos(\xi(\alpha))}, -\alpha\right) & \text{if} \quad \alpha \neq 0, \\ \left(-\frac{\pi}{2}, 0\right) & \text{if} \quad \alpha = 0. \end{cases}$$

The function $\xi$ is characterized by $\xi(0) = \pi/2$ and $\xi(t) = t\tan(\xi(t))$ for $t \neq 0$. We will compute (8) by using that

$$\gamma_\theta(h) = 2\int_0^\infty \frac{\cos(hy)}{|iy + \alpha + \beta e^{iy}|^2}\, dy, \quad h \in \mathbb{R},$$

which follows from (4), (7) and Plancherel's theorem. We choose $(\alpha_0, \beta_0) = (-1, -0.1353)$ in line with [17] and provide statistics similar to those of Tables 1-2 in Table 3.

| $n$ | | 400 | | 1,600 | | 6,400 | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\nu}_n$ | $\hat{\lambda}_n$ | $\hat{\nu}_n$ | $\hat{\lambda}_n$ | $\hat{\nu}_n$ | $\hat{\lambda}_n$ |
| $\nu_0 = 1.3$ $\lambda_0 = 1.1$ | Mean | 1.3869 | 1.1613 | 1.3353 | 1.1271 | 1.3008 | 1.0982 |
| | Bias | 0.0869 | 0.0613 | 0.0353 | 0.0271 | 0.0008 | −0.0018 |
| | Var. ×10 | 1.2143 | 1.5452 | 0.3553 | 0.4501 | 0.0749 | 0.1039 |
| $\nu_0 = 0.9$ $\lambda_0 = 1.1$ | Mean | 1.1460 | 1.4244 | 0.9867 | 1.2151 | 0.9092 | 1.1079 |
| | Bias | 0.2460 | 0.3244 | 0.0857 | 0.1151 | 0.0092 | 0.0079 |
| | Var. ×10 | 2.0205 | 4.7529 | 0.6148 | 1.6267 | 0.0851 | 0.3354 |
| $\nu_0 = 1.3$ $\lambda_0 = 0.5$ | Mean | 1.3202 | 0.5166 | 1.3079 | 0.5060 | 1.2989 | 0.4987 |
| | Bias | 0.0202 | 0.0166 | 0.0079 | 0.0060 | −0.0011 | −0.0013 |
| | Var. ×10 | 0.1910 | 0.1424 | 0.0417 | 0.0333 | 0.0099 | 0.0079 |

**Table 1.** Sample mean, bias and variance based on 500 realizations of $\hat{\theta}_n = (\hat{\nu}_n, \hat{\lambda}_n)$ various choices of $n$. The noise is a Brownian motion.

| $n$ | | 400 | | 1,600 | | 6,400 | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\nu}_n$ | $\hat{\lambda}_n$ | $\hat{\nu}_n$ | $\hat{\lambda}_n$ | $\hat{\nu}_n$ | $\hat{\lambda}_n$ |
| $\nu_0 = 1.3$ $\lambda_0 = 1.1$ | Mean | 1.3638 | 1.1537 | 1.3158 | 1.1234 | 1.2870 | 1.0969 |
| | Bias | 0.0638 | 0.0537 | 0.0158 | 0.0234 | −0.0130 | −0.0031 |
| | Var. ×10 | 1.1162 | 1.5069 | 0.3358 | 0.4505 | 0.0729 | 0.1061 |
| $\nu_0 = 0.9$ $\lambda_0 = 1.1$ | Mean | 1.1339 | 1.3813 | 1.0049 | 1.2249 | 0.9323 | 1.1262 |
| | Bias | 0.2339 | 0.2813 | 0.1049 | 0.1249 | 0.0323 | 0.0262 |
| | Var. ×10 | 1.8303 | 4.3999 | 0.5879 | 1.5714 | 0.0900 | 0.3446 |
| $\nu_0 = 1.3$ $\lambda_0 = 0.5$ | Mean | 1.3017 | 0.5095 | 1.2902 | 0.5000 | 1.2871 | 0.4964 |
| | Bias | 0.0017 | 0.0095 | −0.0098 | 0.0000 | −0.0129 | −0.0036 |
| | Var. ×10 | 0.1615 | 0.1352 | 0.0401 | 0.0319 | 0.0097 | 0.0079 |

**Table 2.** Sample mean, bias and variance based on 500 realizations of $\hat{\theta}_n = (\hat{\nu}_n, \hat{\lambda}_n)$ various choices of $n$. The noise is a centered gamma Lévy process.

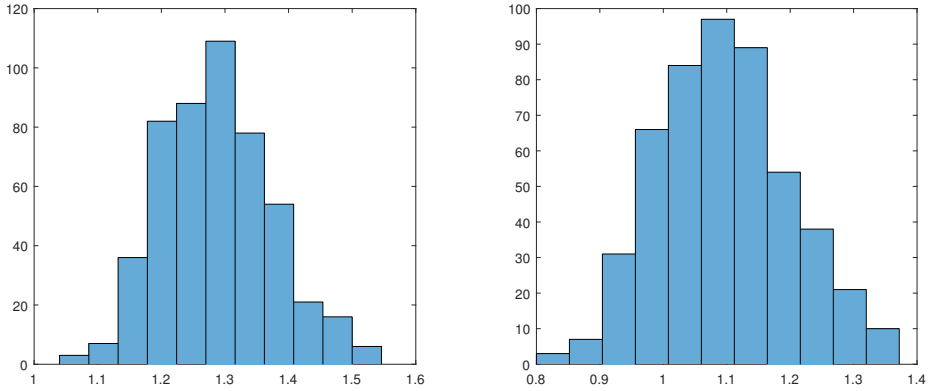



**Fig. 2.** Histograms of 500 realizations of $(\hat{\nu}_{6400}, \hat{\gamma}_{6400})$ when $(\nu_0, \lambda_0) = (1.3, 1.1)$ and the noise is a gamma Lévy process.

| $n$ | 400 | | $1,600$ | | $6,400$ | |
|---|---|---|---|---|---|---|
| | $\hat{\alpha}_n$ | $\hat{\beta}_n$ | $\hat{\alpha}_n$ | $\hat{\beta}_n$ | $\hat{\alpha}_n$ | $\hat{\beta}_n$ |
| Mean | $-0.9980$ | $-0.1654$ | $-1.0127$ | $-0.1508$ | $-1.0132$ | $-0.1459$ |
| Bias | $0.0020$ | $-0.0301$ | $-0.0127$ | $-0.0155$ | $-0.0132$ | $-0.0106$ |
| Var. $\times 10$ | $0.3022$ | $0.0979$ | $0.1165$ | $0.0498$ | $0.0379$ | $0.0189$ |

**Table 3.** Sample mean, bias and variance based on 500 realizations of $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ various choices of $n$ when the true parameters are $\alpha_0 = -1$ and $\beta_0 = -0.1353$. The noise is a centered gamma Lévy process.

## References

1. Barndorff-Nielsen, O. E. (2012). Notes on the gamma kernel. *Thiele Research Reports, Department of Mathematics, Aarhus University*.
2. Barndorff-Nielsen, O. E. and A. Basse-O'Connor (2011). Quasi Ornstein–Uhlenbeck processes. *Bernoulli 17*(3), 916–941.
3. Basse-O'Connor, A., M. S. Nielsen, J. Pedersen, and V. Rohde (2018). Multivariate stochastic delay differential equations and CAR representations of CARMA processes. *To appear in Stochastic Processes and their Applications*.
4. Benth, F. E., J. Šaltytė Benth, and S. Koekebakker (2007). Putting a price on temperature. *Scandinavian Journal of Statistics 34*(4), 746–767.
5. Brockwell, P. (2014). Recent results in the theory and applications of CARMA processes. *Annals of the Institute of Statistical Mathematics 66*(4), 647–685.
6. Brockwell, P. J. and R. A. Davis (2013). *Time series: theory and methods*. Springer Science & Business Media.
7. Brockwell, P. J., R. A. Davis, and Y. Yang (2011). Estimation for non-negative Lévy-driven CARMA processes. *Journal of Business & Economic Statistics 29*(2), 250–259.

8. Cohen, S. and A. Lindner (2013). A central limit theorem for the sample autocorrelations of a Lévy driven continuous time moving average process. *Journal of Statistical Planning and Inference 143*(8), 1295–1306.
9. Fasen-Hartmann, V. and S. Kimmig (2018). Robust estimation of continuous-time ARMA models via indirect inference. *arXiv preprint arXiv:1804.00849*.
10. Francq, C. and J.-M. Zakoïan (1998). Estimating linear representations of nonlinear processes. *Journal of Statistical Planning and Inference 68*(1), 145–165.
11. García, I., C. Klüppelberg, and G. Müller (2011). Estimation of stable CARMA models with an application to electricity spot prices. *Statistical Modelling 11*(5), 447–470.
12. Gorodetskii, V. (1978). On the strong mixing property for linear sequences. *Theory of Probability & Its Applications 22*(2), 411–413.
13. Gushchin, A. A. and U. Küchler (2000). On stationary solutions of delay differential equations driven by a Lévy process. *Stochastic Processes and their Applications 88*(2), 195–211.
14. Guttorp, P. and T. Gneiting (2005). On the Whittle-Matérn correlation family. *National Research Center for Statistics and the Environment-Technical Report Series, Seattle, Washington*.
15. Karhunen, K. (1950). Über die struktur stationärer zufälliger funktionen. *Arkiv för Matematik 1*(2), 141–160.
16. Küchler, U. and B. Mensch (1992). Langevins stochastic differential equation extended by a time-delayed term. *Stochastics: An International Journal of Probability and Stochastic Processes 40*(1-2), 23–42.
17. Küchler, U. and M. Sørensen (2013). Statistical inference for discrete-time samples from affine stochastic delay differential equations. *Bernoulli 19*(2), 409–425.
18. Marquardt, T. (2007). Multivariate fractionally integrated CARMA processes. *Journal of Multivariate Analysis 98*(9), 1705–1725.
19. Marquardt, T. and R. Stelzer (2007). Multivariate CARMA processes. *Stochastic Processes and their Applications 117*(1), 96–120.
20. Nielsen, M. S. and J. Pedersen (2019). Limit theorems for quadratic forms and related quantities of discretely sampled continuous-time moving averages. *To appear in ESAIM: Probability and Statistics*.
21. Nze, P. A., P. Bühlmann, and P. Doukhan (2002). Weak dependence beyond mixing and asymptotics for nonparametric regression. *The Annals of Statistics 30*(2), 397–430.
22. Schlemm, E. and R. Stelzer (2012). Quasi maximum likelihood estimation for strongly mixing state space models and multivariate Lévy-driven CARMA processes. *Electronic Journal of Statistics 6*, 2185–2234.
23. Straumann, D. and T. Mikosch (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics 34*(5), 2449–2495.
24. Todorov, V. and G. Tauchen (2006). Simulation methods for Lévy-driven continuous-time autoregressive moving average (CARMA) stochastic volatility models. *Journal of Business & Economic Statistics 24*(4), 455–469.

# Backtesting Basel III: Evaluating the Market Risk of Past Crises through the Current Regulation

Marcelo Zeuli[1] and André Carvalhal[2]
(1) BCB (Brazilian Central Bank), (2) BNDES

## 1 Abstract

Are the recommendations from the Bank for International Settlements (BIS) effective to a broad set of financial crises?
We submitted two of the main Basel III recommendations for market risk to a back test: the capital requirements and the Value at Risk (VaR) methodology that includes the BIS's Stressed VaR. We tested the main Brazilian currency exchange (U.S. Dollar to Brazilian Reais) and currency exchange swaps contracts through volatility-based VaR methodologies in the period that comprises the so-called Brazilian confidence crisis, which occurred in the second half of 2002.
While the Stressed VaR revealed inapplicable, due to historical data shortage, the capital requirements level appeared innocuous, due to the high levels of daily volatility - daily oscillation limits may have a significant role on crisis mitigation. To circumvent the lack of either historical information or
optimal window for stress patterns, we suggest to calibrate the Stressed VaR or the recently announced Expected Shortfall with a historical VIX (Volatility Index, Chicago Board Options Exchange), working as a volatility scale.
We suggest modeling with other densities, apart from the BIS recommended standard normal.

# Testing normality for unconditionally heteroscedastic macroeconomic variables

October 10, 2017

*Abstract:* In this paper the testing of normality for unconditionally heteroscedastic macroeconomic time series is studied. It is underlined that the classical Jarque-Bera test (JB hereafter) for normality is inadequate in our framework. On the other hand it is found that the approach which consists in correcting the heteroscedasticity by kernel smoothing for testing normality is justified asymptotically. Nevertheless it appears from Monte Carlo experiments that such a methodology can noticeably suffer from size distortion for samples that are typical for macroeconomic variables. As a consequence a parametric bootstrap methodology for correcting the problem is proposed. The innovations distribution of a set of inflation measures for the U.S., Korea and Australia are analyzed.

Keywords: Unconditionally heteroscedastic time series; Jarque-Bera test.

JEL: C12, C15, C18

## 1 Introduction

In the econometric literature, the Jarque Bera (1980) test is routinely used to assess the normality of variables. The properties of this test are well documented for stationary con-

ditionally heteroscedastic processes. For instance Fiorentini, Sentana and Calzolari (2003), Lee, Park and Lee (2010) and Lee (2012) investigated the JB test in the context of GARCH models. However few studies are available on the distributional specification of time series in presence of unconditional heteroscedasticity. Drees and Stărică (2002), Mikosch and Stărică (2004) and Fryźlewicz (2005) investigated the possibility of modelling financial returns by nonparametric methods. To this end, Drees and Stărică (2002) and Mikosch and Stărică (2004) examined the distribution of S&P500 returns corrected from heteroscedasticity. On the other hand Fryźlewicz (2005) pointed out that large sample kurtosis for financial time series may be explained by non constant unconditional variance. In general we did not found references that specifically address the problem of assessing the distribution of unconditionally heteroscedastic time series. Note that non constant variance constitutes an important pattern for time series in general, and macroeconomic variables in particular. Reference can be made to Sensier and van Dijk (2004) who found that most of the 214 U.S. macroeconomic time series they studied have a time-varying variance. In this paper we aim to provide a reliable methodology for testing normality for small samples time series with non constant unconditional variance.

The structure of the paper is as follows. In Section 2 we first set the dynamics ruling the observed process. In particular the unconditional heteroscedastic structure of the errors is given. The inadequacy of the standard JB test in our framework is highlighted. The approach consisting in correcting the errors from the heteroscedasticity for building a JB test is presented. We then introduce a parametric bootstrap procedure that is intended to improve the normality testing for unconditionally heteroscedastic macroeconomic time series. In Section 3 numerical experiments are conducted to shed some light on the finite sample behaviors of the studied tests. In particular it is found that when estimating the non constant variance structure by kernel smoothing, a correct bandwidth choice is a necessary condition for the good implementation of the normality tests based on heteroscedasticity correction. We illustrate our outputs by examining the distributional properties of the U.S.,

Korean and Australian GDP implicit price deflators.

## 2 Testing normality in presence of unconditional heteroscedasticity

We consider processes $(y_t)$ which can be written as

$$\begin{aligned} y_t &= \omega_0 + x_t, \\ x_t &= \sum_{i=1}^{p} a_{0i} x_{t-i} + u_t, \end{aligned} \tag{2.1}$$

where $y_{1,n}, \ldots, y_{n,n}$ are available, $n$ is the sample size and $E(x_t) = 0$. The conditional mean of $x_t$ is driven by the autoregressive parameters $\theta_0 = (a_{01}, \ldots, a_{0p})'$, which fulfill the following condition.

**Assumption A0:** The $a_{0i} \in \mathbb{R}$, $1 \leq i \leq p$, are such that $\det(a(z)) \neq 0$ for all $|z| \leq 1$, with $a(z) = 1 - \sum_{i=1}^{p} a_{0i} z^i$.

In the assumption **A1** below, the well known rescaling device introduced by Dahlhaus (1997) is used to specify the errors process $(u_t)$. For a random variable $v$ we define $\|v\|_q = (E|v|^q)^{1/q}$, with $E|v|^q < \infty$ and $q \geq 1$.

**Assumption A1:** We assume that $u_t = h_t \epsilon_t$ where:

(i) $h_t \geq c > 0$ for some constant $c > 0$, and satisfies $h_t = g(t/n)$, where $g(r)$ is a measurable deterministic function on the interval $(0, 1]$, such that $\sup_{r \in (0,1]} |g(r)| < \infty$. The function $g(.)$ satisfies a Lipschitz condition piecewise on a finite number of some sub-intervals that partition $(0, 1]$.

(ii) The process $(\epsilon_t)$ is iid and such that $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = 1$, and $(E(\|\epsilon_t\|^{8\nu}) < \infty$ for some $\nu > 1$.

The non constant variance induced by **A1(i)** allows for a wide range of non stationarity patterns commonly faced in practice, as for instance abrupt shifts, smooth changes or cyclical behaviors. Note that in the zero mean AR case, tools needed to carry out the Box and Jenkins specification-estimation-validation modeling cycle, are provided in Patilea and Raïssi (2012), Patilea and Raïssi (2013) and Raïssi (2015). For $\omega_0 \neq 0$ define the estimator $\hat{\omega} = n^{-1}\sum_{t=1}^{n} y_t$, and $x_t(\omega) = y_t - \omega$ for any $\omega \in \mathbb{R}$. Writing $\hat{\omega} - \omega_0 = n^{-1}\sum_{t=1}^{n} x_t$, it can be shown that

$$\sqrt{n}(\hat{\omega} - \omega_0) = O_p(1), \tag{2.2}$$

using the Beveridge-Nelson decomposition. Now let

$$\hat{\theta}(\omega) = (\Sigma_{\underline{x}}(\omega))^{-1}\,\Sigma_x(\omega), \tag{2.3}$$

where

$$\Sigma_{\underline{x}}(\omega) = n^{-1}\sum_{t=1}^{n} \underline{x}_{t-1}(\omega)\underline{x}_{t-1}(\omega)' \quad \text{and} \quad \Sigma_x(\omega) = n^{-1}\sum_{t=1}^{n} \underline{x}_{t-1}(\omega)x_{t-1}(\omega),$$

with $\underline{x}_{t-1}(\omega) = (x_{t-1}(\omega), \ldots, x_{t-p}(\omega))'$. With these notations define the OLS estimator $\hat{\theta}(\hat{\omega})$ and the unfeasible estimator $\hat{\theta}(\omega_0)$. Straightforward computations give $\sqrt{n}(\hat{\theta}(\hat{\omega}) - \hat{\theta}(\omega_0))$ $= o_p(1)$, so that using the results of Patilea and Raïssi (2012) we have

$$\sqrt{n}(\hat{\theta}(\hat{\omega}) - \theta_0) = O_p(1). \tag{2.4}$$

Once the conditional mean is filtered in accordance to (2.2) and (2.4), we can proceed to the test of the following hypotheses:

$$H_0 : \epsilon_t \sim \mathcal{N}(0,1) \quad \text{vs.} \quad H_1 : \epsilon_t \text{ has a different distribution,}$$

with the usual slight abuse of interpretation inherent of the use JB test for normality testing. Clearly the skewness and kurtosis of $u_t$ correspond to those of $\epsilon_t$. However in practice $E(u_t^3) = 0$ and $E(u_t^4)/E(u_t^2)^2 = 3$ is checked using the JB test statistic:

$$Q_{JB}^{u} = n\left[Q_{JB}^{S,u} + Q_{JB}^{K,u}\right], \tag{2.5}$$

where

$$Q_{JB}^{S,u} = \frac{\hat{\mu}_3^2}{6\hat{\mu}_2^3} \quad \text{and} \quad Q_{JB}^{K,u} = \frac{1}{24}\left(\frac{\hat{\mu}_4}{\hat{\mu}_2^2} - 3\right)^2,$$

with $\hat{\mu}_j = n^{-1}\sum_{t=1}^{n}(\hat{u}_t - \bar{\hat{u}})^j$ and $\bar{\hat{u}} = n^{-1}\sum_{t=1}^{n}\hat{u}_t$. The $\hat{u}_t$'s are the residuals obtained from the estimation step. Let us denote by $\Rightarrow$ the convergence in distribution. If we suppose the process $(u_t)$ homoscedastic ($g(.)$ is constant), then the standard result $Q_{JB}^{u} \Rightarrow \chi_2^2$ is retrieved (see Yu (2007), Section 2.2). However under **A0** and **A1** with $g(.)$ non constant (the unconditionally heteroscedastic case) we have:

$$Q_{JB}^{K,u} = \frac{1}{24}\left[\kappa_2\left(E(\epsilon_t^4)\right) - 3\right) + 3\left(\kappa_2 - 1\right)\right] + o_p(1), \tag{2.6}$$

where $\kappa_2 = \frac{\int_0^1 g^4(r)dr}{\left(\int_0^1 g^2(r)dr\right)^2}$. Hence if we suppose the errors process unconditionally heteroscedastic with $E(\epsilon_t^4) = 3$, we obtain $Q_{JB}^{u} = Cn + o_p(n)$ for some strictly positive constant $C$. As a consequence, the classical JB test will tend to detect spuriously departures from the null hypothesis of a normal distribution in our framework. This argument was used by Fryźlewicz (2004) to underline that unconditionally heteroscedastic specifications can cover financial time series that typically exhibit an excess of kurtosis.

In order to assess the distribution of S&P500 returns, Drees and Stărică (2002) considered data corrected from heteroscedasticity, using a kernel estimator of the variance. We will follow this approach in the sequel considering

$$\hat{h}_t^2 = \sum_{i=1}^{n} w_{ti}(\hat{u}_i - \bar{\hat{u}})^2, \qquad 1 \le t \le n,$$

with $w_{ti} = \left(\sum_{j=1}^{n} K_{tj}\right)^{-1} K_{ti}$, $K_{ti} = K((t-i)/nb)$ if $t \neq i$ and $K_{ii} = 0$, where $K(\cdot)$ is a kernel function on the real line and $b$ is the bandwidth. The following assumption is needed

for our variance estimator.

**Assumption A2:** (i) The kernel $K(\cdot)$ is a bounded density function defined on the real line such that $K(\cdot)$ is nondecreasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$ and $\int_{\mathbb{R}} v^2 K(v)dv < \infty$. The function $K(\cdot)$ is differentiable except a finite number of points and the derivative $K'(\cdot)$ satisfies $\int_{\mathbb{R}} |xK'(x)|dx < \infty$. Moreover, the Fourier Transform $\mathcal{F}[K](\cdot)$ of $K(\cdot)$ satisfies $\int_{\mathbb{R}} |s|^{\tau} |\mathcal{F}[K](s)| \, ds < \infty$ for some $\tau > 0$.

(ii) The bandwidth $b$ is taken in the range $\mathfrak{B}_n = [c_{min}b_n, c_{max}b_n]$ with $0 < c_{min} < c_{max} < \infty$ and $nb_n^{4-\gamma} + 1/nb_n^{2+\gamma} \to 0$ as $n \to \infty$, for some small $\gamma > 0$.

Let $\hat{\epsilon}_t = (\hat{u}_t - \bar{\hat{u}})/\hat{h}_t$. We are now ready to consider the following JB test statistic:

$$Q_{JB}^{\epsilon} = n \left[ Q_{JB}^{S,\epsilon} + Q_{JB}^{K,\epsilon} \right],$$

where

$$Q_{JB}^{S,\epsilon} = \frac{\hat{\nu}_3^2}{6\hat{\nu}_2^3} \quad \text{and} \quad Q_{JB}^{K,\epsilon} = \frac{1}{24} \left( \frac{\hat{\nu}_4}{\hat{\nu}_2^2} - 3 \right)^2,$$

with $\hat{\nu}_j = n^{-1} \sum_{t=1}^{n} \hat{\epsilon}_t^j$. The following proposition gives the asymptotic distribution of $Q_{JB}^{\epsilon}$.

**Proposition 1.** *Under the assumptions* **A0**, **A1** *and* **A2**, *we have as* $n \to \infty$

$$Q_{JB}^{\epsilon} \Rightarrow \chi_2^2, \tag{2.7}$$

*uniformly with respect to* $b \in \mathfrak{B}_n$.

Proposition 1 can be proved using the same arguments given in Yu (2007), together with those of the proof of Proposition 4 in Patilea and Raïssi (2014). Therefore we skip the proof. For building a test using the above result, we suggest to choose the bandwidth by minimizing the cross-validation (CV) criterion (see Wasserman (2006,p69-70)). On the other hand several kernels available in the literature can be used. In the numerical experiments section

below, we consider the Gaussian kernel and choose the bandwidth by CV unless otherwise specified. The test obtained using (2.7) and choosing the bandwidth by cross-validation is denoted by $T_{cv}$. The standard test that does not take into account the unconditional heteroscedasticity is denoted by $T_{st}$.

For high frequency time series it is reasonable to suppose that the approximation (2.7) is satisfactory when the bandwidth is carefully chosen. Nevertheless considering the above sophisticated procedure for small $n$ is questionable. Therefore we propose to apply the following parametric bootstrap algorithm inspired from Francq and Zakoïan (2010,p335).

1- Generate $\epsilon_t^{(b)} \sim \mathcal{N}(0,1)$, $1+p \leq t \leq n$, build the bootstrap errors $u_t^{(b)} = \epsilon_t^{(b)} \hat{h}_t$, and the bootstrap series $y_t^{(b)}$ using (2.1), but with $\hat{\omega}$ and $\hat{\theta}(\hat{\omega})$ (see (2.2) and (2.3)).

2- Estimate the autoregressive parameters and a constant as in (2.1), but using the $y_t^{(b)}$'s. Build the kernel estimators $\hat{h}_t^{(b)}$ from the resulting residuals $\hat{u}_t^{(b)}$.

3- Compute $\hat{\epsilon}_t^{(b)} = \hat{u}_t^{(b)} / \hat{h}_t^{(b)}$ for $t = 1+p, \ldots, n$. Compute $Q_{JB}^{\epsilon,(b)}$.

4- Repeat the steps 1 to 3 B times for some large B. Use the $Q_{JB}^{\epsilon,(b)}$'s to compute the p-values of the bootstrap JB test.

Of course the $\hat{h}_t^{(b)}$'s in step 2 are always supposed to be obtained in the same way as the $\hat{h}_t$'s (i.e. same bandwidth selection method and kernel). The test obtained using the above parametric bootstrap procedure is denoted by $T_{boot}$.

# 3 Numerical illustrations

The finite sample properties of the $T_{st}$, $T_{cv}$ and $T_{boot}$ tests are first examined by means of Monte Carlo experiments. The distribution of the U.S., Korean and Australian GDP implicit price deflators is then investigated. Throughout this section the asymptotic nominal level of the tests is 5%. In the sequel, we fixed $B = 499$.

### 3.1 Monte Carlo experiments

We simulate $N = 1000$ trajectories of AR(p) processes:

$$y_t = \sum_{i=1}^{p} a_{0i} y_{t-i} + u_t,$$

so that $\omega_0 = 0$. Note however that in all our experiments, the mean is treated as unknown. More precisely the AR parameter is estimated using $y_t - \hat{\omega}$, where $\hat{\omega}$ is given in (2.2), and then the resulting centered residuals are used to compute the test statistics.

Concerning the errors terms, we set $u_t = h_t \epsilon_t$ with $\epsilon_t$ iid(0,1), accordingly to **A1**. Of course under the null hypothesis we have $\epsilon_t \sim \mathcal{N}(0,1)$. Under the alternative hypothesis $\epsilon_t = \cos(\delta) v_t + \sin(\delta) w_t$, with $0 < \delta \leq \frac{\pi}{2}$. In such a case $v_t \sim \mathcal{N}(0,1)$, and we examine three alternatives:

a) $(\sqrt{2} w_t + 1) \sim \chi^2_1$

b) $w_t \sim$ *Laplace*(0,1)

c) $w_t \sim t_9$.

The random variables $v_t$ and $w_t$ are mutually independent. On the other hand, two cases are investigated for the variance structure. The $h_t$'s are set $h_t = 1$, when the homoscedastic case is studied. For the heteroscedastic case, we take

$$h_t = 1 + 2\exp(t/n) + 0.3(1 + t/n)\sin(5\pi t/n + \pi/6). \tag{3.1}$$

In (3.1) the variance exhibits a global monotone behavior together with a cyclical/seasonal component that is common in macroeconomic data (see e.g. Trimbur, and Bell (2010) for seasonal effects in the variance).

### 3.1.1 Empirical size

The outputs obtained for the $T_{cv}$ test using the Gaussian kernel are first analyzed. The results with $p = 1$ and $a_{01} \in \{0.05, 0.5, 0.95\}$ are given in Table 1 for the homoscedastic case, and in Table 2 for the heteroscedastic case. Table 3 displays the relative rejection frequencies for large $p$ ($p = 4$ with $a_{01} = 0.7$, $a_{02} = -0.5$, $a_{03} = 0.3$ and $a_{04} = -0.2$). Noting that macroeconomic time series with noticeable heteroscedasticity are relatively large but smaller than $n = 400$ in general, a special emphasis will be put on interpreting results for samples $n = 100, 200, 400$. Since $N = 1000$ processes are simulated, and under the hypothesis that the finite sample size of a given test is 5%, the relative rejection frequencies should be between the significant limits 3.65% and 6.35% with probability 0.95. The outputs outside these confidence bands will be displayed in bold type.

From Table 1, it appears that the $T_{cv}$ is oversized for small samples ($n = 100$). This could be explained by the fact that this test is too much sophisticated for the homoscedastic case. When the samples are increased the relative rejection frequencies become close to the 5% ($n = 200$, $n = 400$ and $n = 800$). We also found some slight size distortions for the $T_{st}$ test when $n = 100$. It turns out from Table 1 that the $T_{boot}$ test has good results for all the samples. Of course if there is no evidence of heteroscedasticity, the simple $T_{st}$ should be used. However Table 1 reveals that in case of doubt, the use of the $T_{boot}$ is a good alternative. On the other hand in the heteroscedastic case, it is seen from Table 2 that the $T_{st}$ test fails to control the type I error as $n \to \infty$. This was expected from (2.6). Next it seems that the relative rejection frequencies of the $T_{cv}$ test are somewhat far from the nominal level 5% for small samples. From Table 2 it also emerges that the $T_{boot}$ control reasonably well the type I error. Finally note that the results for $p$ large, in Table 3, lead to the same conclusions as those drawn for Tables 1 and 2. Therefore it turns out that the $T_{boot}$ gives a substantial improvement for samples that are typical for heteroscedastic macroeconomic variables. Nevertheless from the outputs of Tables 1, 2 and 3, it seems that

the improvements of the $T_{boot}$ in comparison to the $T_{cv}$ become slight as $n \to \infty$. For this reason if high frequency time series are analyzed, with typically $n \gg 1000$, the $T_{cv}$ should certainly be preferred to the computationally intensive $T_{boot}$.

In general it is important to point out that the bandwidth must be carefully selected to ensure a good implementation of the $T_{boot}$ and $T_{cv}$ tests. Its turns out from our experiments that selecting the bandwidth by cross-validation leads to relatively correct results. Indeed we found that the rejection frequencies of the $T_{cv}$ converge to the 5%, and that the rejection frequencies of the $T_{boot}$ remain close to the nominal level in such a case. However other choices can deteriorate the control of the type I errors. For instance let us consider the $T_f$ test which consists in correcting the heteroscedasticity, but with fixed bandwidth as $\gamma(\hat{\sigma}^2/n)^{0.2}$, where $\hat{\sigma}^2$ is the sample variance and $\gamma$ is a constant. The corresponding bootstrap test will be denoted by $T_{f,boot}$. Here the normal kernel is kept, and we only study the heteroscedastic case. The results given in Table 4 show that the rejection frequencies are strongly affected by this way of selecting the bandwidth.

In the previous experiments the results were obtained using the Gaussian kernel. Such a kernel verifies $K(z) > 0$ for all $x \in \mathbb{R}$. Noting that the most commonly used kernels in the literature are either verifying $K(z) > 0$ for all $z \in \mathbb{R}$, or $K(z) > 0$ only for $|z| < 1$, the finite sample behavior of the $T_{cv}$ and $T_{boot}$ tests are studied using the Epanechnikov kernel. From Table 5 some changes can be noted when comparing the results obtained with the Epanechnikov kernel to those obtained with the Gaussian kernel. However the general conclusions remain the same, as we also note a good control of the type I error for the $T_{boot}$, while the relative rejections frequencies of the $T_{cv}$ become close to the nominal level as $n \to \infty$ in both cases.

### 3.1.2 Empirical power

Now we turn to the analysis of the behavior of the tests under the alternative hypothesis. The homoscedastic and heteroscedastic cases are again considered, but with the alternatives a), b) and c). Note that the alternative c) is the most difficult to detect, as its skewness is equal to zero and its kurtosis is relatively close to 3. On the other hand the alternative a) is the easiest alternative to detect as both skewness and kurtosis are quite different from those of the Gaussian distribution. The sample size $n = 200$ is fixed and recall that the parameter $\delta$ defines the departures from the null hypothesis. We set $p = 1$ with $a_0 = 0.5$. The outputs are given in Figures 1 and 2.

A quick look at Figure 1 could suggest that the $T_{cv}$ is slightly more powerful than the $T_{boot}$ and $T_st$ tests in the standard case. Actually the comparison made in such a case is not fair because the actual level of the $T_{cv}$ seems greater than the 5% level. It also emerges that the $T_{boot}$ test does not suffer from a lack of power in comparison to the $T_{st}$. In a similar way for Figure 2 note that the $T_{cv}$ and $T_{st}$ tests are oversized, so a direct comparison with the $T_{boot}$ under the alternative is not fair. At least it can be seen from Figures 1 and 2 that the slopes of the three detection curves seem similar. This can suggest that the abilities of the three tests for detecting the alternatives are essentially identical. In conclusion it turns out that the $T_{boot}$ improves the distribution analysis, in the sense that it ensures a good control of the type I error, but without entailing noticeable loss of power.

## 3.2 Real data analysis

The inflation measures data are commonly used to analyze macroeconomic facts. Reference can be made to the numerous empirical papers studying the relation between price levels and money supply (see e.g. Jones and Uri (1986)). On the other hand inflation is of great importance in finance, as many central banks adjust their interest rates in view of targeting a certain inflation level. Accordingly, constructing valid confidence intervals for inflation forecasts may be often crucial. In such a kind of investigations clearly the distributional

analysis can help to build a model for the data. In a stationary setting, authors aimed to detected ARCH effects assessing asymmetry and/or leptokurticity in inflation variables following Engle (1982) (see Broto and Ruiz (2008p22) among others). In the same way, it is reasonable to think that a test for normality taking into account the time-varying variance, can help to choose between a deterministic specification, as in **A1**, and the case where in addition to unconditional heteroscedasiticity, second order dynamics are present (as in the case of spline-GARCH processes introduced by Engle and Rangel (2008)). In other words, once the unconditional heteroscedasiticity is removed from $u_t = h_t\epsilon_t$, the JB tests can help to decide whether ARCH effects are present or not in $(\epsilon_t)$.

In this part we will study the normality of the log differences of the quarterly GDP implicit price deflators for the U.S., Korea and Australia from 10/01/1983 to 01/01/2017 ($n = 132$). More precisely we use $y_t = 100 \log(GDP_t/GDP_{t-1})$. The data can be downloaded from the webpage of the research division of the federal reserve bank of Saint Louis: https://fred.stlouisfed.org. The studied variables plotted in Figure 3 seem to show cyclical heteroscedasticity. In the case of Korea we can suspect a global decreasing behavior leading to a stabilization after the Asian crisis. The times series are first filtered according to (2.1). The non correlation of the residuals is tested using the adaptive portmanteau test of Patilea and Raïssi (2013). On the other hand we applied tests for second order dynamics developed by Patilea and Raïssi (2014). The outputs (not displayed here) show that the hypothesis of no ARCH effects cannot be rejected. Hence the deterministic specification of the time-varying variance in **A1** seems valid. Once the linear dynamics of the series seem captured in an appropriate way, the tests considered in this paper are applied to the residuals. The results are given in Table 6. When the null hypothesis of normality is rejected at the 5% level, the p-value is displayed in bold type. It emerges that the outputs of the $T_{boot}$ test are in general clearly different from those of the $T_{cv}$ and $T_{st}$ tests. The p-values of the $T_{cv}$ are all lower than those of the $T_{boot}$. Note that in the case of the U.S. GDP implicit price deflator, the difference between the $T_{boot}$ on one hand, and the $T_{st}$, $T_{cv}$ tests on the other

hand, lead to different conclusions. In view of the outputs obtained from the simulations experiments, it is reasonable to decide that the normality assumption cannot be rejected for the U.S. data. It is likely that rejecting normality will suggest more sophisticated models, and could entail misspecifications for the confidence intervals of the forecasts by fitting a heavy tailed distribution to the U.S. data.

# 4 Conclusion

When a standard Jarque-Bera test is applied to the residuals of a linear time series model, the observed $p$-value can be low because the errors do not have a normal distribution, but also possibly because the errors are simply unconditionally heteroscedastic. In short the classical JB test fail to distinguish between normally and non normally distributed heteroscedastic time series. In time series econometrics, the spurious rejection of the null hypothesis of a time series could lead to consider unnecessary nonlinear structures.

In this paper the approach that consists in correcting the data from the unconditional heteroscedasticity previously of applying the JB test is investigated. It is found that the corresponding test is fully justified for large samples, provided that the bandwidth is selected in a correct way. In particular its turns out that choosing the bandwidth using the cross-validation criterion lead to an improvement of the control of the type I error as the sample size is increased. However for small samples we observed severe size distortions for the above test. As a consequence a parametric bootstrap JB test is proposed to solve this problem. It is emerges from our simulation study that the resulting test constitutes an adequate tool for analyzing the normality of unconditionally heteroscedastic macroeconomic time series.

# References

Broto, C., and Ruiz, E. (2008) Testing for conditional heteroscedasticity in the components of inflation. Working document, banco de España.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Annals of Statistics* 25, 1-37.

Drees, H., and Stărică, C. (2002) A simple non stationary model for stock returns. Preprint. Universität des Saaland.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987-1007.

Engle, R.F., and Rangel, J.G. (2008) The spline GARCH model for unconditional volatility and its global macroeconomic causes. *Review of Financial Studies* 21, 1187-1222.

Fiorentini, G., Sentana, E., Calzolari, G. (2004) On the validity of the Jarque-Bera normality test in conditionally heteroscedastic dynamic regression models. *Economic Letters* 83, 307-312.

Francq, C., and Zakoïan, J-M. (2010) *GARCH models : structure, statistical inference, and financial applications.* Wiley, Chichester.

Fryźlewicz, P. (2005) Modelling and forecasting financial log-returns as locally stationary wavelet processes. *Journal of Applied Statistics* 32, 503-528.

Jarque, C.M., and Bera, A.K. (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6, 255-259.

Jones, J.D., and Uri, N. (1986) Money, inflation and causality (another look at the empirical evidence for the USA, 1953-84). *Applied Economics* 19, 619-634.

Lee, T. (2012) A note on Jarque-Bera normality test for ARMA-GARCH innovations. *Journal of the Korean Statistical Society* 41, 37-48.

Lee, S., Park, S., and Lee, T. (2010) A note on the Jarque-Bera normality test for GARCH innovations. *Journal of the Korean Statistical Society* 39, 93-102.

Mikosch, T., and Stărică, C. (2004) Stock market risk-return inference. An unconditional non-parametric approach. Research report, the Danish national research foundation: Network in Mathematical Physics and Stochastics.

Patilea, V., and Raïssi, H. (2012) Adaptive estimation of vector autoregressive models with time-varying variance: application to testing linear causality in mean. *Journal of Statistical Planning and Inference* 142, 2891-2912.

Patilea, V., and Raïssi, H. (2013) Corrected portmanteau tests for VAR models with time-varying variance. *Journal of Multivariate Analysis* 116, 190-207.

Patilea, V., and Raïssi, H. (2014) Testing second order dynamics for autoregressive processes in presence of time-varying variance. *Journal of the American Statistical Association* 109, 1099-1111.

Raïssi, H. (2015) Autoregressive order identification for VAR models with non-constant variance. *Communications in Statistics: Theory and Methods* 44, 2059-2078.

Sensier, M., and van Dijk, D. (2004) Testing for volatility changes in U.S. macroeconomic time series. *Review of Economics and Statistics* 86, 833-839.

Trimbur, T.M., and Bell, W.R. (2010) Seasonal heteroscedasticity in time series data: modeling, estimation, and testing. In W. Bell, S. Holan, and T. Mc Elroy (Eds.), *Economic Time Series: Modelling and Seasonality*. Chapman and Hall, New York.

Wasserman, L. (2006) *All of Nonparametric Statistics*. Springer, New-York.

Yu, H. (2007) High moment partial sum processes of residuals in ARMA models and their applications. *Journal of Time Series Analysis* 28, 72-91.

# Tables and Figures

Table 1: Empirical size (in %) of the studied tests for normality. The homoscedastic case with $p = 1$.

| $a_0 = 0.05$ | | | | |
|---|---|---|---|---|
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | 3.7 | 3.8 | 5.3 | 5.0 |
| $T_{cv}$ | **7.1** | 5.8 | 6.1 | 5.4 |
| $T_{boot}$ | 5.4 | 4.5 | 5.5 | 4.9 |
| $a_0 = 0.5$ | | | | |
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | **3.4** | 3.8 | 5.7 | 4.9 |
| $T_{cv}$ | **6.8** | 5.9 | **6.4** | 5.6 |
| $T_{boot}$ | 4.9 | 4.7 | 5.4 | 5.4 |
| $a_0 = 0.95$ | | | | |
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | **3.5** | 3.8 | 5.1 | 5.0 |
| $T_{cv}$ | **6.6** | 5.9 | **6.4** | 5.4 |
| $T_{boot}$ | 4.0 | 5.1 | 5.6 | 5.1 |

Table 2: Empirical size (in %) of the studied tests for normality. The heteroscedastic case with $p = 1$.

| $a_0 = 0.05$ | | | | |
|---|---|---|---|---|
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | 5.9 | **9.7** | **11.5** | **15.0** |
| $T_{cv}$ | **8.5** | **7.9** | **7.7** | **6.7** |
| $T_{boot}$ | 4.5 | 5.9 | **6.5** | 5.9 |
| $a_0 = 0.5$ | | | | |
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | **6.5** | **10.1** | **11.7** | **15.3** |
| $T_{cv}$ | **8.4** | **8.1** | **8.2** | **6.6** |
| $T_{boot}$ | 4.2 | 5.5 | 6.3 | 6.3 |
| $a_0 = 0.95$ | | | | |
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | **7.6** | **11.3** | **11.3** | **15.5** |
| $T_{cv}$ | **8.4** | **9.4** | **7.9** | **6.9** |
| $T_{boot}$ | 3.7 | 6.1 | 6.2 | 5.8 |

Table 3: Empirical size (in %) of the studied tests for normality. The lag length is $p = 4$.

| The homoscedastic case | | | | |
|---|---|---|---|---|
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | 5.0 | 3.9 | 4.9 | 5.3 |
| $T_{cv}$ | **7.9** | 6.2 | 6.1 | 5.7 |
| $T_{boot}$ | 5.6 | 4.1 | 5.0 | 5.1 |
| **The heteroscedastic case** | | | | |
| $n$ | 100 | 200 | 400 | 800 |
| $T_{st}$ | **7.7** | **8.3** | **12.0** | **16.1** |
| $T_{cv}$ | **10.4** | **7.9** | **7.3** | **6.6** |
| $T_{boot}$ | 5.0 | 4.8 | 5.5 | 5.7 |

Table 4: Empirical size (in %) of the $T_{cv}$ and $T_{boot}$ tests for normality with fixed bandwidth. The heteroscedastic case with $p = 1$, $a_0 = 0.4$.

| $\gamma$ | 1 | | 1.5 | |
|---|---|---|---|---|
| $n$ | 100 | 200 | 100 | 200 |
| $T_f$ | **11.3** | **14.0** | **12.0** | **14.3** |
| $T_{f,boot}$ | **6.8** | **10.2** | **7.6** | **11.1** |

Table 5: Empirical size (in %) of the studied tests for normality. The lag length is $p = 1$, $a_0 = 0.5$. The Epanechnikov kernel is used for variance estimation on the left, while the corresponding results for the Gaussian kernel are recalled on the right.

| The homoscedastic case | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | 100 | 200 | 400 | 800 | 100 | 200 | 400 | 800 |
| $T_{cv}$ | **6.9** | 5.7 | **6.4** | 5.5 | **6.8** | 5.9 | **6.4** | 5.6 |
| $T_{boot}$ | 5.1 | 4.4 | 5.4 | 5.2 | 4.9 | 4.7 | 5.4 | 5.4 |
| The heteroscedastic case | | | | | | | | |
| $n$ | 100 | 200 | 400 | 800 | 100 | 200 | 400 | 800 |
| $T_{cv}$ | **12.6** | **9.6** | **8.7** | **6.4** | **8.4** | **8.1** | **8.2** | **6.6** |
| $T_{boot}$ | 3.8 | 5.4 | 5.3 | 5.5 | 4.2 | 5.5 | 6.3 | 6.3 |

Table 6: The p-values (in %) of the tests for normality for GDP implicit price deflators for the U.S., Korea and Australia.

| | U.S. | Korea | Australia |
|---|---|---|---|
| $T_{st}$ | **3.8** | 16.4 | 50.9 |
| $T_{cv}$ | **2.3** | 82.0 | 21.0 |
| $T_{boot}$ | 8.2 | 87.0 | 49.0 |

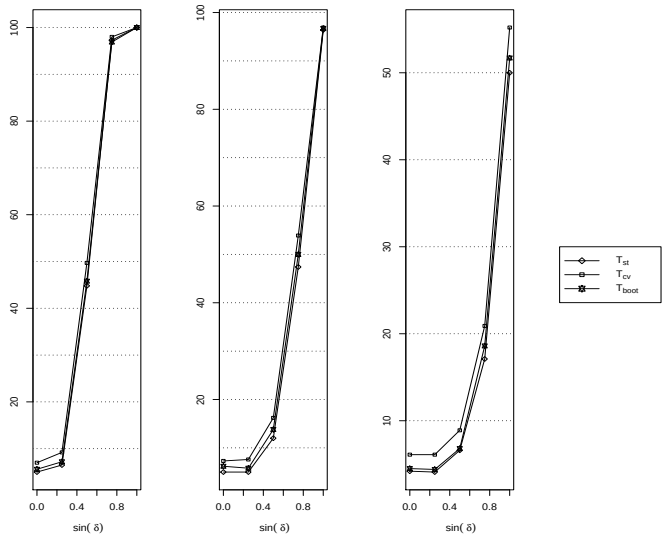



Figure 1: Empirical power (in %) of the $T_{st}$, $T_{cv}$ and $T_{boot}$ tests in the homoscedastic case. The chi-square alternative is displayed on the left panel, the Laplace alternative is displayed on the middle panel, while the Student alternative is given on the right panel.

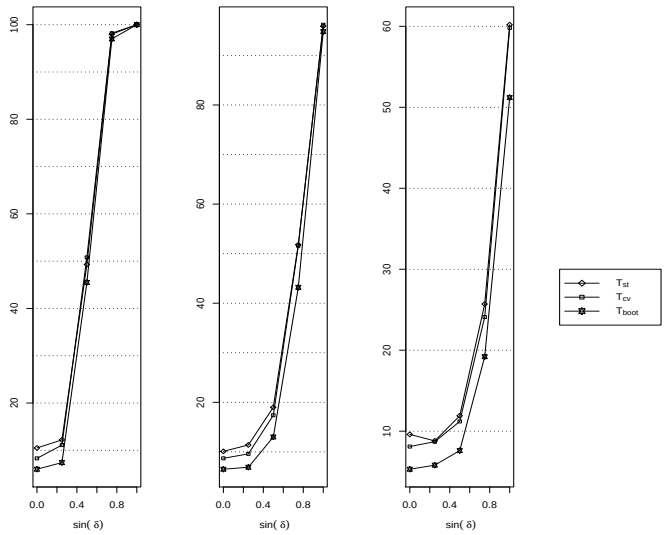



Figure 2: Empirical power (in %) of the $T_{st}$, $T_{cv}$ and $T_{boot}$ tests in the heteroscedastic case. The chi-square alternative is displayed on the left panel, the Laplace alternative is displayed on the middle panel, while the Student alternative is given on the right panel.

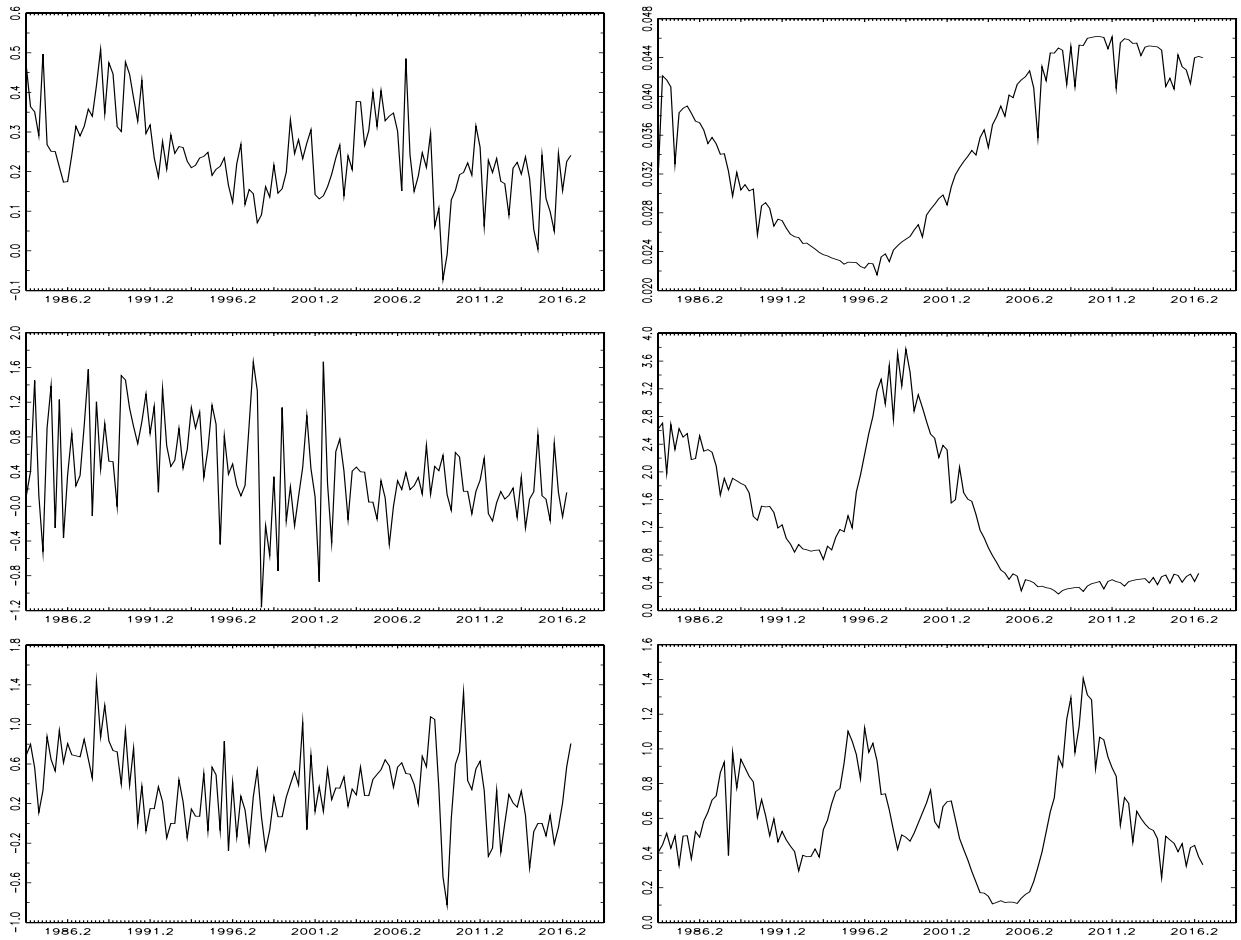



Figure 3: The log differences of the quarterly U.S. (top left panel), Korean (middle left panel) and Australian (bottom left panel) GDP implicit price deflators from 10/01/1983 to 01/01/2017 ($n = 132$). The corresponding estimations of the innovations variances are on the right. Data source: The research division of the federal reserve bank of Saint Louis, fred.stlouisfed.org.

# Regional Development and Inequalities in Latin American Countries: Econometric Analysis

Evgeniya Muzychenko

Moscow, Russia
Jennymuz123@gmail.com

**Abstract.** For centuries, problems of socio-economic development of Latin American countries, in particular, the issues of poverty and income inequality have been a central topic in the studies. At the same time, the aspect of regional (sub-national) development of Latin America appears to be quite new and understudied as there is a lack of literature on this topic. Therefore, the aim of this research is to examine the dynamics of regional development from 2000 to 2016 and determine the main factors of regional growth of Latin America. In order to analyze regional development and assess the dynamics, the classification of selected countries' regions will be conducted. The results of the study have shown that despite the improvements in multiple socio-economic indicators, the problem of regional inequality is still relevant for the countries in the region, including the most developed ones. Moreover, some factors of territorial development were detected and the hypothesis of industrial regional growth was confirmed.

**Keywords:** Latin America, socio-economic development, inequalities, regional development

## 1 Introduction

For many centuries the issues of poverty, income inequality and other socio-economic discrepancies in emerging countries have been a primary concern for economists and politicians and countries of Latin America and the Caribbean have not become an exception. Although recently countries of the region have made great strides towards economic development and considerable progress has been made in combating numerous socio-economic imbalances, there is still a huge room for improvement for Latin America and the Caribbean. The Gini coefficient as an indicator of inequality remains much higher than that of the developed countries and, on average, 25 percent of Latin American population is living below the poverty line [1, 2].

This research focuses on the recent changes and dynamics of socio-economic development of Latin American countries with a certain emphasis on the regional coherence within the states. Despite the fact that there are numerous academic articles on the topic of socio-economic issues in Latin American countries, a research gap still exists in the context of the regional development within the states and in terms of regional economic development convergence and divergence. Hence, in the present study, first, we aim to address this gap and assess the factors of regional development in Latin American countries. Moreover, through the multi-factor analysis we will attempt to identify the stage of economic development of the chosen countries.

This research will be contributive for scientific community as a basis for broader analysis of issues of regional and territorial coherence in Latin American countries.

## 2 Literature Review and Problems Background

Throughout the history of economic development, many scientists and researchers tried to determine the factors that can lead the country to economic prosperity and sustainable growth. Nowadays there are a number of theories, from classic to neo-institutional, which explain positive tendencies of economic development from different perspectives [3, 4, 5]. It is a quite well known fact that there are three main factors of production: land, labour and capital. While economic thinking was developing, scientists started to look for the new ways to explain why even though some countries have those resources their levels of development differ. For instance, a widely famous Solow model [3], apart from traditional factors, is as well based on investments into technological advances. The presence of technological factor in the model might have explained the difference in economic inequality between the countries but later it was proven that technology has a diminishing return to scale which means that at some point marginal utility of technological investment will be equal to 0, therefore it cannot be the factor of constant growth [3, 4]. Along with technological advances, another important factor, which is supposed to strengthen economic development, seems to be a highly efficient human capital. The endogenous theory of Paul Romer is based on the prerequisite that investments in human capital are characterized by permanently increasing return to scale due to accumulated knowledge and skills which raises labour productivity and final output [3, 5].

However, none of traditional theories of economic growth is able to explain the differences in the level of socio-economic development of the countries to the full extent, which means that factors of economic prosperity are not limited to labour and capital. Another possible factor that can cause inequalities between and within countries is regional coherency. There are some studies delving into the importance of equal regional development to the sustainable economic development of the country. More specifically, this aspect seems to be essential to be considered in the states with a big area and complex territorial and administrative division.

Existing literature on this issue shows that a number of authors [7, 8, 12, 13, 14] agree on the positive effect of even territorial development on socio-economic prosperity of the country whereas inequalities trigger such economic discrepancies as income inequality and poverty [7, 13, 15]. For example, Scott points out that there is a need for so-called “regional push”, some sort of government intervention and programs in the sub-national territories, to foster the development of “backward” regions and thus of a country itself [15]. That is particularly important as the studies show that the more prosper the economy is, the more evident all the regional discrepancies become [7, 15].

Although there is a series of studies on the topic of socio-economic and regional development in general, little is known about the dynamics of regional development in Latin American countries. The focus of social policy in Latin American countries is still concentrated in the transfers to the most disadvantages groups of the population in

order to combat poverty and reach an equal distribution of income and not in the regional growth promotion.

However, some studies on this subject show that regional discrepancies are relevant to Latin American countries, even the most developed ones. First of all, the existence of regional disproportions manifests itself in the income gap between the most and the least developed regions of the country. For examples, in Mexico in 2010 the largest GRP (Gross Regional Product - GRP) per capita surmounted the smallest one y 9.3 times. In Argentina, this gap reached 7.35, in Brazil – 7.35 [13]. More specifically, the differences between the regional development levels are seen in the human development index that varies a lot depending on the economic conditions in the region demonstrating that there is a gap in the quality of live among the regions [16].

## 3 Methodology

First, in order to find the main characteristics and identify the recent changes in regional development of Latin American countries we are going to divide the regions of four Latin American countries (Brazil, Chile, Columbia and Mexico) into four groups:

1. Group 1 (G1) – capitals and financial centers;
2. Group 2 (G2) – developed regions;
3. Group 3 (G3) – average-developed regions;
4. Group 4 (G4) – less developed regions.

According to our initial hypothesis, *developed regions* mostly include the regions with high share of services and manufacturing industry. *Average-developed regions* are represented by industrial centers with extractive industry. *Less developed regions* are in their majority are agricultural areas.

Second, methodological grounds to this study consist of multi-factor regression model. The model includes both growth variables and 3 dummy-variables deriving from the economic nature of the region and can be presented as follows:

*regional growth = c – a1 initial level + a2 country growth + a3 dummy capital + a4 dummy industry – a5 dummy services + e*, where

- *regional growth* illustrates how many times GRP (PPP – purchasing power parity) per capita increased or decreased from 2000 to 2016;
- *initial level* stands for GRP (PPP) per capita in 2000;
- *country growth* represents the country's average growth rate from 2000 to 2016;
- *dummy capital* is set for the capitals of the chosen countries, i.e. Mexico, Bogota DC, Santiago and Brasilia;
- *dummy industry* states regions with high presence of industry in the economic structure of the region;
- *dummy service* – regions with the predominance of tertiary sector in the economy;

- $e$ – standard error.

This analysis is expected to confirm the hypothesis that Latin American countries are on the industrial stage of economic development, therefore the regional GDP growth is concentrated mostly in the areas with prevalence of industrial sector in the economy of the given region.

## 4 Data description

Regional statistics in each of the four countries is represented in national currencies, which makes it difficult to compare and analyze the data. That is why, in order to conduct our regressive analysis and make the countries comparable we need to convert the existing data to international dollars.

The method we use to convers GRP of given countries is based on the allocation of GDP among the regions in proportion with the share of GRP in the aggregate GRP of a country. Per capita indicators are calculated by dividing imputed GRP by the population of the given region. This method allows us to compare the regional data for different countries due to the elimination of the variations in the aggregate GRP and GDP of the country.

Moreover, the GDP will be re-calculated by the means of the parity of purchasing power (PPP), which should eliminate the influence of exchange rates and make the countries even more comparable to each other [17].

One more difficulty to be added is that regional statistics in Latin America is presented in quite a limited volume, which makes it hard to estimate a more complex set of variables in the model. By and large, the data published by national statistics agencies is restricted to GRP and its components (industry, trade, tourism, financial services, etc.) or sectors of the economy (primary, secondary and tertiary). That fact, to a certain extent, has determined the choice of depending variables in our model.

The timeline of the research will be limited to the year of 2016 as national statistics agencies provide the data with the lag of 1,5-2 years (by now the latest comparable figures are dated the year of 2016).

## 5 Results

After classifying the regions into four groups, we have found several peculiarities of regional development of the chosen countries:

1. Regional discrepancies can be primarily tracked looking into the shares of the most and the least developed regions in the aggregate income of the country. For instance, in Columbia the contribution of 5 most rich regions in the GDP accounts for more than 60 percent whereas 10 least developed departments make up only 2 percent. The same pattern can be observed in all the countries that were under our review. The most even distribution of income can be found in Mexico, but still not enough to be considered

equal. These findings one more time confirm that regional inequalities are relevant to Latin American countries.

2. In each country, except Brazil, financial centers coincide with capitals of the countries. In Brazil, San-Paulo and Rio-de-Janeiro were also attributed to the first group (capitals and financial centers) as they are classifies as international financial centers by the rating of Long Finance [18].
3. The considerable number of citizens in those countries live in the capitals and financial centers (see Table 1). In Chile, the share of population living in Santiago makes up 40,5 percent. The least populous capital was noted in Mexico where this proportion constitutes 8 percent but if taking into account the whole state Mexico, the share will increase to almost 25 percent. Similar studies of BRICS and EAEU show that this situation is typical for developing countries in general because migration flows are directed into capitals with more opportunities and higher standards of living [19].

**Table 1**. Population share (%) and average GRP (PPP) per capita (thousand dollars) by regional development groups, 2000 and 2016

| | | ***Brazil*** | | ***Columbia*** | | ***Mexico*** | | ***Chile*** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Pop-ula-tion** | **GRP per cap-ita** | **Popu-la-tion** | **GRP per cap-ita** | **Popu-la-tion** | **GRP per cap-ita** | **Pop-ula-tion** | **GRP per cap-ita** |
| *G1* | 2000 | 31,5 | 20,4 | 16,3 | 9,8 | 8,8 | 28,3 | 41,6 | 11,2 |
| | 2016 | 29,8 | 31,6 | 16,4 | 21,8 | 7,9 | 44,6 | 40,5 | 27,3 |
| *G2* | 2000 | 23,9 | 9,9 | 39,2 | 11,0 | 28,6 | 14,6 | 7,8 | 16,4 |
| | 2016 | 31,5 | 18,3 | 43,2 | 16,7 | 18,9 | 35,0 | 7,9 | 43,0 |
| *G3* | 2000 | 27,9 | 6,5 | 19,6 | 5,1 | 33,2 | 8,7 | 16,1 | 8,8 |
| | 2016 | 20,6 | 11,0 | 23,3 | 10,3 | 25,9 | 18,7 | 16,1 | 22,7 |
| *G4* | 2000 | 16,7 | 4,1 | 24,9 | 3,5 | 29,4 | 6,1 | 34,5 | 6,3 |
| | 2016 | 18,2 | 6,8 | 17,1 | 6,0 | 47,3* | 12,2 | 35,5 | 15,0 |

*the increase in the share of population of less developed states of Mexico deals with the transition of the state Mexico from the group of average-developed to the group of less developed regions in 2016

Source: compiled by author based on INEGI, Banco Central de Chile, DANE, IBGE, and INE

4. In three of the four countries, the highest per capita income was detected in the regions with the predominance of extracting industry. In Brazil, the leading positions were

taken by Brasilia and two financial centers Rio-de-Janeiro and San-Paulo. In general, developed regions in those countries represent the industrially developed territories.

5. The previous point is followed by the conclusion that unlike our initial hypothesis the development of the tertiary sector is not the necessary condition for a region to be highly developed in Latin America. Otherwise, in average, the most significant shares of services in the economy were found in the group of less developed regions. In Columbia and Brazil, it might be connected with the interventions of the governments in the less developed regions. In Mexico, to the influence of the government sector we can add the impact of touristic branch, which is quite developed in the states with low income.

Judging from our findings, we can make the conclusion that our initial hypothesis should be partly rejected as not the services were the trigger of economic growth of the most developed regions. This structure of economic activity is typical for developing countries that have not yet undergone the transformation to the post-industrial stage of development. In continuation, to confirm that Latin American countries are now on an industrial stage of economic development, we perform multi-factor regressive analysis with the dummy variables.

**Table 2**. Results of correlation analysis

| | *Regional growth* | *Initial level* | *Country growth* | *Dummy capital* | *Dummy industry* | *Dummy services* |
|---|---|---|---|---|---|---|
| *Regional growth* | 1 | | | | | |
| *Initial level* | -0,13 | 1 | | | | |
| *Country growth* | 0,23 | -0,13 | 1 | | | |
| *Dummy capital* | -0,07 | 0,41 | -0,03 | 1 | | |
| *Dummy industry* | 0,25 | 0,31 | 0,13 | -0,12 | 1 | |
| *Dummy services* | -0,31 | 0,00 | -0,11 | 0,24 | -0,38 | 1 |

Source: compiled by author based on INEGI, Banco Central de Chile, DANE, IBGE, and INE

From table 2 we can observe that correlations between the variables that we plan to include in the model are not high. The highest one between initial level of development and dummy set on capitals equals 0.41, as capitals tend to be the most developed high-income regions while those set on services and industries may include regions throughout the whole GRP per capita distribution (see Table 2). Low correlations between the factors and dependent variable means that there is no need to exclude any of them from the regressive analysis.

Having conducted multi-factor analysis, it was found out that all the key factors of regional economic growth that were included in the model (initial stage of regional economic development, average growth rate for the given period and dummy-variables) are statistically significant. The proposed equation with calculated coefficients (see Table 3) can be seen as follows:

*regional growth = 1.62 – 0.03 initial level + 0.34 country growth +*
*+ 0.32 dummy capital + 0.4 dummy industry – 0.33 dummy services + 0.64*

**Table 3**. Results of econometric analysis: the industrial stage of development

| **Regression statistics** | |
|---|---|
| *Multiple R* | 0,43 |
| *R-squared* | 0,18 |
| *Adjusted R-squared* | 0,14 |
| *Standard error* | 0,64 |
| *Observations* | 105 |

| | df | SS | MS | F | Significance F |
|---|---|---|---|---|---|
| *Regression* | 5 | 9,28 | 1,85 | 4,47 | 0,001 |
| *Residual* | 99 | 41,11 | 0,41 | | |
| *Total* | 104 | 50,39 | | | |

| | Coefficients | Standard error | t-statistics | P-Value | Lower 95% | Higher 95% |
|---|---|---|---|---|---|---|
| *Constant* | 1,62 | 0,45 | 3,57 | 0,00 | 0,72 | 2,52 |
| *Initial level* | -0,03 | 0,01 | -2,04 | 0,04 | -0,05 | -0,00 |
| *Country growth* | 0,34 | 0,22 | 1,57 | 0,12 | -0,09 | 0,77 |
| *Dummy capital* | 0,32 | 0,32 | 1,02 | 0,31 | -0,30 | 0,95 |
| *Dummy industry* | 0,39 | 0,19 | 2,11 | 0,04 | 0,02 | 0,77 |
| *Dummy services* | -0,33 | 0,14 | -2,29 | 0,02 | -0,62 | -0,04 |

Source: compiled by author based on INEGI, Banco Central de Chile, DANE, IBGE, and INE

The R-squared of the model is quite low along with a high standard error. However, considering that we are working with heterogeneous group of observations (105 regions of 4 countries with their differences in administrative and territorial division and the concentration of population in the regions) such a low R-squared and high standard error values can be predictable but the results of the model still might be statistically significant and meaningful.

So what impact do the chosen factors have on the economic development from the regional perspective? Primarily, we can see that there is a reverse influence of initial level of development of the country on the growth rates in the region. Although the connection between these two variables is not that strong and the coefficient equals only -0.03, it is still significant by the means of t-statistics; therefore, we can conclude that the initial level of GRP has certain impact on growth. As we can observe from our model, the less initial level of development is, the faster regional growth is which might signify that the concept of "catching-up" took place in Latin American region.

The results of econometric analysis also show that our suggestion about the industrial stage of development has its own grounds. As long as we see that there is a positive connection between the dummy on industrial regions and regional growth rate while dummy on services proves to have a reverse effect on regional growth, we can assume that Latin America is still on the early industrial stage (the most developed regions are the ones with extracting industry dominating).

The next question is what does the industrial stage imply? Firstly, we can see that Latin American economies are to a certain extent dependent on the raw materials and minerals. Secondly, this overreliance of economic growth on natural resources may interfere with further development of the countries. Thirdly, it does not allow Latin American countries to vanish regional inequalities, as those regions with the diversity and high concentration of commodities tend to grow more rapidly than the ones with the lack of resources.

Having analyzed the factors of regional growth, we have also estimated the dynamics and the changes in the distributions of GRP from 2000 to 2016. We testified the equality of that distribution by the means of standard deviation. The results show that income has become more unequally allocated between regions because of the rapid growth of several highly developed regions. The standard deviation declined only in Columbia but we still cannot imply that as a positive tendency as the decrease owes to a fall in GRP per capita of some developed regions in this country.

## 6 Conclusion

What conclusions can be drawn from analyzing regional development of Latin American countries?

First, we have been able to identify the factors of economic growth from the regional perspective. Thus, we have noted that there is a positive connection between industrial production and regional growth rate and reverse connection between the latter and the large share of services in the economy. It means that in terms of regional development Latin American continent remains to be on an industrial stage of development with the high relevance of extracting activity and consequently the dependence on commodity prices. At the same time, the intensity of these factors' influence was not very high, that is why, we do not deny that there are other factors that had a stronger influence on the regional growth.

Second, we consider that "catching up" takes place in Latin American countries considering regional development, which is evident by the negative coefficient in front of

the *initial level* variable. It means that since 2000 less developed regions have been growing at a faster rate than the most developed ones. However, we have notes that the gap between the most and the least developed regions in each country has widened which, by and large, is due to the "breakaway" of industrial centers such as Campeche in Mexico and Antofagasta in Chile.

Here we highlight the importance of the results obtained to determine the stage of development of Latin American countries through regional analysis and therefore the factors that influence regional growth the most.

Further, we intend to develop the analysis with a greater variety of indicators and try to implement an econometric model based on time series forecasting to provide the in-depth analysis of regional dynamics in Latin America.

## 7 References


1. López-Calva, L., & Lustig, N. (2010). Declining inequality in Latin America: A decade of progress? Washington, D.C.: Brookings Institution Press.
2. Gasparini, L., & Cruces, G. (2013). Poverty and inequality in Latin America: A story of two decades. Journal of International Affairs, 66(2), 51-63.
3. Barro, R. (1999). Economic growth (2nd Ed.). Cambridge: MIT Press.
4. McQuinn, K., & Whelan, K. (2007). Solow (1956) as a model of cross-country growth dynamics. Oxford Review of Economic Policy, 23(1), 45-62.
5. Rivera-Batiz, L., & Romer, P. (1991). Economic Integration and Endogenous Growth. The Quarterly Journal of Economics, 106(2), 531-555.
6. ECLAC. (2015). Inclusive Social Development. Santiago: United Nations.
7. Buitelaar, R., Perico, R., Lira, I., & Perez, L. Estrategias y Politicas Nacionales para la Cohesion Territorial. Santiago de Chile: Naciones Unidas.
8. Commission of the European Communities. (2008). Green Paper on Territorial Cohesion: Turning territorial diversity into strength. Brussels: Commission of the European Communities.
9. ECLAC. (2018). Social Panorama of Latin America. Santiago: United Nations.
10. Savoie, D. (1992). Regional economic development reconsidered. In Regional Economic Development: Canada's Search for Solutions (2nd Edition). London: University of Toronto Press.
11. ECLAC. (2017). Linkages between the Social and Production Spheres: Gaps, Pillars and Challenges. Santiago: United Nations.
12. Teitz, M. (2012). Regional Development Planning. In Sanyal B., Vale L., & Rosan C. (Eds.), Planning Ideas That Matter: Livability, Territoriality, Governance, and Reflective Practice. Cambridge: MIT Press.
13. Cuadrado-Roura, J.R., & Aroca, P. (2013). Regional Problems and Policies in Latin America. N.Y.: Springer.
14. Martin, R., & Sunley, P. (1998). Slow Convergence: The New Endogenous Growth Theory and Regional Development // Economic Geography, 74, 201-227.
15. Scott, A. J. (2002). Regional Push: Towards a Geography of Development and Growth in Low-and Middle-Income Countries. Third World Quarterly, 23, 137-161.
16. Subnational Human Development Index (SD-2019). Retrieved from: https://globaldatalab.org/shdi/
17. Grigoryev, L. M., & Pavlyushina, V. A. (2018). Social inequality in the world: The tendencies of 2000-2016. Voprosy Ekonomiki, 10, 1-24.
18. The Global Financial Centers Index. Retrieved from: https://www.long-finance.net/programmes/financial-centre-futures/global-financial-centres-index/gfci-25-explore-data/gfci-25-rank/
19. Pavlyushina, V., Brilliantova, V., & Bondarenko, K. (2018). BRICS countries: Region's classification. Byulleten' o tekuschih tendentsyyah mirovoy economiki, 34, 1-20.

# STRUCTURAL STABILITY OF INFINITE-ORDER REGRESSION

Abhimanyu Gupta[1] and Myunghwan Seo[2]
1.-University of Essex (United Kingdom); 2.-Seoul National University (South Korea)

## Abstract

We develop a class of tests for the structural stability of infinite order regression models, when the time of a structural change is unknown. Examples include the infinite order autoregressive model, the nonparametric sieve regression and many others whose dimensions grow to infinity. When the number of parameters diverges, the traditional tests such as the supremum of Wald, LM or LR statistic or their exponentially weighted averages diverge as well. However, we show that a suitable transformation of these tests converges to a proper weak limit as the sample size n and the dimension p grow to infinity simultaneously. In general, this limit distribution is different from the sequential limit, which can be obtained by increasing the order p of the standardized tied-down Bessel process in Andrews (1993). More interestingly, our joint asymptotic analysis discovers that the joint asymptotic distribution depends on a higher order serial correlation. We also establish a weighted power optimality property of our tests under certain regularity conditions. A new result on partial sums of random matrices is established. We examine finite-sample performance in a Monte Carlo study and illustrate the test with a number of empirical examples..

# Customers of Future: How do They Spend their Bitcoins

Huber Nieto-Chaupis

Universidad Autonoma del Perú
Programa de Ingenierí a de Sistemas
hubernietochaupis@gmail.com

**Abstract.** Once the Bitcoins are established in the worldwide economy, from the point of view of markets would appear the question: what are the products that a potential customer would acquire with Bitcoins. Particularly developing economies might have not clear what market or product follow to. In this paper we present a Monte Carlo simulation that allows us to have a rough idea about the preferences of future customers. Thus, our results would indicate that electronics might be a favorable market for Peruvians with a confidence of order of 97%.

## 1 Introduction

Recently, is has been suggesting the usage of the so-called bitcoin that is defined as a kind of cryptocurrency as a unit of money that in the same manner as cash money does it, the sell and buy can be achievable without any prejudice [1]. Although most of the operations seem to be applicable by using an Internet network, the fundamental concepts of economy variables and theory of markets, apply well. In contrast of the existence of a Bank of Reserve that guarantees the value of the paper money, cryptocurrency [2][3] only needs of a peer-to-peer connection through a Internet network. In this paper, we focus on the prospective and massive usage of bitcoin in emergent economies such as Peru, in South American. Nowadays the equivalence in Peruvian currency 1 Bitcoin = 17 Peruvian soles. Being Peru one of the countries with a sustainable economy and an annual growth of order of 3.8The Peruvian economy has experienced various jumps along its pathway to be considered meanwhile as stable economy in the region. Furthermore, the robustness of the internal economy where there is a strong predominance of real state and the aperture of middle and large size malls, all this is seen as an interesting scenario to use Bitcoin as unit of currency instead the Peruvian sol. It is noteworthy that Bitcoin dynamics in a economy like Peru it might be reflected and projected to similar countries like Chile, Ecuador and Uruguay that demonstrating an interesting sustainability constitute solid markets to the implementation of modern currencies such as Bitcoins.

Thus, by considering the replication of this study, the implementation of a mathematical model appears to be necessary to make sustainable simulations

[4]. Once the mathematical model is proposed the usage of an advanced algorithm is mandatory to test the dynamics of a prospective dynamics of Bitcoin in Peru [5]. While fluxes of capitals might be well described by stochastic, another angle is that the commerce transactions would depend on the decision to perform any invest. We use the Monte Carlo method to simulate the different commerce transactions that might be well adjusted to the Bitcoin dynamics. This is because stochastics governs the rules of using Bitcoin along the diverse actions of buy and sell. The main Monte Carlo action simplified in the step Accept or Reject is used throughout the algorithm. Mainly this action is associated to the decision to perform any action towards, otherwise it is left and not any action is again performed unless that it is of fully satisfaction of the customer [6]. Essentially, we focus in the different actions that characterizes the ongoing Peruvian economy: real state, housing, motorized vehicles and credits to people. Certainly, the applicability of the Bitcoins is strongly limited to be continuously engaged to an Internet network. It might be perceived as a disadvantage in the sense that the lack or lost of Internet signal is seen as a questionable issue inside the Bitcoin dynamics. In fact, the usage of Internet networks would have to be adjusted to the requirement of the unstoppable transactions with Bitcoins. This seen from the angle of the probabilities, there is a certain probability for each Bitcoin transaction to be aborted or discarded, in according to the Quality-of-Service of the network. This is understood in terms in critic variables inside of the territory of the Telecommunications. In other words, the finance transactions are entirely dependent on the QoS of network. Thus, more than a fully independent process, the Bitcoin dynamics if governed by Probability Distribution Functions, that would determine the validity of nullity of transactions.

Despite that Bitcoin dynamics is entirely determined by people that require to accomplish a commerce operation, it is open also to other agents that can be for example:

(i) Intelligence Artificial Agents

(ii) Intelligent Machines

(iii) Advanced Commerce Algorithms

(iv) Any eGovernment

In fact, the existent intelligent systems or so-called autonomous machines with capabilities to carry out decisions on buy or sell are also potential candidates to be a well-defined customer. Even more the electronic capability makes them with enough artificial criterion to accomplish in parallel to the action of buy or sell, the encryption of the nominal Bitcoin values.

The fact that the encryption process is an internal process or sub process inside of the commerce transaction, probabilities are supporting the rules per each sell or buy. The rest of this paper is as follows: in second part all the mathematical machinery is presented. Once the formalism is done, we pass to formulate the Monte Carlo Algorithms as developed in the third section. In fourth section, we present the results using current data belonging to the Peruvian economy. Finally, the conclusion of this paper is presented.

## 2 Theoretical Model

Consider the flux of Bitcoin that is convolution of ordinary currency (dollar, euros, etc.) cash money governed by the integral of convolution following the well-known equation of Hammerstein:

$$\int_a^b h_Q(t,s)u(x_R(s))ds \tag{1}$$

where $h_Q(t,s)$ denotes the Kernel and $u(x_K(s))$ the functional of the $x_K(s)$ incoming fluxes that is actually a distribution of ordinary cash. The integration limits define the net amount of money to be converted in bitcoins. The meaning of this integration is as follows: given a certain flux of cash (ordinary money), then the convolution with the actual function that is expected to be a link between ordinary money and Bitcoin, would have as result the exact amount of money but in terms of Bitcoins in the time t. The solution of (1) demands of various ways about the usage of the mathematical methodologies. One manner which can be as an orthodox one, becomes the usage of orthogonal functions by which it can help us to some extent to get a fidelity in the solution of (1). In this manner, the incoming flux of cash is analytically described by the orthogonal functions and reads,

$$u(x_K(s)) = \sum_K^{\text{KMAX}} \mathcal{C}_K x_K(s) \tag{2}$$

with KMAX the number of fluxes of cash to be converted into Bitcoins, for any market this integer number depends in the offer and demands dynamics. The coefficient $\mathcal{C}_k$ is recognized as the confidence of the transaction and it is ranging between -1 to 1. On the other side, the kernels as written in (1) can also be worked out in through the separation of variables as seen in their dependence of t and s. With this, the kernel is expressed as:

$$h_Q(t,s) = H_Q(t)h_Q(s) \tag{3}$$

The introduction of the integer $Q$ is associated to a concrete transaction between two parties. So far, no any intervention of a third party in the binomial seller-buyer is considered in this analysis. The manner as (3) is written comes from the fact that any time evolution of the flux might be fully independent of the variable integration (such as a market well established is committed to buy and sell using units of Bitcoins in any time). Thus, the full integration reads

$$y(t) = \sum_{j=1}^{j\text{MAX}} \int_{a_J}^{b_{j+1}} \sum_K^{\text{KMAX}} \mathcal{C}_K H_Q(t)h_Q(s)\chi_K(s) \tag{4}$$

the usage of a sum over $j$ to JMAX details the different segment of times by which a $Q$ transaction is achieved. While the coefficients $\mathcal{C}_K$ denote the weights in according to the importance of the transactions. We found a list of criteria for this study

Readiness of the transaction
Pricing
Equity state
Product availability

It should be remarked that given a universe of transactions, we established in this analysis that no any preference to anyone of both parties: seller and buyer is applied. So under this view then the coefficients $\mathcal{C}_K$ are taken to be random, and the requirement of a universal random number generator is required for further simulations. For example, consider the minimal model for a single transaction of type one-to-one, so the resulting convolution would depend on the integration of the product of two continue functions:

$$y(t,j) = C_1 H_1(t) \int_{a_J}^{b_{j+1}} h_1(s)\chi_1(s)ds \tag{5}$$

The evaluation of Eq. (5) requires of the explicit knowledge of $h_1(s)$ that is nonetheless, an orthogonal function for a first troy model. For instance consider $x_1(s)$ having a oscillatory behavior: a deal of cash acquires is a functional of $sin(s)$ or $cos(s)$ functions for example:

$$\chi_1(s) = \sin^m(s) + \cos^n(s) + [\sin(s)\cos(s)]^p \tag{6}$$

Where the integer numbers $m, n$, and $p$ are arbitrary and for ends of a computational simulation, all of them are randomly selected. How $h_1(s)$ behaves as function of the incoming cash, it might be seen as an option, so that

$$y(J,t) = f(t) \int_{a_J}^{b_J} \mathcal{P}_L(s)[\sin(s)\cos(s)]^p ds \tag{7}$$

Where $\mathcal{P}_L(s)$ is a Legendre polynomial while $f(t) = C_1 H_1(t)$ the time evolution shape of the dynamics of bitcoin transactions With all this we can define the efficiency of this transaction in the time based uniquely in the gain or loss of cash:

$$y(J,t) = \frac{y(J+1,t) - y(J,t)}{y(J,t)} \tag{8}$$

That is well understood for a one-to-one transaction. Consider the general case of $Q$ customers performing $Q$ transactions as indicated in (3), then we finally arrive to

$$y_Q(J,t) = \frac{y_Q(J+1,t) - y_Q(J,t)}{y_Q(J,t)} \tag{9}$$

Thus, the full success of bitcoin transactions is defined by those that are presenting a large success probability that is tangible in the sense of - No abort of transactions

- Confidence of both parties
- Unstoppable and continuous networking
- Minimal local inflation

- Products market running over legality
- Negligible probability for complains Therefore, the fulfill of all these items would have as result the continuous increasing of an economy based in bitcoin [7]. However, the full departure of a fiducially currency to one that in based in encrypted currency might have unexpected events that would play the role of disturbs during the time that the transaction is achieved.

TABLE I. **RESULTING EFFICIENCIES FOR BITCOIN TRANSACTION FOR 6 TYPES OF COMMERCE IN PERU**

| Item | Fiduciary Cash (Vol) | Bitcoin Packages | Efficiency Transaction | Order |
|---|---|---|---|---|
| **Automobile** | 730 | 59 | 42% | **1** |
| **Housing** | 1052 | 27 | 75% | **2** |
| **Retail** | 459 | 13 | 93% | **3** |
| **Education** | 218 | 10 | 90% | **4** |
| **Electronics** | 781 | 61 | 96% | **5** |
| **Fashion** | 630 | 58 | 95% | **6** |

Fig. 1

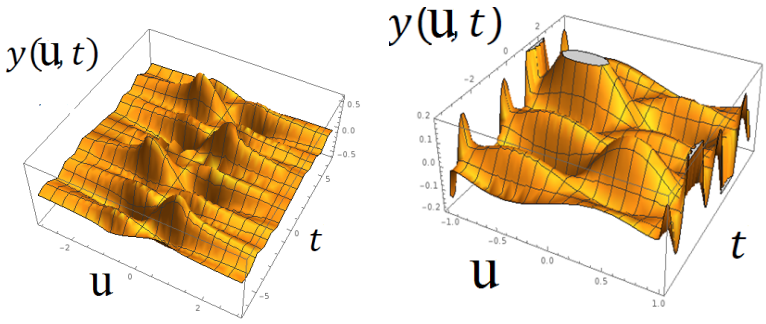



Fig. 2: 3D curves reconstruction of Eq.7. Left: for the case when m=n=0 and p=1 in Eq. 6. Right: m=0, n=1 and p=1.

## 3 Monte Carlo Simulations and Results

Below is written in its simplified version the code corresponding to the Monte Carlo simulation for the calculation of the efficiency of the bitcoin transaction.

Whereas in line 1 the code is fully initialized with the adjudication of value for the parameters and coefficients, in lines1 and 2 loops corresponding to the model accuracy and ordinary cash packets number, respectively. In line 3 we extract a random integer number that is the number of transactions set to 10,000 operations. With all this in line 5 the integrations corresponding to the Tsallis entropy is calculated. From line 4 to line 9 the Monte Carlo analysis is carried out emphasizing the line 6 that shows explicitly the acceptance or rejection. In line 8 we have saved all those curves that satisfy the Monte Carlo step.

**0 INITIALIZES**
**1 DO** K = 1, MAX! MODEL ACCURACY
**2 DO** J = 1, MAX! ORDINARY CASH PACKETS
**3** CALL RANDOM QMAX
**4 DO** Q = 1 QMAX! NUMBER TRANSACTION
**5** Y = Y(Q, J,K)
**6 IF** (Y ¿ YREQ) **THEN**
**7** ACCEPT Y MONTE CARLO ACCEPTANCE
**8** SAVE CURVES Y(Q, J,K)
**9 ENDIF**
**10 ENDDO**
**11 ENDDO**
**12 ENDDO**
**13 END**

In Fig.1, in left and right sides the cases where Eq.7 is plotted for two different scenarios of $m, n$ and $p$. The discrete variable J has passed to be continuous as $u$. Thus, such new variable gives account about the volumes of ordinary cash expressed in packets. Thus the peaks denote the high efficiency of bitcoin-based transactions. In right plot, In left side is seen the apparition of two large peaks that for certain values of $u$ but periodically. It clearly tells us that the efficiency in bitcoin transactions might not be a continuous activity in contrast to ordinary cash [8]. In Fig.2 the efficiency is plotted as a 3D curve and a contour plot. In both presentations one can see the presence of two bi-dimensional portions that to some extent might to indicate the existence of a certain periodicity to carry out bitcoin transactions but with a high efficiency. Although these bitcoin transaction are not exposed to taxes cuts [9], this point should be considered carefully. Table II lists the products used for simulation. Finally, electronics and fashion have turned out to be the more profitable operations with efficiencies reaching up to a 96% for electronics.

## 4 Conclusion

We have reported a Monte Carlo study that anticipated the economic consequences for using bitcoins in transactions in emergent markets, yielding that customers might to opt to use bitcoins to acquire electronics and fashion as yielded by the simulations from a universe of 10K transactions. The plots have also yielded interesting patterns that is translated as the possible periodicity

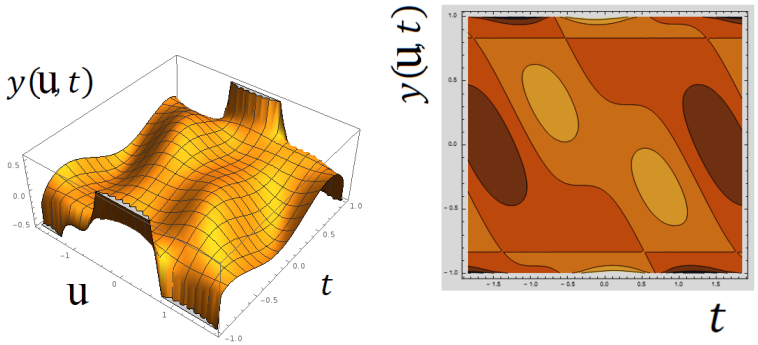



Fig. 3: Left: 3D curves reconstruction of Eq.7 for the cases when m=n=1 and p=2 in Eq. 6. Right: the corresponding contour plot. .

in the conversion of packets of ordinary cash to bitcoins to accomplish market transactions. Efficiency here is seen as an indicator that measures the level of favorability in using bitcoins for the economy of a city or country [10].

## References

1. Pavel Ciaian, Miroslava Rajcaniova and dArtis Kancs, The economics of BitCoin price formation, Applied Economics, Volume 48, 2016 - Issue 19.
2. Fulya Teomete Yalabk and smet Yalabk, Anonymous Bitcoin v enforcement law International Review of Law, Computers and Technology, Volume 33, 2019 - Issue 1.
3. Mark Holub and Jackie Johnson, Bitcoin research across disciplines, The Information Society, Volume 34, 2018 - Issue 2.
4. Yuen C Lo and Francesca Medda, Bitcoin mining: converting computing power into cash flow, Applied Economics Letters, Published Online: 07 Nov 2018.
5. Jying-Nan Wang, Hung-Chun Liu, Shu-Mei Chiang and Yuan-Teng Hsu, On the predictive power of ARJI volatility forecasts for Bitcoin, Applied Economics, Published Online: 12 Apr 2019.
6. Adam S. Hayes, Bitcoin price and its marginal cost of production: support for a fundamental value, Applied Economics Letters, Volume 26, 2019 - Issue 7 Published Online: 20 Jun 2018.
7. Bill Maurer , Taylor C. Nelms and Lana Swartz, When perhaps the real problem is money itself!: the practical materiality of Bitcoin, Social Semiotics, Volume 23, 2013 - Issue 2, Published Online: 12 Mar 2013.
8. P. S. Lintilhac and A. Tourin, Model-based pairs trading in the bitcoin markets, Quantitative Finance, Volume 17, 2017 - Issue 5, Published Online: 04 Nov 2016.
9. Marie Vasek, The Bitcoin Brain Drain: Examining the Use and Abuse of Bitcoin Brain Wallets, International Conference on Financial Cryptography and Data Security FC 2016: Financial Cryptography and Data Security pp 609-618.
10. Bonaiuti G. (2016) Economic Issues on M-Payments and Bitcoin. In: Gimigliano G. (eds) Bitcoin and Mobile Payments. Palgrave Studies in Financial Services Technology. Palgrave Macmillan, London.

# Mimicking the Mechanisms of Language for the Unsupervised Detection of Hierarchical Structure in Time Series

Christopher Josef Rothschedl, Paul O'Leary, and Roland Ritt

Chair of Automation, University of Leoben, Leoben, Austria
christopher-josef.rothschedl@stud.unileoben.ac.at
https://automation.unileoben.ac.at

**Abstract.** The purpose of this paper is to investigate, how mechanisms of natural language can support the analysis of time series data emanating from human-operated physical systems. There may be no physical models to describe human behavior reliably; nevertheless, there is commonly structure in the data acquired from observing this behavior. This paper focuses on investigating a new approach to the unsupervised detection of structure in such data. Symbolic analysis and linear differential operators are adopted, to derive results bearable for resemblance with the metaphorical concept of language and its mechanisms. To support this, phenomenology is introduced and discussed. The general significance of metaphors – and why they may be of specific relevance to analyzing time series data – is discussed and references to exemplary applications are presented to ascertain the validity of the approach.

**Keywords:** Symbolic analysis, time series analysis, unsupervised detection, metaphor of language, phenomenology

## 1 Preamble

This paper focuses on time series data emanating from cyber physical systems, i.e., on the example of machines being operated by humans. In an exploratory manner, analytical approaches are adopted to mimic basic concepts of natural language. Especially in the advent of machine learning and related techniques, it needs to be pointed out that learning is fundamentally function without understanding. Hence, such perceptive approaches alone cannot solve all kinds of problems in a generalized manner, which is why analytical methods – such as in the form of symbolic analysis – are required to acquire understanding. A discussion section at the end of this paper poses an attempt to associate results from mimicking the mechanisms of language with phenomenology; in particular with the Asian views on this and how it could be of support when analyzing time series data of human-operated equipment.

Eventually, the aim of this paper is to initiate discussions on alternative approaches to analyzing time series data emanating from phenomena, i.e., the

observation of physical systems, as a possible precursor to new implementations of what we might wish to call artificial intelligence. We also present some initial results on the emergence of language in the analysis of multi-channel time series relating to physical systems: the results are not final but their quality certainly justify further research.

## 2 Introduction

In the following paragraphs, our approach on symbolic analysis of time series data emanating from human-operated equipment is presented. The results of the proposed concept will then be associated with components of language they potentially correspond to. The subsequent discussion section about the relationship of the presented work to phenomenology provides insights on the background and should pave the way for further discourse about its significance to the matter.

## 3 Symbolic Analysis of Time Series Data Emanating from Human-Operated Machines

Symbolization plays a major role in what we want to achieve with the analysis of time series data of machines, which are operated by humans. Symbolic representations of time series have been used in a number of different manners [1, 2, 10, 13, 16, 15, 17]; most notably among these is the symbolic aggregate approximation (SAX) [9]. The SAX algorithm works directly on the original data stream and uses a set of linear quantization levels to define the alphabet; in this manner, numerosity is reduced with the goal of simplifying the identification of patterns. Furthermore, the approach poses a lower bounding property, giving it a positive semi-definite measure, i.e., the distances between sequences have a strict mathematical meaning. However, SAX does not take the nature of the system being observed into account. They have neither proposed nor offered any additional abstraction relating to the metaphoric nature of language and how it points to implicit structure in the dynamics of the experience of phenomena. For example, in language we have *nouns* and *verbs*, which each refer to a different aspect of an experience and we have implicit structure which is captured by the grammar of a language, in the most general case we may consider grammar as ensuring that the symbols are ordered correctly so that they are a correct and valid reference to the experience of phenomena.

### 3.1 Adding Physical Dynamics

When analyzing physical systems, dynamics need to be considered in our models and approaches. Such dynamics are usually modeled by differential equations. As humans, we intuitively solve certain classes of differential equations, e.g., when crossing the road: we estimate the position, velocity and acceleration of a car to ascertain if it is safe to cross. Such capabilities are also needed in data

analysis. Consequently, we propose adding linear differential operators (LDO; [8], detailed implementation in [3]) as precursors to symbolization of continuous time series data streams. Additionally, the LDO introduce the concept of dynamics, giving a natural link to the linguistic elements called *verbs*[1]. Deriving velocity or acceleration from position signals depicts simple examples for the application of LDO. This enables the generation of additional linguistic elements such as *verbs*, *adjectives*, *adverbs*, etc., which substantiates the emergence of language as a metaphor.

## 4 Introducing the Mechanisms of Language

We now provide a basic definition of language relating to the analysis of multi-channel time series data streams. The model is somewhat restricted; nevertheless it is sufficient to support initial research into the viability of applying the metaphor of language to real physical systems.
In Yogācāra phenomenology [11, 14], it is proposed that repetitions in our sensory excitation, which are significant to our situation, are assigned a language representation, i.e., words. Consequently, meaning is simultaneously both, experiential and contextual. It is the repetition which is considered to be characteristic and not the thinking about the repetition; this process leads to a naming. At this point we may not understand, however, we do perceive. In this manner the use of a word is considered to be a metaphor, which points at a sensory experience rather than describing the experience directly. Monosyllabic words are considered a representation of simpler sensory experiences, while polysyllabic words tend to describe more abstract experiences, which are commonly the result of complex multi-sensory experience.

In Proto-Indo-European languages, each syllable has a specific meaning and more complex experiences are expressed as *poly*syllabic words. Furthermore, predicates – primarily adjectives and adverbs – emerge to define properties of objects and activities. Other linguistic elements, such as punctuation, can also be of metaphorical support in future research. In this work we investigate the usefulness of the metaphor of language when analyzing real-time time series data of machines – particularly, when these are being operated by humans. One interesting result is that, given the "symbolized" data streams (the words), the mechanisms of language, e.g., contraction, compounding, etc., lead to a natural formulation of a hierarchical decomposition of the data, revealing implicit structure embedded in the data (see example in upcoming section 5.4).

For Lakoff and Johnson [7], metaphors exhibit influential significance beyond linguistics. They suggest that human thinking, action and speaking follow metaphorical concepts, in daily life as well as in science. In their book, metaphors are presented to possess self-contained cognitive value, instead of only being linguistic tools for simple comparisons. Furthermore, it is pointed

[1] Beyond that, LDO can be used to compute local time-dependent estimates for the state space vectors of systems, which, if mapped to the pseudo phase space, allow prediction of system behavior.

out that metaphors – instead of being mappings – could also add elements to a domain. This is exactly what we want to achieve with data analysis: integrate domain knowledge to derive understanding from time series data. Metaphors, and in our case especially the metaphor of language, can be of extensive support with regard to how complex time series data from physical systems can be conquered.

## 5 Mimicking the Mechanisms of Language

Time series data of bucket-wheel excavators was used to demonstrate initial ideas regarding the metaphorical relevance of emerging language. The sensor and actor data was sampled at a rate of 1 Hz. These machines used in open-pit mining are operated by humans in a non-continuous manner, making the time series data suitable for analysis in the sense of metaphorical language.

### 5.1 Monosyllables and Polysyllables

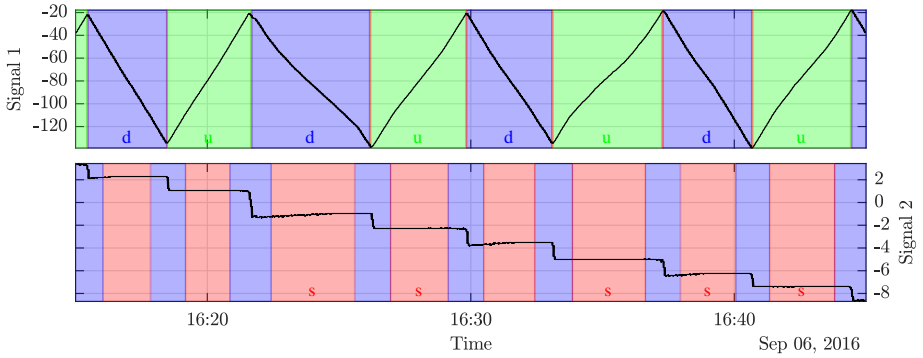
Symbolization of single channels produces words we refer to as *monosyllables*. That are, *nouns* for describing states, and *verbs* for describing activities. The results correspond to sequences of words (see Figure 1a for two individual channels) which are highly suitable for further processing, also for algorithms such as *regular expressions* (*regex*).

Cross-channel combinations of these monosyllables yield *polysyllables*, describing more complex states and/or activities across several channels in a more abstract manner. The example given in Figure 1b illustrates polysyllables gained from the combination of monosyllables of the channel Signal 1 (verbs; **u** (green) and **d** (blue), both describing different directions of boom slewing) and monosyllables of Signal 2 (nouns; **s** (red), identifying non-motion states of boom luffing). The newly generated syllables (**ds** and **us**) now describe machine operations on a more general, abstract level. Such methods support the integration of knowledge from experience of domain experts, as time series data is prepared in a way more practicable for the analysis of (complex) machine processes.

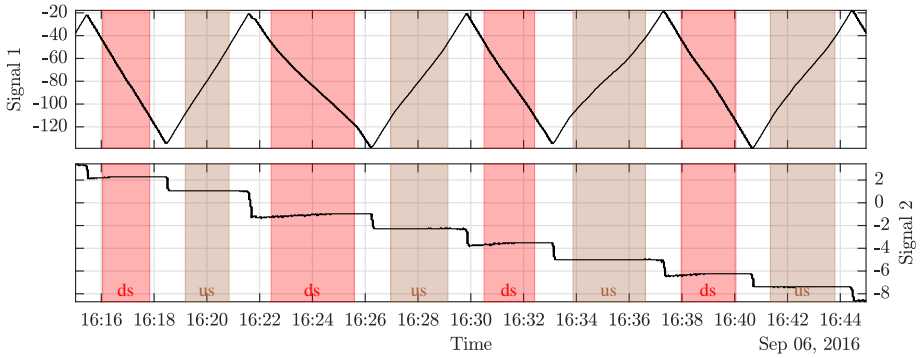
### 5.2 Predicates: Adjectives, Adverbs

By symbolizing time series data, states (nouns) and activities (verbs) are predicated with their run-lengths[2] (*adjectives*, *adverbs*). The example in Figure 2 shows two separate sequences of a time series with their symbolic mapping. The identified symbol sequences along with the corresponding predicates (referring to the run-length) are shown on the right. This twofold structure (nouns/verbs and associated adjectives/adverbs) can be used when analyzing the underlying structure of such sequences, e.g., both signal snippets in the figure show a different time range, 50 and 40 minutes. However, the sequence structure is the

[2] *Runs* of data are sequences of consecutive same symbols within a time series.




(a) Monosyllables of two individual signal channels are shown at the top, where Signal 1 is symbolized by verbs to represent the specific motion (boom slewing left/right) of the machine and Signal 2 shows the position of the boom luffing unit (up/down) as nouns (**s** in red color identifies the state *luffing is stationary*).




(b) The individual monosyllables from above are combined to form polysyllables that refer to all occurrences where Signal 1 exhibits actions **d** or **u** and Signal 2 exhibits the state **s**.

Fig. 1: Monosyllables and their combinations yielding polysyllables

same, although the lengths of the individual runs differ. This form of run-length encoding enables data compression.

The adjectives and adverbs can deliver additional information, such as motion time or speed, depending on what the corresponding nouns or verbs indicate. Furthermore, irregularities can be detected by identifying words with an uncommon length, speed, etc., or certain analyses can be made possible when omitting words which do not reach a certain length or speed limit.

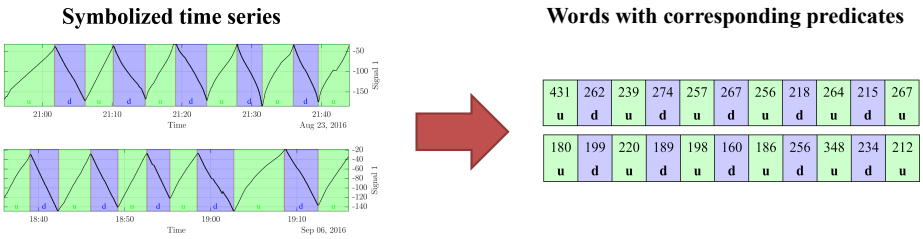



Fig. 2: Symbolized time series of different time ranges are shown on the left (top: 50 minutes; bottom: 40 minutes), while the corresponding word sequences can be seen on the right side. The words (**u**, **d**) are printed along with the associated predicates (run-lengths). Although the predicates of both signal snippets differ, they exhibit the same underlying structure.

### 5.3 Frequency Dictionaries

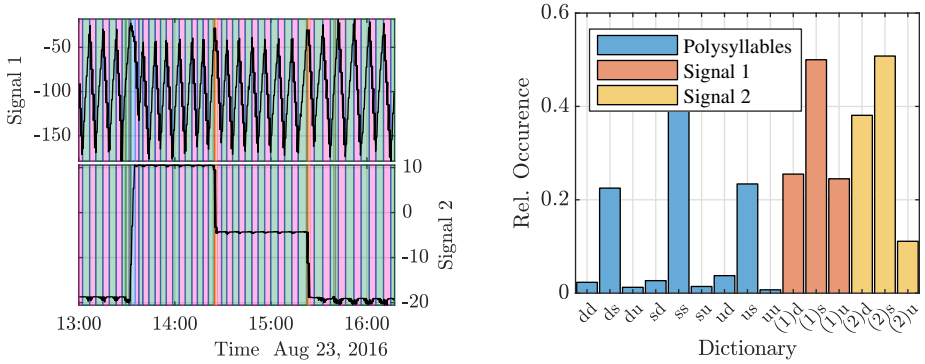
By interpreting time series and the combinations of the individual channels, we not only yield sequences of words, but also the frequency of all word occurrences. Picking up the idea of *frequency dictionaries*, words can be sorted by their frequency to reveal which words are common and which are not. This is especially interesting for machines: if a frequency dictionary is generated during controlled tests in the commissioning phase of a machine, it can build the reference for how the response behavior of the system should look like — a form of *operations recognition*. During operation, deviations can be identified by comparing this reference operation profile with the actual profile. Also outliers, i.e., words that never or rarely occurred during commissioning, or unusual distributions of words, can be detected; further investigations can be performed if necessary.

In Figure 3a, there are two channels characterizing a certain machine operation mode[3] on the left side, while on the right plot the occurrence statistics of the words, mono- and polysyllabic, are given. In contrast, Figure 3b shows another operation, again based on the two channels with the corresponding statistics plot. To distinguish between the two operations, the information regarding the frequency of the polysyllables is essential, as the statistics of the monosyllables of both operations do not differ as distinctively. This indicates that mechanisms of forming polysyllabic words reveal additional information about a system by letting a machine-specific language emerge.

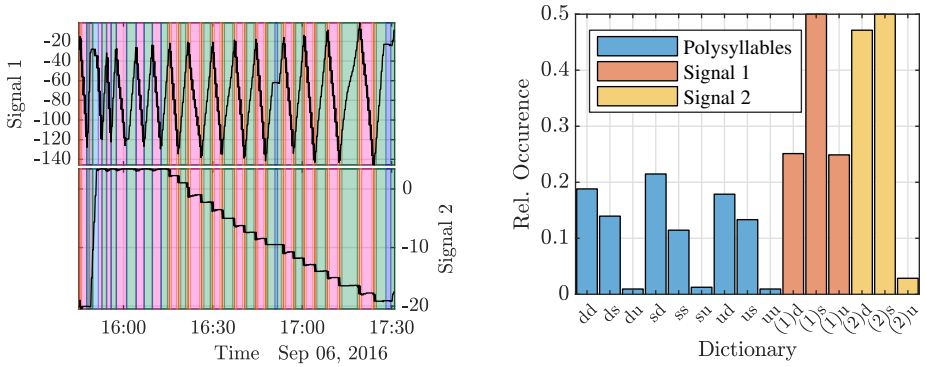
### 5.4 Compounding

In natural language, patterns that occur repeatedly are often combined to form a simpler and less complex identifier, i.e., a metaphor describing what is meant. For instance, offering a cup of coffee to someone will be understood easily while

[3] For demonstration purposes, the count of signals relevant to the identification of operation modes was limited to two.




(a) Operation mode 1: The bucket-wheel excavator is operated in terrace-cut mode.




(b) Operation mode 2: The bucket-wheel excavator is operated in drop-cut mode.

Fig. 3: Two time series signals with their corresponding mapping of mono- and polysyllables are shown on the left side for each operation mode of a bucket-wheel excavator. The statistics for each mode are plotted on the right.

offering "a ceramic container with blackish fluid, white substance from an animal and sweet-tasting carbohydrates" instead might leave most in confusion. Additionally, this example might also imply another level of abstraction where the initiator seeks a more casual means of communication outside his office by sharing a coffee in the cafeteria.

A similar issue arises when analyzing large time series: the simpler a certain system response behavior or machine operation can be represented for subsequent processes, the more efficient it is to draw conclusions. While the initial time series are only present as raw data signals, the compounding of nouns and verbs (from intermediate symbolization) adds a substantial level of abstraction by providing a succinct identifier (metaphor) for a complex operation or event/incident.

A simple example using two signals of a bucket-wheel excavator is illustrated in Figure 4. The top section (Level 1) shows the original symbolic assignments,

resulting in 712 polysyllabic words for the presented time range; the corresponding dictionary contains 9 definitions. All existing symbols are mapped onto the original signals with different colors. The bottom section (Level 11) shows the result after 11 iteration steps, totaling in 67 words. Each iteration combines the most frequent word combinations. The result exhibits clearly the main operating modes (and pauses between them). The level of complexity is decreased significantly, while the level of abstraction is increased to benefit subsequent analyses. The central part of the figure illustrates all the single iterations (one line per iteration) necessary to achieve the desired level of abstraction; this hierarchical compounding process enables the implicit structure within the data to emerge without using a priori knowledge.

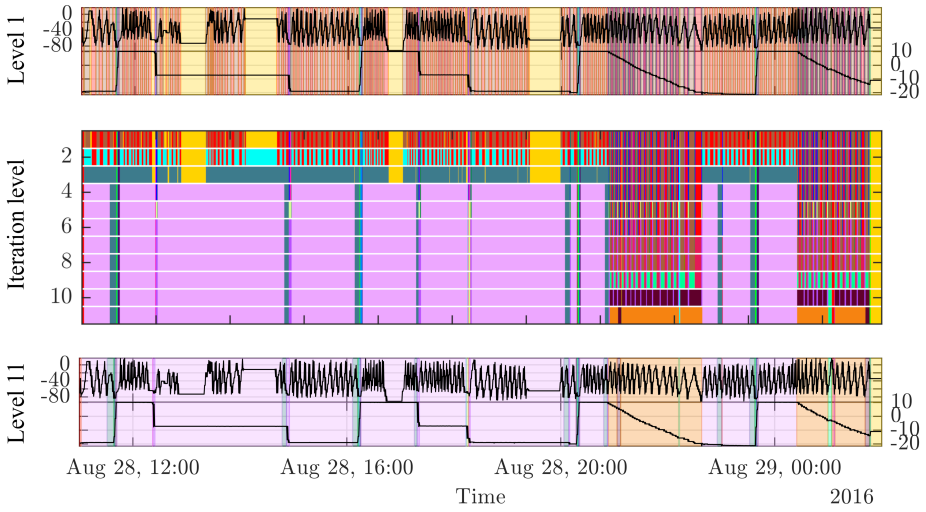



Fig. 4: The top plot represents the initial level (Level 1) with 712 words; the corresponding dictionary contains 9 definitions for the symbolization. The bottom plot shows the result after 11 iterations, with the word count being reduced to 67. The centre part illustrates all steps performed to reach the desired abstraction.

## 6 Discussion: The Metaphorical Concept of Language and its Relationship to Phenomenology

By introducing the metaphor of language, additional results implicit to the data could be drawn from time series, as the results from observations in the previous section show. This can be interpreted as a pointer to phenomenology.

Although the origins date back to the 18$^{th}$ century, Edmund Husserl (1859-1938) is considered the founder of the philosophy of Western phenomenology [14].

In simple terms, phenomenological philosophy is about the structures of experiences and consciousness. Husserl describes *experience as the source of all knowledge* [6]. One of his students, Martin Heidegger, also contributed philosophical insights to phenomenology. Heidegger introduced the concept that lived experiences always consist of more than what can be seen [5] — this is considered as a pointer to hidden models, implicit structures within natural experiences. Another student of Husserl, Maurice Merleau-Ponty, extended phenomenology in regard to how we perceive as a result of experiencing phenomena [12]. In quite simple terms it can be concluded that *we experience first and reflect afterwards.*

In contrast to Western philosophy, the Eastern view of phenomenology is provided by the Yogācāra school, which was founded by Asaṅga and Vasubandhu in the $4^{\text{th}}$ century CE. Although this philosophy has had quite some time to mature, it is still considered valid today [11]. Even Heidegger – a representative of Western phenomenology – indicated particular interest in Eastern positions, although pointing out hindrances between Eastern and Western philosophies in his work *A Dialogue on Language between a Japanese and an Inquirer* [4].
Two main concepts of the Yogācāra phenomenology are the *five skandhas* (also known as five aggregates) and the *eight vijñānas* (also known as eight consciousnesses). The concepts support the view of life as a continuous flow of sensory experience to which meaning and significance are added. The skandhas describe how we get from sensory excitation to discursive thinking in five distinct steps, i.e., the five aggregates. This concept is relevant, as it conceptualizes that we are never in direct contact with objects in the world, but we are always in contact with a model of the world. An overview of the skandhas is given in Table 1 along with their English translations and interpretations of meaning in a technical context.

Table 1: The five skandhas and possible technical interpretations [11, 14]

| **Sanskrit** | **English** | **Technical equivalent** |
|---|---|---|
| rūpa | Form | Context-dependent sensor information |
| vedanā | Feeling (Sensation) | Low level model-based control |
| saṃjñā | Perception | Combination of low level data to identify a situation |
| saṃskārā | Impulse (Association) | Learned situative semi-autonomous behavior |
| vijñāna | Consciousness (Discursive thinking) | Artificial reasoning, e.g., rule systems |

The second main phenomenological model is the concept of the eight vijñānas. This model relates to the dynamics of our experience on how we establish understanding during the process of emergent consciousness. The first five vijñānas

are related to our sensory experiences. The other three are *mano-vijñāna* (modeling sensory experience), *manas-vijñāna* (self-referencing of consequence) and *ālaya-vijñāna*[4] (models for past experience). Language as a metaphorical reference to the content of an experience emerges within these phases as illustrated in Figure 5. Here, language is interpreted as a metaphor for how we understand such an experience. As it emerges throughout the full spectrum of mental models, it can also exhibit variable stages and different strengths of metaphorical references.

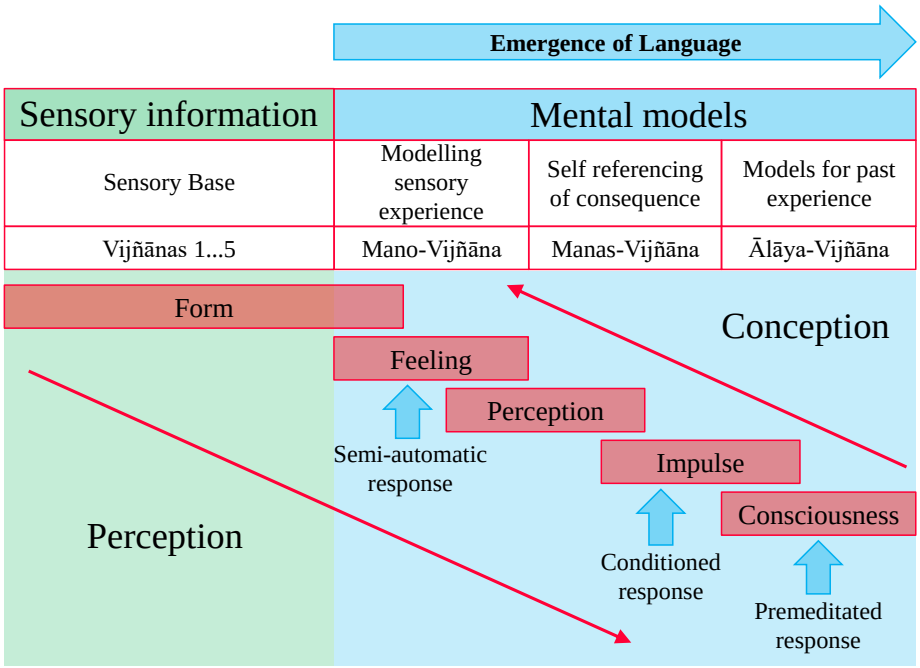



Fig. 5: A schematic overview of a possible interpretation of the eight vijñānas and the five skandhas in regard to the emergence of language. Following this scheme, we can pose a simple example: we hear a sound (*form*); we have a neutral *feeling* about it; we *perceive* it to be an aircraft; it is the plane between Graz and Vienna, just like every day at this very time (*impulse*); we think about whether or not we already checked in for our own flight the next day (*consciousness* or *discursive thinking*).

The different steps in Figure 5 also allow responses along the path from *form* to *consciousness* or vice versa. Examples from everyday life: a child stumbles

[4] In literature, this is often referred to as "store-house consciousness". However, we are careful with translating it into a single English term so as not to neglect any of the semantic nuances as also stated in [18].

besides us and we reach out to prevent falling (*semi-automatic response*); we approach a junction and the car from the left does not decelerate: we are getting slower and more cautious (*conditioned response*); you hurt yourself and you need medical attention, so you make an emergency call (*premeditated response*).

## 7 Conclusions

The phenomenological model of how humans experience and how language emerges proved valuable for analyzing time series data emanating from physical systems. With the metaphor of language, additional levels of abstraction could be derived from time series data. Implicit structures were found in time series data of physical systems, adding information about the behavior of the system being observed by utilizing the metaphorical nature of those. Modeled system dynamics are significant to the analysis processes for obtaining additional metaphorical, linguistic elements such as verbs, adjectives, and adverbs.

Based on the promising results, it can be concluded that the phenomenological approach and the metaphor of language bear the potential of being of significant relevance to time series analysis of physical systems. At this point in time there may be alternatives to what we have proposed; however, this is not the primary issue in this paper. Our goal is to initiate discussion and further research on completely new approaches to analyzing time series data obtained from the observation of human behavior when operating physical systems. In Asian phenomenology, language is considered to be central as to how humans build mental models for their interaction with the world surrounding them. We feel this – combined with the initial results presented here – justify further discussion and research.

## References

1. Beskhyroun, S., Wegner, L.D., Sparling, B.F.: New Methodology for the Application of Vibration-Based Damage Detection Techniques. Structural Control and Health Monitoring (May 2011), n/a–n/a (2011). https://doi.org/10.1002/stc, http://dx.doi.org/10.1002/stc.456
2. Camerra, A., Palpanas, T., Shieh, J., Keogh, E.: iSAX 2.0: Indexing and Mining one Billion Time Series. Proceedings - IEEE International Conference on Data Mining, ICDM pp. 58–67 (2010). https://doi.org/10.1109/ICDM.2010.124
3. Gugg, C., Harker, M., OLeary, P., Rath, G.: An Algebraic Framework for the Real-Time Solution of Inverse Problems on Embedded Systems. In: 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems. vol. V, pp. 1097–1102. IEEE (aug 2015). https://doi.org/10.1109/HPCC-CSS-ICESS.2015.50, http://ieeexplore.ieee.org/document/7336315/

4. Heidegger, M.: Aus einem Gespräch von der Sprache. Zwischen einem Japaner und einem Fragenden (1959)
5. Heidegger, M.: Sein und Zeit. Max Niemeyer Verlag, 11 edn. (1967)
6. Husserl, E.: Philosophie als strenge Wissenschaft. Logos: Zeitschrift für systematische Philosophie pp. 289–341 (1910)
7. Lakoff, G., Johnson, M.: Metaphors We Live By. The University of Chicago Press (2003)
8. Lanczos, C.: Linear Differential Operators. SIAM (1961), http://epubs.siam.org/doi/pdf/10.1137/1.9781611971187.fm
9. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 2–11. ACM, New York, NY, USA (2003)
10. Lin, J., Khade, R., Li, Y.: Rotation-Invariant Similarity in Time Series Using Bag-of-Patterns Representation. Journal of Intelligent Information Systems **39**(2), 287–315 (2012). https://doi.org/10.1007/s10844-012-0196-5
11. Lusthaus, D.: Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Ch'eng Wei-shih Lun. Curzon Critical Studies in Buddhism, Routledge Curzon (2002), http://books.google.at/books?id=IeiwsT-XqwQC
12. Merleau-Ponty, M.: Phenomenology of Perception. Routledge classics, Routledge (2002), http://books.google.at/books?id=oSgaSzvHbaoC
13. Minnen, D., Isbell, C., Essa, I., Starner, T.: Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). pp. 601–606. IEEE (oct 2007). https://doi.org/10.1109/ICDM.2007.52, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4470297 http://ieeexplore.ieee.org/document/4470297/
14. O'Leary, P., Harker, M., Gugg, C.: A Position Paper on: Sensor-Data Analytics in Cyber Physical Systems, from Husserl to Data Mining. In: SensorNets 2015, Le Cresout, France (2015)
15. Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A.P., Chen, C., Frankenstein, S.: Time Series Anomaly Discovery with Grammar-Based Compression. In: Edbt. pp. 481–492 (2015). https://doi.org/10.5441/002/edbt.2015.42
16. Senin, P., Malinchik, S.: SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. In: 2013 IEEE 13th International Conference on Data Mining. pp. 1175–1180. IEEE (dec 2013). https://doi.org/10.1109/ICDM.2013.52, http://ieeexplore.ieee.org/document/6729617/
17. Vespier, U.: Mining Sensor Data from Complex Systems. Ph.D. thesis, Leiden University (2015), http://hdl.handle.net/1887/37027
18. Waldron, W.S.: The Buddhist Unconscious - The Ālaya-Vijñana in the Context of Indian Buddhist Thought. Routledge Curzon

# Prediction of Transformer Temperature for Energy Distribution Smart Grids Using Recursive Neural Networks

F.J. Martinez-Murcia[1], J. Ramirez[2], F. Segovia[2], A. Ortiz[1], S. Carrillo[3], J. Leiva[3], J. Rodriguez-Rivero[2,3], and J.M. Gorriz[2]

[1] Department of Communications Engineering, University of Malaga (Spain)
fjmm@ic.uma.es
[2] Department of Signal Theory, Networking and Communications, University of Granada (Spain)
[3] Endesa Distribución, Madrid (Spain)

**Abstract.** The near future of energy is conformed by a plethora of heterogeneous sources as well as an increasing demand. This poses new challenges for energy production and distribution, in which it will be essential that Medium Voltage/Low Voltage (MV/LV) distribution networks are planned, operated and monitored in an analogous way to what transport networks have been doing for decades. Within this context, an accurate tool for anticipating transformer overload and potential network problems is of paramount importance. Here, a system that can predict transformer temperature –a critical indicator of potential problems in a transformer– is key to the development of versatile and autonomous network control strategies that allow for a smarter distribution of energy. In this work we propose a transformer temperature prediction system based on long short-term memory (LSTM) networks that uses data from the previous 100 minutes to predict the transformer temperature in the next 100 minutes. The system is able to predict with a low error the temperature value using only the active power of the three transformer lines together with the ambient temperature. This makes it possible to discover trends towards anomalous temperature values in different transformers and act accordingly by planning a redistribution of the workload, avoiding possible incidents or service interruptions.

**Keywords:** MV/LV, temperature, time-series prediction, LSTM, regression, energy distribution

## 1 Introduction

The near future of energy is conformed by an increasing presence of electric vehicles, massive penetration of renewable and heterogeneous energy sources with low emission levels and the introduction of varying practices such as self-consumption [1]. Within this scenario, it will be essential that Medium Voltage/Low Voltage (MV/LV) distribution networks are planned, operated and

monitored in an analogous way to what transport networks have been doing for decades, in which the distributor goes from a mere distribution asset manager to being the operator of the network. This inevitably implies that the voltage levels are provided with much more intelligence than hitherto [2], involving a whole spectrum of digital technologies: sensorization -including smart meters-, broadband communications, local controllers based on Internet of Things (IoT) devices, Supervision, Control and Data Acquisition (SCADA) devices and Energy Management Centres (EMCs) that implement advanced data processing software, optimal control, or workload prediction, among others.

In this context, initiatives such as the Monitoring and Advanced Control (MONICA) of MV and LV distribution networks have provided solutions such as an state estimator of MV/LV networks. New initiatives like the spanish Preventive Analysis of Smart Grid with real Time Operation and Renewable Assets Integration (PASTORA) project –a follow-up to MONICA– are key in order to advance in the development of flexible, reliable and efficient networks capable of absorbing the maximum renewable generation at the lowest cost. For this purpose, the project proposes, among others, the development of real-time information processing tools and analysis of historical series for prediction of possible device overload. In this context, a system capable of identifying patterns of anomalous behaviour of the network could act preventively with regard to incidents and breakdowns, improving the quality of service of the energy supplier.

Particularly, the treatment of historical data of the MONICA project could allow the system to predict anomalies in the transformer variable, especially with regards to the temperature. The prediction of possible anomalies in the temperature of the transformers could help prevent network malfunctioning, triggering a series of security containment protocols to prevent overload and optimize energy distribution in this context of heterogeneous power generation and demand. Thus, a series of actions have been directed towards the construction of an anomalous temperature early warning system (SATTA).

In this regard, there exist a vast literature of time-series prediction algorithms, ranging from classical Auto-Regressive (AR) [3] methods to complex machine learning regression techniques like Support Vector Machines (SVM) [4, 5]. The current wave of neural networks architectures has revolutionised the classification and regression paradigm [6, 7], with many applications in fields such as image recognition [7], generative models [8] or biomedical image analysis [9], among others. Within this context, the recent advances in recurrent neural networks –networks with feedback links– have paved the way for newer applications in time series analysis and prediction using either Convolutional Neural Networks (CNNs) [10] or the long short-term memory (LSTM) [11] cells, which have experienced a major growth in the last years with many applications in, among others, stock market prediction [12], speech recognition [13, 14] or even music composition [15].

In this paper we propose a recurrent neural network architecture based on LSTMs in order to predict temperature levels of a transformer from a series

of power and temperature variables. In Section 2 we propose the methodology that combines feature selection via covariance matrices and LSTM networks for prediction, as well as the dataset used. In Section 3, we describe the evaluation procedure and present and analyse the results. Finally in Section 4 we draw some conclusions about the proposed system.

## 2 Data and Methodology

### 2.1 Data Acquisition

Data used in the preparation of this article was provided by ENDESA, the largest electric utility company in Spain. It was obtained during the MONICA (acronym for Advanced Monitoring and Control) project, with the fundamental objective of developing a technology that allows real-time monitoring and diagnosis of medium and low-voltage distribution networks, with an approach similar to that which has traditionally existed in transmission networks (high voltage). The data consists of yearly acquisitions of different variables at the transformation centres in southern Spain. It comprises a large and variable number of measures, including active and reactive power delivered by the transformer, reactive, capacitive and inductive energy, intensity, phase, voltage and temperatures.

In this work, we use the 16 transformers which provide information about Transformer Temperature (TT), a critical variable to measure potential incidences and anomalous behaviour, including transformer overload. Since the signals were recorded with non-uniform period, the data was subsequently resampled to 12 samples/hour (or a time-step $\tau$ of 5 minutes), corresponding to the mode of the data distribution. A simple linear interpolation between consecutive samples was used for this procedure.

### 2.2 Feature Selection

The MONICA dataset contains data from a real-world distribution network that is currently being used. Therefore, the number of variables recorded in each transformer may vary from one to the other, so we must establish a relevant set of variables to be used henceforth. In order to select a set of variables that are both available for all transformers and relevant for the predicted variable, we propose a method based on the computation of the inter-variable covariance matrices using the Ledoit-Wolf shrinked covariance estimate [16, 17].

Figure 1 displays the covariance matrices for transformers 0 and 10 (similar matrices are obtained for all the 15 studied transformers). A greater covariance between two variables may be an indicator of the potential predictive capability of one to the other [17]. Since our variable of interest is the TT, we can observe that the most influential variable is the Ambient Temperature (TA). Other variables such as the active power (PA) supplied and the reactive power (PR) for each of the R, S and T lines also display some influence with TT. The remaining variables in most transformers show very small covariance, so PA, PR and TA will be selected for this prediction task.

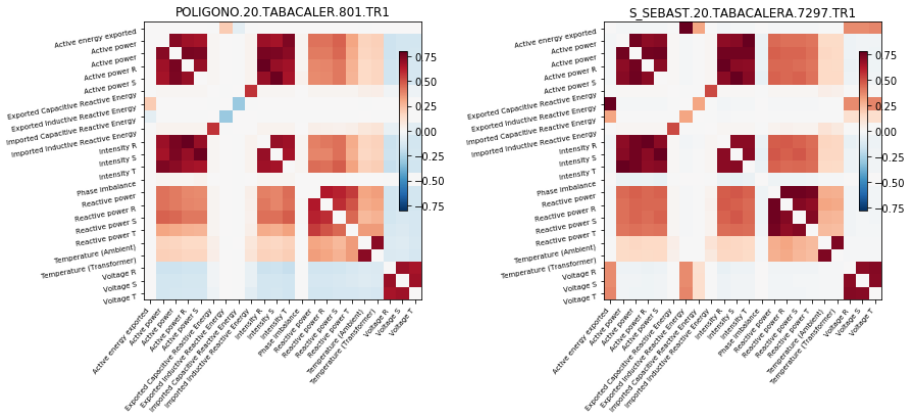



**Fig. 1.** Covariance (connectivity) matrices estimated for TRFs 0 and 10 using the Ledoit-Wolf method. Observe that the highest correlation with the transformer temperature can be found with variables PA, PR and TA.

### 2.3 Recursive Neural Networks

Although there exist many examples of time-series processing using neural networks such as Restricted Bolzmann Machines or CNNs [10, 18], Recursive Neural Networks (RNNs) are the state of the art for time series prediction and analysis. RNNs are a subtype of neural architectures specially designed for temporal processing, in which some type of memory or 'state' is hold within the network. Recursive means that unlike typical feedforward networks [19, 9], it has feedback connections. They are usually arranged in 'cells' that hold some memory of the past events in order to provide activations. Networks architectures based on Long Short Term Memory cells (LSTM) or Gated Recurrent Units (GRU) are becoming commonplace in applications such as stock market prediction [12], speech recognition [13, 14] or even music composition [15].

**Long Short-Term Memory cell** The long short-term memory (LSTM) [11] is a recurrent architecture. The architecture contains a memory activated via a "forget" gate that, together with an input and output gates, regulate the flow of the information and whether they are relevant for the output or not. A schema of a LSTM cell is shown at Figure 2.

LSTM networks are particularly good for the analysis and prediction of time series data. Within the architecture proposed in Figure 2, the equations that govern the behaviour of the unit can be summarizes as follows:

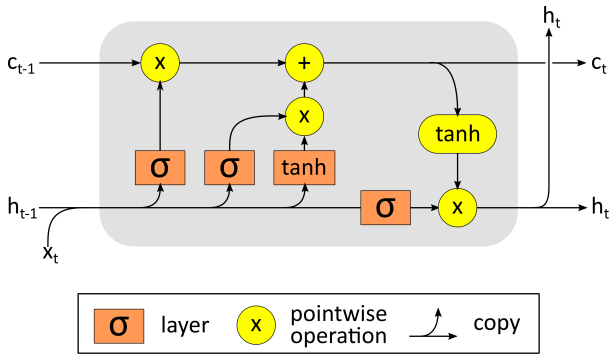



**Fig. 2.** Structure of a Long Short-Term Memory (LSTM) cell. Refer to the legend for understanding the layers and operations applied.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{4}$$
$$h_t = o_t \circ \sigma_h(c_t) \tag{5}$$

where $\sigma_g(x)$ and $\sigma_c(x)$ are the sigmoid and hyperbolic tangent activation functions. $x_t \in \mathbb{R}^d$ and $h_t \in \mathbb{R}^h$ are the input and output (also known as hidden state) vector of the LSTM unit, of length $d$ and $h$ respectively, $f_t \in \mathbb{R}^h$ the forget gate's activation vector, $i_t \in \mathbb{R}^h$ the input gate's activation vector, $o_t \in \mathbb{R}^h$ the output gate's activation vector, $c_t \in \mathbb{R}^h$ the cell state vector and $W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$ the weight and bias parameters of the different layers implemented in each gate.

**Network Architecture** In this work we used a LSTM network composed of two LSTM cells of 100 and 50 units connected to a dense layer of 20 units. The network is trained with different combinations of the PA, PR and TA variables in the last $\tau$, and the output is intended to predict TT with a maximum lapse of 20 $\tau$. The predicted TT is therefore in a range between 5 and 100 minutes from the current instant, fed by the PA, PR and TA data of the 100 minutes previous to the current instant.

## 3 Results and Discussion

### 3.1 Evaluation

Our system is trained with the data of each of the 15 available transformers, using the selected variables over the 20 previous timesteps. The trained system is then tested with two different approaches:

– Predictive ability **on the training transformer**. A time-series cross validation (CV) [20, 21], in which the timeseries is divided into progressive sequential batches (see Fig. 3) and all previous batches are used to predict the next one, is used to estimate the performance.
– Generalization ability **over other transformers** (GEN). In this case, the system is trained with one transformer, and then the predictive ability over other transformer's data is tested.

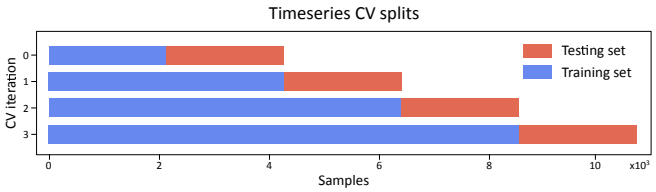



**Fig. 3.** Example of a timeseries 4-fold cross-validation split.

As for the transformer variables used, we have used different combinations of PA, PR and TA (see Section 2.2), always predicting the values between 1 and 20 $ts$ from $t = 0$. The Root-Squared Mean error (RMSE) and its standard deviation (when several values are aggregated, e.g. within folds or over different transformers) is provided as measure the quality of prediction of the model.

### 3.2 Results

As aforementioned, we have used each of the 16 transformers to train a temperature LSTM model that tries to predict the next 20 ts, which is equivalent to a period ranging from 5 to 100 minutes ahead. The performance of this model over all transformers under the two experiments proposed is shown in Table 3.2.

From this table, we can find two main trends. The first tendency is that the generalization error (GEN) is always higher than the CV error, which is coherent with what was expected. TRF-1 and TRF-3 are clearly outliers in both experiments, as it can be seen for an extremely low or high CV error (for TRF-3 and TRF-1 respectively) and an anomalous GEN error over the rest of transformers.

As for the features selected, even in the case of these two transformers, it looks like the addition of TA always enhances the performance of the prediction model. This is even more clear when looking at the AVG* performance. A closer look to the predictions of the model (see Figure 4) reveals a possible explanation. Whereas the PA or PR are closely related to the transformer peaks of temperature, a major contribution to the temperature of the transformer seems to be the weather conditions at the specific location of the transformer. That may explain why the higher-frequency variations are correctly modelled in the left figure (only PA), but the lower-frequency trends of the TA allow our system to

| Exp. | TRF | PA | PA+TA | PR | PR+TA | PA+PR | PA+PR+TA |
|---|---|---|---|---|---|---|---|
| | TRF-0 | 3.76 (1.09) | 1.14 (0.59) | 3.78 (0.85) | 1.36 (0.69) | 3.77 (0.78) | 1.23 (0.60) |
| | TRF-1 | 5.56 (6.33) | 4.78 (5.59) | 5.74 (6.76) | 4.49 (5.14) | 5.69 (6.68) | 4.95 (5.77) |
| | TRF-2 | 3.40 (0.66) | 1.22 (0.74) | 2.97 (0.63) | 1.34 (0.64) | 2.81 (0.52) | 1.38 (0.82) |
| | TRF-3 | **0.76 (1.41)** | **0.80 (1.40)** | **0.73 (1.43)** | **0.72 (1.43)** | **0.74 (1.42)** | **0.74 (1.42)** |
| | TRF-4 | 3.70 (0.79) | 1.10 (0.68) | 2.97 (0.53) | 1.54 (0.75) | 3.32 (0.71) | 1.57 (1.28) |
| | TRF-5 | 3.84 (1.25) | 1.83 (1.25) | 4.11 (1.25) | 2.37 (0.77) | 3.88 (1.14) | 2.36 (1.25) |
| | TRF-6 | 4.88 (1.71) | 2.67 (1.33) | 4.06 (0.57) | 2.67 (1.07) | 4.33 (1.18) | 2.95 (1.45) |
| | TRF-7 | 3.45 (1.01) | 2.08 (1.13) | 3.25 (0.95) | 2.15 (1.14) | 3.15 (1.17) | 2.26 (1.24) |
| CV | TRF-8 | 4.16 (0.99) | 1.78 (0.89) | 4.15 (0.86) | 1.76 (0.69) | 3.97 (0.94) | 1.92 (0.97) |
| | TRF-9 | 4.00 (1.10) | 1.47 (1.01) | 4.42 (1.02) | 1.40 (1.00) | 4.15 (0.86) | 1.56 (1.04) |
| | TRF-10 | 3.73 (1.12) | 1.76 (1.29) | 3.01 (1.06) | 1.98 (1.11) | 2.97 (0.99) | 1.82 (1.36) |
| | TRF-11 | 3.89 (1.12) | 1.68 (1.04) | 3.87 (1.23) | 2.11 (0.78) | 3.75 (1.21) | 1.83 (1.24) |
| | TRF-12 | 3.56 (1.17) | 2.04 (1.02) | 3.70 (1.40) | 2.70 (0.67) | 3.65 (1.32) | 2.51 (1.39) |
| | TRF-13 | 3.34 (0.90) | 1.40 (0.68) | 3.51 (0.84) | 1.81 (0.51) | 3.06 (0.88) | 1.70 (0.81) |
| | TRF-14 | 3.91 (0.98) | 2.13 (1.29) | 3.77 (1.28) | 2.17 (1.33) | 3.66 (1.19) | 2.17 (1.41) |
| | TRF-15 | 3.93 (1.13) | 2.15 (1.22) | 3.48 (0.54) | 2.18 (1.19) | 3.43 (0.68) | 2.41 (1.28) |
| | *AVG** | *3.82 (1.15)* | *1.75 (1.12)* | *3.65 (1.07)* | *1.97 (1.01)* | *3.56 (1.08)* | *1.98 (1.26)* |
| | TRF-0 | 7.49 (3.59) | 4.11 (2.60) | 7.25 (3.05) | 4.15 (3.25) | 7.48 (4.04) | 4.19 (2.63) |
| | TRF-1 | 13.87 (4.20) | 8.04 (2.18) | 14.10 (4.32) | 6.41 (1.90) | 14.09 (4.29) | 10.29 (2.56) |
| | TRF-2 | 7.21 (4.44) | 4.28 (3.78) | 6.95 (4.40) | 4.38 (3.49) | 7.38 (4.68) | 5.56 (4.34) |
| | TRF-3 | 16.08 (3.36) | 13.73 (2.69) | 10.97 (7.48) | 10.15 (6.64) | 14.27 (3.48) | 10.71 (4.13) |
| | TRF-4 | 7.41 (5.02) | 4.20 (3.67) | 7.06 (5.17) | 4.20 (2.52) | 7.55 (5.26) | 4.42 (3.94) |
| | TRF-5 | 8.20 (5.97) | 6.18 (5.72) | 8.72 (6.06) | 4.69 (4.36) | 9.08 (5.91) | 7.66 (6.06) |
| | TRF-6 | 7.15 (2.98) | 5.54 (1.64) | 7.49 (2.99) | 6.06 (1.55) | 7.05 (2.92) | 5.69 (1.63) |
| | TRF-7 | 8.48 (5.84) | 7.60 (5.93) | 8.42 (6.32) | 8.35 (6.34) | 8.89 (7.26) | 8.45 (6.38) |
| GEN | TRF-8 | 7.58 (5.25) | 4.93 (4.26) | 7.35 (5.50) | 4.55 (4.02) | 7.47 (5.38) | 5.11 (4.43) |
| | **TRF-9** | **6.97 (4.17)** | **4.09 (3.22)** | **6.94 (4.80)** | **4.07 (3.00)** | **7.02 (5.36)** | **4.15 (3.20)** |
| | TRF-10 | 8.10 (5.79) | 4.93 (4.89) | 8.36 (6.99) | 5.65 (5.38) | 8.49 (6.81) | 5.35 (4.91) |
| | TRF-11 | 8.39 (6.01) | 5.79 (5.35) | 8.38 (6.68) | 4.65 (4.22) | 8.44 (6.24) | 6.76 (5.59) |
| | TRF-12 | 9.57 (6.97) | 7.18 (6.05) | 8.74 (6.49) | 6.54 (5.40) | 9.36 (6.71) | 8.31 (6.23) |
| | TRF-13 | 7.70 (5.26) | 5.66 (5.50) | 7.71 (5.58) | 5.76 (5.18) | 7.88 (5.57) | 6.16 (5.55) |
| | TRF-14 | 8.38 (2.48) | 5.98 (1.70) | 9.36 (2.59) | 5.91 (1.73) | 8.94 (2.56) | 5.89 (1.69) |
| | TRF-15 | 7.60 (5.23) | 5.07 (4.56) | 7.27 (5.57) | 5.18 (4.81) | 7.49 (6.01) | 5.54 (5.05) |
| | *AVG** | *7.87 (5.11)* | *5.40 (4.56)* | *7.86 (5.38)* | *5.30 (4.33)* | *8.04 (5.55)* | *5.95 (4.85)* |

** AVG: Average RMSE excluding TRF-1 and TRF-3*

**Table 1.** RMSE and average RMSE (AVG) results for all the experiment and transformers, using different combinations of the input features (PA, PR and TA).

provide a much better prediction, even when generalizing to another transformer (trained with TRF-9, predicting TRF-2).

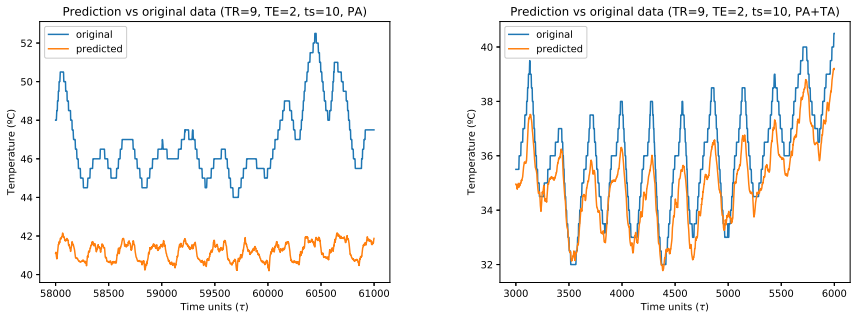



**Fig. 4.** Prediction of the time series of Transformer 0 test set and extrapolation of the model to TRF-2, when using a time-step prediction of 10 (50 minutes).

For a given transformer, the prediction error hardly varies with the prediction steps. This can be checked at Figure 5, where the error varies as expected within the CV experiment, since it increases over time (above), but this effect is far less evident in the GEN experiment (below).

From Figure 5 and Table 3.2 we can observe that the variables of PA (and PR) are useful for the modelling of the high frequency components (fast variations) in the temperature. However, the addition of TA corrects general temperature trends primarily related to climate conditions, achieving a much more accurate prediction. The combination of PA+TA seems to provide the best results, although very similar to PR+TA and PR+PA+TA. Given that PA+PR does not pose any improvement over using solely either PA or PR, it could be concluded that they do not provide additional information about the temperature variations in the transformer.

As for the prediction step $\tau$, we can observe that the predictions are safer enough with more than $10\tau$, equivalent to 50 minutes. Therefore, in the case of an abnormal temperature rise, our system could provide early warnings that could help reconfigure the energy distribution network in order to avoid a transformer overload and subsequent problems.

This model makes it possible to predict with a low error the temperature value using only the PA variables of the three lines, together with the ambient temperature at this time, at least 50 minutes in advance. This would make it possible to discover trends towards anomalous temperature values in different transformers and act accordingly by planning a redistribution of the workload, avoiding possible incidents or service interruptions.

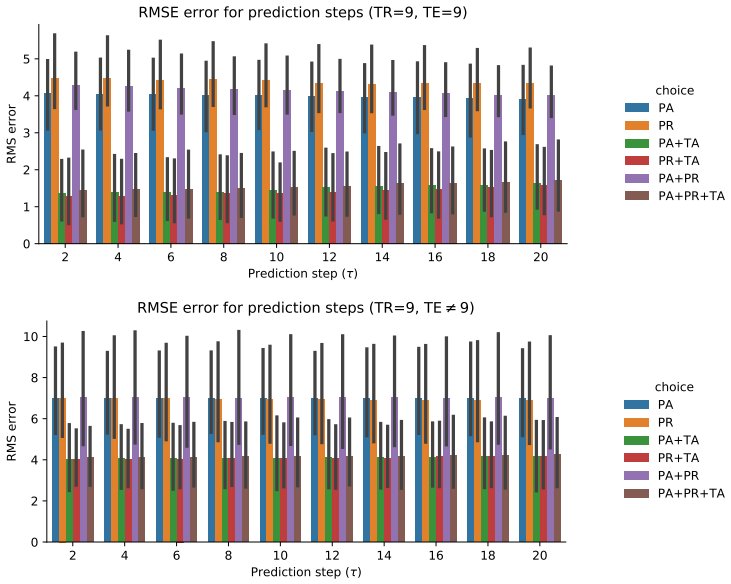



**Fig. 5.** RMSE error for the different prediction steps when trained and tested in transformer 9 (above, using TS-CV) and when extrapolating the model fitted with TRF=9 to all remaining transformers (below).

## 4 Conclusions

The near future of energy is conformed by a plethora of heterogeneous sources as well as an increasing demand. This poses new challenges for energy production and distribution, in which it will be essential that Medium Voltage/Low Voltage (MV/LV) become smart distribution grids. This works tackles the problem of predicting the transformer temperature at each node of these grids, a fundamental tool to create an intelligent control system that can discover trends towards anomalous temperature values in different transformers and act accordingly by planning a redistribution of the workload, avoiding possible incidents or service interruptions. In this context, we proposed a system based on LSTM networks, a very recent advance in the neural network field, from which we could predict with a low error the temperature value using only the active power of the three lines at each transformer, together with the ambient temperature at every instant. A prediction up to 100 minutes –using the last 100 minutes– was possible with a small RMSE, proving the ability of this architecture, allowing for a fast integration in smart planning and control of energy distribution networks.

**Acknowledgments.** This work was supported by the PASTORA: PREVENTIVE ANALYSIS OF INTELLIGENT NETWORKS IN REAL TIME AND INTEGRATION OF RENEWAL RESOURCES (ITC-20181102) funded by the

Centre for Industrial Technological Development (CDTI), supported by the Ministry of Science, Innovation and Universities and co-financed by the European Union with ERDF funds through the "Plurirregional Operational Programme of Spain 2014-2020". It was also partially supported by the MINECO/ FEDER under the TEC2015-64718-R, RTI2018-098913-B-I00 and PGC2018-098813-B-C32 projects. Work by F.J.M.M. was supported by the MICINN "Juan de la Cierva" Fellowship.

## References

1. T. Stetz, F. Marten, and M. Braun. Improved Low Voltage Grid-Integration of Photovoltaic Systems in Germany. *IEEE Transactions on Sustainable Energy*, 4(2):534–542, April 2013.
2. A. Kahrobaeian and Y. A. I. Mohamed. Analysis and Mitigation of Low-Frequency Instabilities in Autonomous Medium-Voltage Converter-Based Microgrids With Dynamic Loads. *IEEE Transactions on Industrial Electronics*, 61(4):1643–1658, April 2014.
3. S. L. Patil, H. J. Tantau, and V. M. Salokhe. Modelling of tropical greenhouse temperature by auto regressive and neural network models. *Biosystems Engineering*, 99(3):423–431, March 2008.
4. Bao Rong Chang and Hsiu Fen Tsai. Forecast approach using neural network adaptation to support vector regression grey model and generalized auto-regressive conditional heteroscedasticity. *Expert Systems with Applications*, 34(2):925–934, February 2008.
5. N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. Predicting solar generation from weather forecasts using machine learning. In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 528–533, October 2011.
6. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
7. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
8. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
9. F. J. Martinez-Murcia, A. Ortiz, J.M. Gorriz, J. Ramirez, and D. Castillo-Barnes. Studying the Manifold Structure of Alzheimer's Disease: A Deep Learning Approach Using Convolutional Autoencoders. *IEEE Journal of Biomedical and Health Informatics*, Online(0), June 2019.
10. Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.
11. Sepp Hochreiter and Jrgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

12. K. Chen, Y. Zhou, and F. Dai. A LSTM-based method for stock returns prediction: A case study of China stock market. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2823–2824, October 2015.
13. A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, May 2013.
14. Hasim Sak, Andrew Senior, and Francoise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. page 5.
15. Douglas Eck and Jrgen Schmidhuber. Learning the Long-Term Structure of the Blues. In Jos R. Dorronsoro, editor, *Artificial Neural Networks ICANN 2002*, Lecture Notes in Computer Science, pages 284–289. Springer Berlin Heidelberg, 2002.
16. Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, December 2003.
17. Jonathan D. Power, Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154, February 2012.
18. Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.
19. Francisco J Martinez-Murcia, Juan M Gorriz, Javier Ramirez, and Andres Ortiz. Convolutional Neural Networks for Neuroimaging in Parkinson's Disease: Is Preprocessing Needed? *International journal of neural systems*, pages 1850035–1850035, 2018.
20. Jeffrey D. Hart. Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-Validation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):529–542, 1994.
21. Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

# Knowledge Extraction (KnoX) in Deep Learning: Application to the *Gardon de Mialet* Flash Floods Modelling

Bob E. Saint Fleur [1,2], Guillaume Artigue [1], Anne Johannet [1], Severin Pistre [2]

[1] IMT Mines Alès, Laboratoire de Génie et de l'Environnement Industriel (LGEI), Alès, France
[2] Hydrosciences, Univ Montpellier, CNRS, IRD, 34090 Montpellier, France

Corresponding author Guillaume ARTIGUE (guillaume.artigue@mines-ales.fr)

**Abstract.** Flash floods frequently hit Southern France and cause heavy damages and fatalities. To better protect persons and goods, official flood forecasting services in France need accurate information and efficient models to optimize their decision and policy. Since heavy rainfalls that cause such floods are very heterogeneous, it becomes a serious challenge for forecasters. Such phenomena are typically nonlinear and more complex than classical floods events. That problem leads to consider complementary alternatives to enhance the management of such situations. For decades, artificial neural networks have been very efficient to model nonlinear phenomena, particularly rainfall-discharge relations in various types of basins. They are applied in this study with two main goals: first modelling flash floods on the *Gardon de Mialet* basin; second, extract internal information from the model by using the Knowledge eXtraction method to provide new ways to improve models. The first analysis shows that the kind of nonlinear predictor influences strongly the representation of information: e.g. the main influent variable (rainfall) is more important in the recurrent and static models than in the feed-forward one. For understanding flash floods genesis, recurrent and static models appear thus as better candidates, even if their results are better.

## 1 Introduction

In the Mediterranean regions, flash floods due to heavy rainfalls frequently occur and cause numerous fatalities and costly damages. During the last few years, the south of France has been particularly exposed to these catastrophic situations. In such cases, damages can reach more than one billion euros, and, in only one event, there can be more than 20 fatalities [1]. Facing these issues, authorities need reliable forecasts for early warning purposes. Unfortunately, both the short-term rainfall forecasts and the processes leading to the discharge response remain poorly known at the space and time scales required. It is thus difficult to provide forecasts using the traditional coupling between a meteorological model and a physically based hydrological model.

Artificial Neural Networks therefore appear as an alternative paradigm as they are able to provide forecasts of an output (discharge) without making any other hypothesis

on the system than the causality between rainfall and discharge. ANN have been applied in a wide variety of domains as they are essentially based on data and training [2]. They appear as particularly suitable for identifying the generating processes in hydrological time series because of their ability to model nonlinear dynamic systems [3,4]. However, due to their statistical origin, it is difficult to associate meaning to their internal parameters and they are rightly considered as black-box models. For this reason and to enhance the understanding of the behavior of the model, several works have been done to bring more transparency in the operating mode and introduced concepts of gray-box and transparent-box models [5,6]. In hydrology, several works have been conducted to make neural networks models more physically meaningful [6, 7, 8].

To be considered as gray-box (or transparent-box), ANN internal information or data must be accessible. In this paper, it will not be discussed *deep learning* itself, but an intermediate method to analyze the meaningful of internal information about neuronal models in hydrology operating on deep models. That method is termed ***Knowledge eXtraction*** (KnoX), it has been proposed by [7]. It was proved efficient on a fictitious basin, before being applied, by simulation, to estimate contributions and response times of various parts of a karst aquifer: the *Lez* aquifer (Southern France). It was later used by [8] for better apprehend the contributions of surface or underground processes in generation of floods on the Lavallette basin (Southern France).

Several studies were performed on the *Mialet* basin: first [4] showed that flash flood discharge can be forecasted by a multilayer perceptron with reasonable quality up to two-hours lead time; second, [9] showed that the initial value of the neural network parameters in flash floods forecasting has a major impact on the result. The purpose of this work is thus to better understand how the main variables influence the basin's outflow, regarding the model scheme used, in order to diminish the sensitivity of the model to the initialization of its parameters.

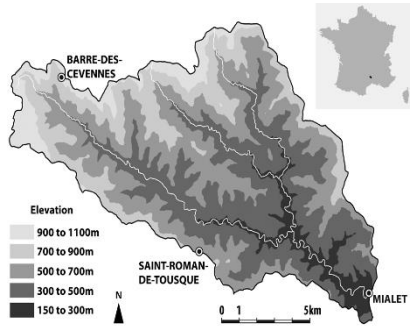
In the next sections, we will briefly present neural networks, their operating principles in hydrology, the deep multilayer perceptron used, as well as a reminder about the KnoX method and the models designed. The focus is set on a discussion about the behavior of the variable's weights according to the model type used, by applying the KnoX method to extract that information.

## 2 Materials and methods

### 2.1 Study area: location and general description

The *Gardon de Mialet* basin covers 220 sq.km in southern France. It is part of the *Cévennes* range which is known as a preferential location for the well-known meteorological phenomenon named *cevenols episodes* (Fig. 1). These episodes consist in short duration (less than 2 days) very heavy rainfall events. The elevation of Mialet basin ranges from 150 *m.a.s.l.* to 1170 *m.a.s.l.* and its mean slope is about 33 %. As for the most of basins of the *Cévennes*, these characteristics lead to limited infiltration or underground flow and thus to a high drainage density. Its response time is relatively short: between 2-4 hours [4]. The area is dominated by a metamorphic formation essentially with 95 % of mica-schist and gneiss, which lead to a poorly porous and impermeable

rocky sub-soil. The land use is almost homogeneous while covered by natural vegetation (chestnut trees, conifers, mixed forest and bush) for 92 %. The rest is shared between rocks and urban areas.




Typically, in Mediterranean regions, heavy rainfalls sometimes exceed 500 mm in only 24 h, to be compared to the 600 mm that fall on Paris annually. They are mainly produced by convective events, triggered either by relief, by a wind convergence, or by both. For example, in September 2002, the Gard (France) department has registered 687 mm of rainfall in 24h with 137 mm in only one hour at *Anduze* (a few km distant from *Mialet*).

**Fig. 1.**The study area (by Artigue, 2012)

## 2.2 Database

### 2.2.1. Presentation.

The database used in this study is essentially compounded with hourly observations from 1992 to 2002 and 5 minutes time step observations from 2002 to 2008 on three rain gauges and one hydrometric station at the outlet at *Mialet* (Fig. 1). From upstream to downstream, these stations are: BDC (*Barre des Cévennes*), SRDT (*Saint-Roman de Tousque*) and *Mialet* which coincide with the discharge station. They are all managed by the local Flood Forecasting Service (*SPC Grand Delta*). 58 events were extracted at 30 min time-step (based on rainfall events having at least 100 mm accumulation in 48 h on any of the rain gauges). Data description is synthetized in Tables 1 & 2.

**Table 1.** Data description

| | Rainfall (mm) | | | Discharge | |
|---|---|---|---|---|---|
| | ***BDC*** | ***SRDT*** | ***Mialet*** | ***($m^3s^{-1}$)*** | ***($m^3s^{-1}km^{-2}$)*** |
| *Maximum (30 min)* | 33.3 | 41.8 | 62.0 | 819.3 | 3.72 |
| *Median (30 min)* | 0.3 | 0.3 | 0.2 | 29.3 | 0.13 |
| *Moy* | 1.0 | 1.3 | 1.2 | 43.4 | 0.20 |
| *Min* | 0 | 0 | 0 | 2.13 | 0.010 |

**Table 2.** Test event description

| Event | Date | Duration | Maximum of discharge ($m^3s^{-1}$) | Mean discharge ($m^3s^{-1}$) | Cumulative rainfall (mm) | Intensity ($mm.h^{-1}$) |
|---|---|---|---|---|---|---|
| 13 | Sept. 00 | 26 h | 454,2 | 70 | 230 | 40 |

## 2.3 Artificial Neural network

### 2.3.1. General presentation.

A neural network is a combination of parametrized functions called neurons that calculate their parameters thanks to a database using a training process [10]. The most

popular model is the multilayer perceptron (MLP), which generally contains one or more hidden layers of nonlinear neurons and one output linear neuron. Each hidden neuron computes a non-linear function of a weighted sum of the input variables, then the output neuron computes the linear combination of the outputs of the hidden ones.

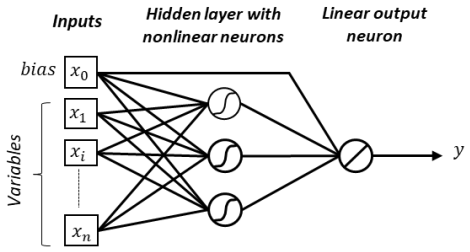



**Fig. 2.** Multilayer perceptron with a single hidden layer

The MLP is very popular due to its two main properties: universal approximation and parsimony. The first one states the capability to successfully approximate any differentiable function with an arbitrary level of accuracy [11]. The latter states how the multilayer perceptron needs fewer parameters to successfully fit a non-linear function, compared to others statistic model that linearly depend on their parameters [12]. The more general model of neuron calculates it output *y* as following:

$$y = f\left(\sum_{j=1}^{n} c_j \,.\, x_j\right) = f(x_1, \dots, x_n;\ c_1, \dots, c_n), \tag{1}$$

with $x_j$, the input variable *j*; $c_j$, the parameter linking the variable $x_j$ to the neuron; $f(.)$, the activation function (usually a sigmoid). The dynamic properties of the identified process can be considered thanks to three kinds of models [13].

- ***Static model***

The static model is a digital filter with a finite impulse response. It calculates the following equation:

$$\hat{y}(k) = \varphi(\,\mathbf{x}(k), \dots, \mathbf{x}(k - n_r + 1), \mathbf{C}) \tag{2}$$

with $\hat{y}(k)$**,** the estimated output at the discrete time *k*; $\varphi_{\mathrm{rn}}$, the non-linear function implemented by the model; **x** is the input vector**;** ***n*,** the sliding time-windows size defining the length of the necessary exogenous data**; C**, the vector of the parameters. This model is known for having more parameters than the following models.

- ***Recurrent model***

The recurrent model allows identification of dynamical processes (Infinite Impulse Response), it is implemented following the equation (3).

$$\hat{y}(k) = \varphi(\,\hat{\mathbf{y}}\,(k-1), \dots, \hat{\mathbf{y}}\,(k-r);\ \mathbf{x}(k), \mathbf{x}(k-1) \dots, \mathbf{x}(k - n_r + 1);\ \mathbf{C}) \tag{3}$$

With *r*, the order of the recurrent model; $n_r$, the depth of the sliding time-window used to consider the input variables. Ones must distinguish the recurrent variable (*y*) from the exogenous variables (**x**). This model can deliver forecasts for an undetermined forecasting horizon providing the availability of the exogenous variables.

- ***Feed-forward model***

In the feed-forward model, the recurrent input is substituted by the measurements of the process output at previous times step. This model is non recurrent; but it can identify dynamical processes. This model is the most used and generally provides the best results. Nevertheless, we have observed that it generally has difficulties to model the dynamics of the process (cited in Artigue et al 2012). It calculates:

$$\hat{y}\,(k) = \varphi(\mathbf{y}\,(k-1), \ldots, \mathbf{y}(k-r);\ \mathbf{x}(k), \mathbf{x}(k-1)\ \ldots, \mathbf{x}(k-n_r+1);\ \mathbf{C}) \qquad (4)$$

with $y(.)$, the observed value of the modelled variable at the discrete time *k*.
These three categories of models will be compared in this study.

**2.3.2. Training**

As data-driven models, neural networks design is based on a database. Training consists in calculating the set of parameters of the model in order to minimize the least square cost function on the training set [10]. Because the model is non linear, this minimization is iteratively calculated.

Nevertheless, as the goal of the model is to be able to generalize the trained behavior to any set of data never seen, the quality of the model must be validated on another set, independent from the training set that is called "test set". The bias-variance dilemma [14] shows an important limitation: the training error is not representative of the test error, and the difference increases with the complexity of the model (*i.e.* the number of free parameters of the model). The bias-variance dilemma may be avoided using regularization methods.

**2.3.2. Regularization methods**

*Early stopping*

Early stopping was presented by [15] as a regularization method. It consists in stopping the training before the full convergence. To this end a supplementary subset, called stop set, is defined those goal is to evaluate the ability of generalization of the model during the training. This subset is independent from the training set. Training is stopped when the error on the stop set begins to increase. The stop set is used to stop the training, the performances of this set are thus overestimated compared to those of the test set. Nevertheless, this set is usually (improperly) called "validation set" in the literature.

*Cross validation*

Proposed by [16], cross validation allows to select a model having the lower variance. To this end the training set is divided in *K* subset and the error is calculated on the remaining (*K*-1) subsets in the training set. After *K* trainings, the cross-validation score is calculated, for example by the mean of the previously obtained errors. Based on the cross validation score it is possible to select the model that has the lowest variance, minimizing by this way the bias on the training set and the variance on validation sets. This method allows to select input variables, the order (r), and the number of hidden neurons.

*Ensemble model*

Darras et al. [9] showed that, surprisingly, cross validation was not able to successfully select the best initialization of parameters. In order to diminish the sensitivity of the output to the parameter's initialization, they propose to create an ensemble model of $\mathcal{M}$ members [17] and to calculate the output of the ensemble, at each time step, by the median of the $\mathcal{M}$ members.

**2.3.3 Design of the model**

In this study, regularization methods are applied by: (*i*) dividing the dataset in three subsets (training, stop and test sets), (*ii*) using cross correlation to select the architecture of the model in the following succession: inputs ($n_r$) except for rain gauges, order ($r$), number of hidden neurons ($h$), and (*iii*) using 20 members in the ensemble.

Three kinds of sliding window widths are tried based on the rainfall-runoff cross-correlogram.

**2.3.4. Performance criteria**

Several criteria are used to assess the performance of a model. The determination coefficient $R^2$ [18]; the *Synchronous percentage of the peak discharge ($S_{PPD}$)* and the *Peak delay* as two peak assessment criteria [4]. They are briefly detailed below:

- **$R^2$ criterion**

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{n}(y_k - \overline{y_o})^2}, \quad (5)$$

with the same notations as previously.

The nearest than 1 the Nash-Sutcliff efficiency is, the best the results are. Nevertheless, this criterion can reach good values even if the model proposes bad forecasts.

- **Peak analysis**

The quality of the flood prediction is analyzed regarding the quality of the peak using two criteria defined by [4].

Synchronous percentage of the peak discharge: SPPD

The synchronous percentage of the peak discharge: SPPD [4] is a relevant criterion to assess flash flood modeling performance of a model on the peak discharge. It shows the simulation quality at the peak discharge through the ratio between the observed and simulated discharges at the observed peak discharge moment ($k_o^{max}$).

$$S_{PPD} = 100\frac{\hat{y}^{k_o^{max}}}{y^{k_o^{max}}}, \quad (6)$$

Peak delay ($P_D$)

The peak delay [4] indicates the duration between the maximum of simulated peak and measured peak. When the estimated peak is in advance, the peak delay is negative.

$$P_D = k_s^{max} - k_o^{max}, \quad (7)$$

with $k_{\cdot}^{max}$ the instant of the peak of discharge (simulated or observed).

### 2.5. Extracting information: KnoX method

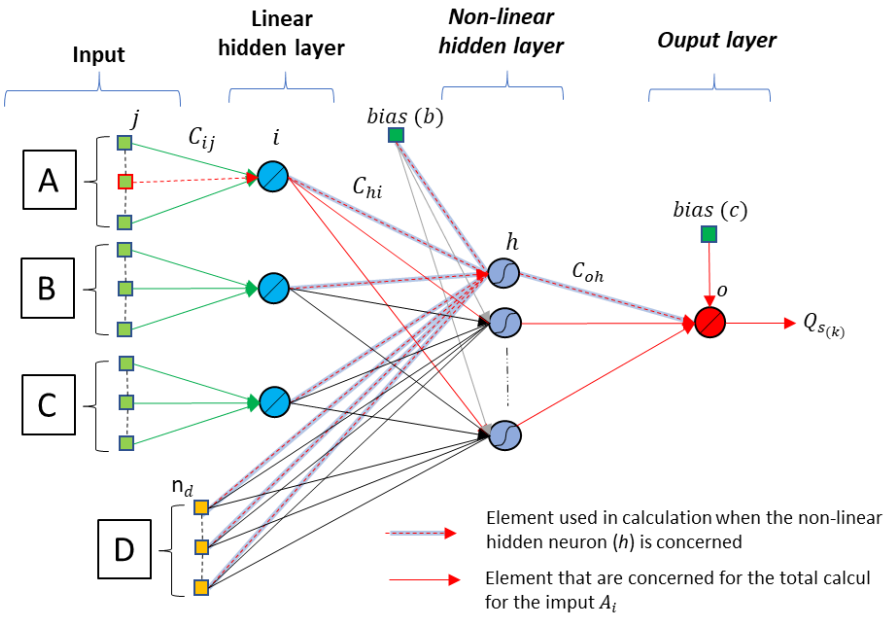



**Fig. 3.** Application of the KnoX method on the deep multilayer perceptron

$$P_{A(j)} = \frac{{}^{M}|C_{ij}|}{\sum_{i=1}^{n_A} {}^{M}|C_{ij}|} \sum_{h=1}^{H} \left( \frac{{}^{M}|C_{hi}|}{\sum_{i=1}^{n_i} {}^{M}|C_{hi}| + \sum_{d=1}^{n_d} {}^{M}|C_{hd}| + b_h} \right) \left( \frac{{}^{M}|C_{oh}|}{\sum_{h=1}^{H} {}^{M}|C_{oh}| + c_o} \right), \quad (8)$$

and: $P_A = \sum_{j=1}^{n_A} (P_{A(j)})$ (9)

The KnoX method [8, 19] allows to calculate a simplified contribution of each input to the model output. This method is described for the general deep model (2 hidden layers) shown in Fig. 3. The principle of the method is that a contribution of an individual input variable can be quantified after training, by the product of the parameters linking this input to the output. The considered parameters are (*i*) "normalized" by the sum of the parameters linked to the same targeted neuron, and (*ii*) regularized by calculating the median of absolute values of their values for 20 different random initializations. This regularized value is noted as ${}^{M}|C_{ij}|$ for the parameter $C_{ij}$ linking the neuron (or input) *j* to the neuron *i*.

Regarding the model shown in Fig. 3, the contribution ($P_A$) of the input *A* (group of several delayed inputs) is the sum of the contributions of each individual delayed input of the group A. The equation calculating the contribution for just one element of the

input A is provided in eq. (8). It is not possible to explain more comprehensively the method in the short present paper, so we suggest to the reader to refer to [8].

## 3 Results

Starting from previous works of [4] we chose the following exogenous variables: (*i*) *Barre des Cevennes* rain gauge, *Saint-Roman de Tousque* rain rauge and *Mialet* rain gauge, each one with a sliding window length $\{k, \dots k\text{-}n_r\text{+}1\}$, (*ii*) the sum of the mean rain (over the three gauges) fallen from the beginning of the event. Of course, a bias input is used; several values were tried in order to evaluate the sensitivity of the KnoX method to its value.

### 3.1. Window widths selection thanks to correlation analysis

Widths of the rainfall windows applied to the model are selected thanks to cross correlation. Initially proposed by [20] Jenkins and Watts (1968), [1] generalizes the application of cross correlation in hydrology. The used equation in this study is presented in eq. (9).

$$C_{xy}(k) = \frac{Cov\,(x_i, y_{i+k})}{\sigma_x\,\sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(x_i - \bar{x})(y_{i+k} - \bar{y})}{\sigma_x\,\sigma_y} \quad , \tag{9}$$

With $k = 0, 1, \dots$ ;. where $m$ is the truncation which is recommended to be $m=n/3$ (Mangin 1984). [20]) indicated that 2 hydrological variables can be considered as statistically independent if their cross-correlation is superior to 0.2. We thus select three possible lengths for the sliding windows of rain gauges inputs: (*i*) the number of time step between $C_{xy}$ =0 and $C_{xy}$ =0.2, that defines the memory effect; (*ii*) the window between $C_{xy}$ =0.2 (positive slope) and $C_{xy}$ =0.2 (negative slope) and (*iii*) all the $m$ positive values of $C_{xy}$. Based on [20] the correlations between gauges as well as response times are indicated in Table 3.

**Table 3.** Correlation analysis of the data

| *Rain gauge* | | Mialet (h) | SRDT (h) | *BDC* (h) |
|---|---|---|---|---|
| Average response-time | | 2 | 3 | 4.5 |
| Response-time range | | 1 – 3.5 | 2.5 – 4.5 | 4 – 5.5 |
| Rainfall-discharge average cross-correlation | | 0.40 | 0.455 | 0.44 |
| Rain gauge cross-correlation | Mialet | -- | 0.58 | 0.45 |
| | SRDT | -- | -- | 0.61 |

### 3.2. Model selection

A partial cross-validation score was operated on a subset of 17 most intense events in the database [3]. The number of hidden neurons was increased from 1 to 10. The best

model was chosen according to the highest cross-validation score $S_v$ estimated as following:

$$S_v = \frac{1}{K}\sqrt{\sum_{i=1}^{K}|E_i|^2}, \quad (9)$$

Where $E_i$ is the validation error of the subset *i* used in partial cross validation.

The output values are the result of the median of the outputs of an ensemble of 20 members differing only by their initialization before training.

Three bias values are considered (0.01; 0.1; 1), three depths of sliding windows (see section 3.1) and three kinds of models (see section 2.3), 27 different models have been designed following the procedure indicated in section 2.3.3. The best one in each kind of models has been chosen, regarding the test event, in order to have efficient models to analyze. Architectures presented in Table 4 were thus selected.

**Table 4.** Selected models

| **Input variables** | **Static** | **Recurrent** | ***Feed-Forward*** |
|---|---|---|---|
| Rain-gauge window width ($n_r$) (BDC/SRDT/Mialet) | 32/32/23 | 27/28/20 | 32/32/23 |
| Rain cumul window width | 3 | 3 | 3 |
| Order (r) | x | 3 | 3 |
| Number of hidden nonlinear neurons | 2 | 10 | 5 |
| Bias value | 1 | 0.01 | 1 |

### 3.3 Results

Obtained test set hydrographs are shown in the Fig 4 and their performances described in Table 5. It appears in
**Fig. 4** and **Table 5** that the best results are provided by the feed-forward model. This is usual because the feedforward model uses the previous observations of the modelled variable in input. The recurrent model is usually not as efficient but exhibits better dynamics, which is also frequently observed [4]. The static model presents an acceptable performance, being able to generate 63% of the peak discharge.

**Table 5.** The models performances on the test set

| *Model* | *$R^2$* | *SPPD %* | *$P_D$(0.5h)* |
|---|---|---|---|
| Static | 0.83 | 63,3 | 1 |
| Recurrent. | 0.89 | 78.5 | 0 |
| Feed-Forward | 0.99 | 99.3 | 1 |

After having verified that the models are convenient, it is possible to apply the KnoX method. The extracted contributions are presented in **Table 5**.

Regarding the rainfalls, one can note that in general, SRDT is the station with the highest contribution. The contributions do not change significantly for *Mialet* through all the models. BDC and *Mialet* are probably affected by their location close to the border of the basin whereas SRDT is close to the middle of the basin.

Regarding the balance between the state variables and the rainfalls, it appears that when the previous observed discharge is used as an input variable, it brings almost 50

% of the contribution to the output. This observation means that the model does not pay enough attention to rain inputs and this could be the reason of the sensitivity to parameters initialization. Beside this, it also appears that the state variables in the static model have lesser contribution than they do in the other two models. In general, from the static model to the feed-forward one, the total contributions of the state variables are respectively 45%, 61 % and 65 %, where the biggest parts are imputed to the previous observed discharge (feed-forward). These observations are fully consistent and the results seem highly interpretable.

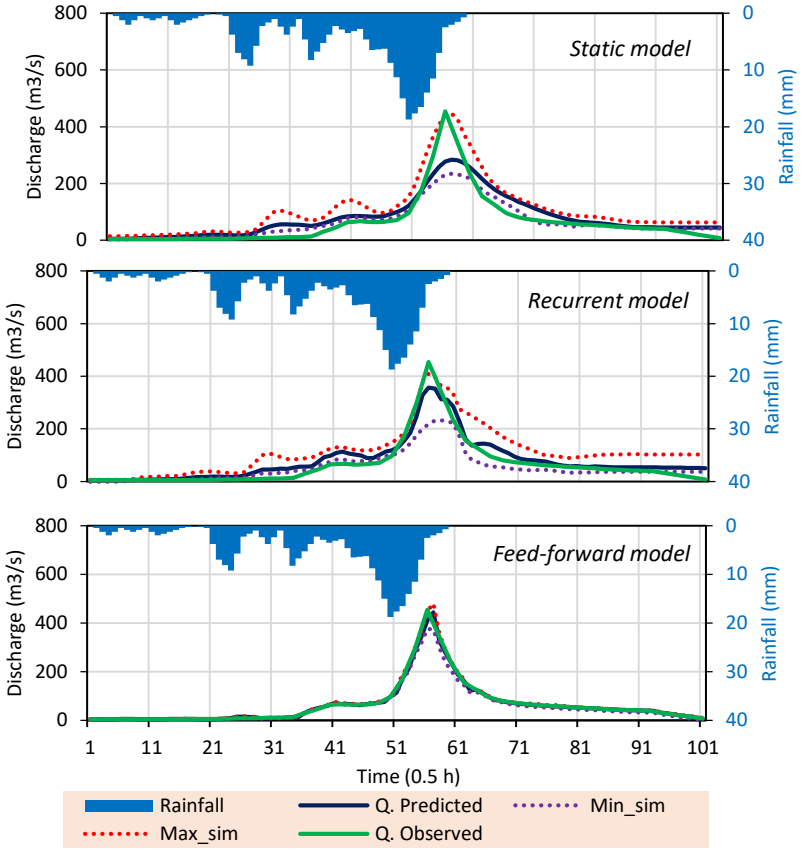



**Fig. 4**. Hydrographs for the test set. Min_sim and Max_sim correspond to the minimum and maximum values of the ensemble model. Q is the median of the 20 members of the ensemble.

**Table 6.** Contributions ($P_A$) for the variables, from each model, expressed in %.

| ***Name of variable*** | ***Static*** | ***Recurrent*** | ***Feed-forward*** |
|---|---|---|---|
| *BDC* | 13 % | 12 % | 5 % |
| *SRDT* | 31 % | 17 % | 22 % |
| *Mialet* | 11 % | 11 % | 9 % |
| *Cumulated rainfall* | 31 % | 20 % | 12 % |
| *Previous Q. obs* | -- | -- | 45 % |

| | | | |
|---|---|---|---|
| *Previous Q. calc* | -- | 25 % | -- |
| *bias* | 14 % | 16 % | 8 % |

## 4 Interpretation

These results show how the kind of model can modify the contribution of explanatory variables on an observed phenomenon. Thus, some kind of models must be preferred when it comes to represent physical relations. It is also shown that the mean cumulative rainfall used here as a state variable plays a great role in models where the previous discharge is not used as input. This state variable seems to have a great interest in hydrologic modelling. The value of the bias, surprisingly, seems to have a role. It is usually interpreted as the base flow. Nevertheless, its behavior is consistent: it shows more involvement when the previous observed discharges are not used as input; then by complementarity with the humidity information, it guides the models to acceptably approximate the real discharge information.

## 5 Conclusion

Prediction of flash flood events is a very challenging task in the *Cévennes* range. It was previously realized using neural networks but sometimes appeared difficult to understand because of the specific behaviors of the models. In order to be able to improve these models, the present work takes steps to better understand the processes involved in such events. To this end, the KnoX method, developed to extract information from a neural network model was applied to the *Gardon de Mialet* Basin. The obtained results show that by using relevant variables properly combined on whatever the network used here, efficient model can be built out. Besides, the KnoX method allows to see how the variables are handled by the model to approximate the phenomenon. There has been evidence that the variables do not express themselves in the same way through the different models used. As it is understandable, sometimes, the choice for a model is commanded by the situations in presence. The information extracted from the network can probably be used to compare to some physical meaningful characteristics of watershed or events, such as the Thiessen polygons, the response time, the cross correlation etc. It provided also some guidelines to deal with the sensitivity of the model to the parameter's initialization.

## 6 Aknowledgement

The authors thank the METEO-France weather agency, the SPGD flood-forecasting agency for providing rainfall datasets. Our gratitude is extended to Bruno Janet for the stimulating collaboration shared with the SCHAPI Unit, and to Roger Moussa and Pierre Roussel-Ragot for the helpful discussions and support. The constant effort made by Dominique Bertin and the Geonosis Company to enhance and develop the neural network software RNF Pro are thereby acknowledged as well.

## 7 References


1. Rouzeau, M., Xavier M., and Pauc, J.C. 2010. “Retour d’expériences des inondations survenues dans le departement du Var les 15 et 16 juins 2010.” http://cgedd.documentation.developpement-durable.gouv.fr/documents/cgedd/007394-01_rapport.pdf.
2. J. Roberts, Stephen, and Will Penny. 1981. Neural Networks: Friends or Foes? Sensor Review. Vol. 17. London: MCB University Press.
3. Toukourou M., Johannet A., Dreyfus G.,Ayral P.A. 2011. Rainfall-runoff Modeling of Flash Floods in the Absence of Rainfall Forecasts: the Case of "Cévenol Flash Floods", App. Intelligence, 35 2,178-189.
4. Artigue, G,et al. 2012. "Flash Flood Forecasting in Poorly Gauged Basins Using Neural Networks: Case Study of the Gardon de Mialet Basin (Southern France)". NHESS, 12(11): 3307-24.
5. Oussar, Yacine, and Gérard Dreyfus. 2001. "How to Be a Gray Box: Dynamic Semi-Physical Modeling." Neural Networks 14 (9): 1161-72. https://doi.org/10.1016/S0893-6080(01)00096-X
6. Johannet, Anne, B Vayssade, and Dominique Bertin. 2007. "Neural Networks: From Black Box towards Transparent Box - Application to Evapotranspiration Modelling." Int. Journal of Comp. Int. 24 (1): 162.
7. Kong-A-Siou, L., et al, S.: KnoX method, or Knowledge eXtraction from neural network model. Case study on the Lez karst aquifer (southern France), J. Hydrol., 507, 19–32.
8. Darras, T., et al. 2015. Identification of spatial and temporal contributions of rainfalls to flash floods using neural network modelling: case study on the Lez basin (southern France) Hydrol. Earth Syst. Sci., 19, 4397–4410, 2015
9. Darras, T., Johannet, A., Vayssade, B., Kong-A-Siou, L. and Pistre, S. (2014). Influence of the Initialization of Multilayer Perceptron for Flash Floods Forecasting: How Designing a Robust Model, (ITISE 2014), Ruiz, IR, Garcia, GR Eds, 687-698.
10. Dreyfus, G. 2005. Neural networks, methodology and applications, Springer, Berlin.
11. Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." Neural Networks 2 (5): 359-66.
12. Barron, A R. 1993. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function." IEEE Trans. Inf. Theor. 39 (3): 930-45. https://doi.org/10.1109/18.256500.
13. Nerrand, O., P. Roussel-Ragot, L. Personnaz, G. Dreyfus, and S. Marcos. 1993. "Neural Networks and Nonlinear Adaptive Filtering: Unifying Concepts and New Algorithms." Neural Comp 5 (2): 165-99.
14. Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. "Neural Networks and the Bias/Variance Dilemma." Neural Computation 4 (1): 1-58.
15. Sjöberg, J., et al. 1995. "Nonlinear Black-Box Modeling in System Identification: A Unified Overview." Automatica 31 (12): 1691-1724.
16. Stone, M. 1976. "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)." Journal of the Royal Statistical Society: Series B (Methodological) 38 (1): 102-102.
17. Dietterich T.G., 2000. Ensemble Methods in Machine Learning, in J. Kittler and F. Roli (Ed.), First Int. Workshop on Multiple Classifier Systems, LNCS, p. 1-15, New York: Springer Verlag.
18. Nash, J.E., and J.V. Sutcliffe. 1970. "River Flow Forecasting through Conceptual Models Part I - A Discussion of Principles." Journal of Hydrology 10 (3): 282-90.
19. Jenkins, G.-M., Watts, D.-G. 1969 “Spectral analysis and its applications”, Holden-Day, - 525 pages
20. Mangin, Alain. 1984. "Pour Une Meilleure Connaissance Des Systèmes Hydrologiques à Partir Des Analyses Corrélatoire et Spectrale." Journal of Hydrology 67 (1-4): 25-43.

# The Study of Recurrent Neuron Networks based on GRU and LSTM in Time Series Forecasting

Tatiana Afanasieva[1][0000-0003-3779-7992] and Pavel Platov[1]

[1] Ulyanovsk University, Ulyanovsk, Severny Venec, 32, Russia
tv.afanasjeva@gmail.com

**Abstract.** The article focuses on a comparative study of the prediction accuracy of the above mention configurations of recurrent neural networks for time series, having a trend, or seasonality, or their combination. The study was conducted on a dataset of 30 artificial time series. A methodology for such an investigation and corresponding software for time series modelling and forecasting have been proposed. LSTM shows a slightly better accuracy (within 1 percent) on average and in most cases compared to GRU in predicting time series. At the same time GRU has demonstrated a better prediction accuracy for 40% of time series Experimental study showed that forecasting accuracy for time series having a trend and random component was the best on average for both LSTM and GRU in comparison with time series with seasonal component. The obtained results expand our understanding of the applicability of recurrent neural networks with GRU or LSTM in forecasting of time series with different behavior**.**

**Keywords:** Recurrent neural networks, Prediction, Time series, Accuracy.

## 1 Introduction

Recurrent neural networks (RNNs) are defined as a class of supervised machine learning models, made of artificial neurons with one or more feedback loops [1].

In general, these networks are made from nonlinear but simple units, enabled to store, remember, and process past complex signals for long time periods. In this way RNNs can learn the temporal context of input sequences, map an input sequence to the output sequence at the current timestep and predict the sequence in the next timestep. The special units in the hidden layer with controlled elements of state "memorization / forgetting" is a feature of RNN structures. This property is useful in predicting time series, especially in learning their temporal dependences. Currently, various configurations of recurrent neural networks are proposed based on different units in hidden layers, e.g. BRNN (a bidirectional NN), LSTM (a long short-term memory), GRU (a gated recurrent unit) [2-4]. The successful implementation of RNN and LSTM as a component of forecasting methods for time series sets leads to increasing interest in them [5-6], [21-23]. In [4], ten LSTM architectures were considered, and their main disadvantages were mentioned: a higher memory demand and computational complexity than a simple RNN due to the many memory cells.

Compared to LSTM, the GRU is characterized by some simplifications leading to study of a smaller number of weights. While the LSTM is commonly used for time series forecasting, the RNN based on GRU is more novel Therefore, the comparative study of their effectiveness attracts more and more attention. In the comparative study [7] for speech recognition RNNs based either on LSTM or GRU units have been demonstrated to perform well. Several similarities and differences between GRU networks and LSTM networks are outlined in [8]. This study found that both models performed better than the other only in certain tasks, so, it is hard to tell, which one is the better choice for a given problem. In the paper [9] Laptev et al. studied RNNs in event forecasting and found that neural networks might be a better choice in comparison with classical time series methods when the number, the length and the correlation of the time series are high. Che et al. in the report [10] described a GRU-based model with a decay mechanism to capture informative missingness in multivariate time series. A methodology (namely, DeepAR) for producing probabilistic forecasts, based on training an auto-regressive recurrent network model on a large number of related time series was proposed in [16]. Instead of forecasting raw time series the authors focused on probabilistic forecasting, i.e. estimating the probability distribution of a time series' future given its past. It was showed through empirical evaluation on several real-world forecasting data sets accuracy improvements of around 15% compared to state-of-the-art methods, in particular, ETS model [17] with automatic model selection. Experimental results provided in forecasting of Financial Data show that GRU can be more suitable to use due to their training time performance and can yield similar accuracy to LSTM [18]. However, these results also show no significant performance difference with simpler traditional forecasting models. The comparison of LSTM with ARIMA in forecasting of financial and economical time series was provided in [19]. As follows from the work, the LSTM showed significantly better prediction accuracy than ARIMA according to the RMSE criterion. In the work [20] it was carried out an empirical study in time series forecasting using both LSTM and GRU networks. To compare LSTM and GRU the real time series set referred to bike sharing was used. In this work the prediction technique was proposed, included bootstrap samples of sequences and preliminary presentation of time series properties such as cyclicality, seasonality, the beginning of the month, holidays or working days and some others. Although the author used many performance indicators, the conclusion was that two networks produce very similar forecasts.

The amount of related works demonstrates the interest in analysis of RNNs with GRU and LSTM units from different point view. At the same time the RNNs based on GRU and LSTM are not sufficiently studied according to accuracy in forecasting of time series in particular for time series having different behavior. An analysis of the literature allowed us to conclude that there is no rigorous empirical assessment of the prediction accuracy of LSTM and GRU for time series that have either a trend, or seasonality, or a combination of both. Filling such gap is useful for obtaining a meaningful and accurate predictions of the time series on the basis of a reasonable choice of the type of RNN configuration. The paper focuses on evaluating and comparing the accuracy of LSTM and GRU units of RNN in forecasting of time series based on different models. For this study, we developed a methodology, created software for

generating the artificial time series set and for prediction these time series. The experimental results were evaluated on MAPE criterion for two different types of RNN configurations in forecasting of time series having either a trend, or seasonality, or a combination of both.

The structure of the paper includes six Sections. The second section briefly discusses the basics of the functioning of the LSTM and GRU units of RNN. The models of time series used for comparison of two configurations of RNN are presented in the Section 3. The Section 4 incudes the methodology for evaluating and comparing of LSTM and GRU units of RNN in time series forecasting. The results of comparison of the predicting accuracy of two configurations of LSTM and GRU on modeled time series are described at the Section 5. The discussions and conclusions are depicted at the Section 6.

## 2 Concepts of LSTM and GRU units of RNN

A recurrent neural network (RNN) is an extension of a conventional feedforward neural network which is one of the best tools for solving image recognition problems. In time series analysis the RNN is focused on learning the function of the input $x_t$ and previous output $h_{t-1}$. The simplified form of this function can be expressed as follows:

$$h_t = \mathrm{f}(x_t, h_{t-1}) \tag{1}$$

This RNN is primarily suited for one-step-ahead forecasting when a single output immediately follows the corresponding sequence of inputs at the next time moment. For multi-step-ahead forecasting an iterative (recursive) technique should be applied. However, it is difficult to train RNNs to learn long-term dependencies because the gradients tend to vanish [2].

To solve the problem the recurrent unit, called Long Short Term Memory [2] and later its modification, called Gated Recurrent Unit [11] were proposed.

### 2.1 Long Short-Term Memory

Below the concepts of LSTM is described as used in [12]. In comparison to major RNN each recurrent unit in LSTM architecture includes the variable for store in the memory cell the previous state $c_{t-1}$. Thus, the output at a given time is calculated not only on the basis of the previous output and input, but also using the previous state:

$$h_t = \mathrm{f}(x_t, h_{t-1}, c_{t-1}) \tag{2}$$

LSTM unit includes the forget (f), the input (i), the output (o) gates respectively and new memory context (g) which is used to update the previous state $c_{t-1}$.

The expressions are used to calculate the output h and the new state c of LSTM [12]:

$$f = \sigma\left(W_f h_{t-1} + U_f x_t\right) \qquad g = \tanh\left(W_g h_{t-1} + U_g x_t\right)$$
$$i = \sigma\left(W_i h_{t-1} + U_i x_t\right) \qquad c_t = \left(c_{t-1} \otimes f\right) + \left(g \otimes i\right)$$
$$o = \sigma\left(W_o h_{t-1} + U_o x_t\right) \qquad h_t = \tanh\left(c_t\right) \otimes o$$

So, at each time LSTM unit updates the two variables, the output $h_t$ and the state $c_t$, using controlled by gates composition of functions on previous values of these variables and learned weights. Including the additional variable and function to modify it makes the architecture of RNN with LSTM more complex for understanding in comparison to simple RNN. The various architectures of LSTM and comparison of their advantages are described in [4], where was noted that for identical size of hidden layers, a typical LSTM has about four times more parameters than a simple RNN.

The main advantages of LSTM are it capability of modeling long-term sequential dependencies and more robustness to vanishing gradients than a RNN.

### 2.2 Gated Recurrent Unit

To make LSTM less complex the gated recurrent units (GRUs) were proposed in the work [11]. Like to major unit of RNN the GRU computes the output as a function of the input $x_t$ and previous output $h_{t-1}$, but using gates that modulate the flow of information inside the unit. The latter makes GRU similar to LSTM. In contrast to LSTM, the GRU does not process the previous state and computes a linear sum:

$$z = \sigma\left(W_z h_{t-1} + U_z x_t\right) \qquad c = \tanh\left(W_c\left(h_{t-1} \otimes r\right) + U_c x_t\right)$$
$$r = \sigma\left(W_r h_{t-1} + U_r x_t\right) \qquad h_t = (z \otimes c) + \left((1 - z) \otimes h_{t-1}\right)$$

In addition to the advantages associated with LSTM compared to RNN, GRU requires less memory and time to learn than LSTM.

## 3 Models of Time Series

In this Section the models of time series used for comparison of GRU and LSTM units in configurations of RNN are presented. We considered three major time series models as they correspond to real time series in many domains, such as medicine, economics, manufacturing, ecology and etc.

Let $X = \{x_t \in \mathbb{R}, t = 1,2, \ldots, n\}$ be a given numerical time series. We consider an additive model in according to time series model [13] containing a trend $f_t$ and fluctuation components, the latter includes a regular $s_t$ and an irregular $\varepsilon_t$ fluctuations:

$$x_t = \alpha \cdot f_t + \beta \cdot s_t + \gamma \cdot \varepsilon_t \tag{3}$$

where $x_t$ is a value of a given time series $X$, $s_t$ is a seasonal part of a time series and $\varepsilon_t$ is a random part of a time series, $n$ is a length of a given time series, $t$ = 1, 2, …, n. In the model (3) three coefficients were added: $\alpha \in \{-1,0,1\}$, $\beta \in \{0,1\}$, $\gamma \in \{0,1\}$,

their values determine the presence of the corresponding component in the model (3). If $\alpha$ = 0 the time series does not involve a trend, when $\alpha$ = -1 the time series has a decreasing trend, if $\alpha$ = 1 the time series has an increasing trend; if $\beta$ = 0 the time series model (3) does not include the seasonal fluctuations; $\varepsilon_t$ is regarded as a sequence of serially uncorrelated random values with zero mean and finite variance. Time series with different behavior could be generated changing the values of the coefficients in the model (3). In the study three types of time series behavior was considered:

$$x_t = \alpha \cdot f_t + \gamma \cdot \varepsilon_t \quad (4)$$

$$x_t = \beta \cdot s_t + \gamma \cdot \varepsilon_t \quad (5)$$

$$x_t = \alpha \cdot f_t + \beta \cdot s_t + \gamma \cdot \varepsilon_t \quad (6)$$

In our comparison of RNN with LSTM and RNN with GRU the following representation of components of time series model (3) was used:

$$f_t = k \cdot t + q, \quad (7)$$

$$s_t = y \cdot \sin(w \cdot t) \quad (8)$$

Here $k, q, y, w$ designate the parameters, which values are some samples from predefined numeric intervals: $k \in [1{,}10], q \in [1,\ 40], w \in [0.5,\ 1.5], t = 1{,}2, \ldots, n$. The coefficient $y$ is calculated using the expression (9) to make seasonal fluctuations more expressive in the time series with trend:

$$y = (|q - k \cdot n|)/10 \quad (9)$$

## 4 Methodology and Software for Study of GRU and LSTM in time series forecasting

We focused on assessing accuracy of two of RNN configurations, having in the hidden layer different units (see Section 2) on the dataset of artificial time series. The artificial time series with different behavior were produced on the basis of varying the coefficients of the models (4-8). Although there are a lot of descriptions of methods of time series prediction, where RNN with LSTM was successfully applied to real time series, the research question, which type of units in RNN should be used to which type of TS behavior to obtain the best accuracy, is still open. In this Section the methodology and software tool for study and evaluating the learning ability of GRU and LSTM units of RNN in time series forecasting are proposed. The comparison was made in equal conditions. Equal conditions mean that the parameters of RNN, such as number of RNN layers, quantity of units in these layers, probability of dropout between layers, activate function, optimizer are the same, except the type of units (GRU or LSTM), at the same dataset of time series and on the equal prediction interval.

Dataset of artificial time series includes 30 time series was created using time series models (3-8): five TS with $\alpha = -1,\ \beta = 1, \gamma = 1$, five time series with $\alpha = 1,\ \beta = 1, \gamma = 1$, ten time series with $\alpha = 0,\ \beta = 1, \gamma = 1$, five time series with $\alpha = -1,\ \beta = 0, \gamma = 1$, five time series with $\alpha = 1,\ \beta = 0\gamma = 1$. Thus, we formed noisy time series with trend, time series with seasonal component and time series having both trend and seasonal components. All time series in the dataset include random components which are assumed to be a white noise. Each of time series has equal length, that is n=200, the splitting on training and testing parts was established as 90:10.

Below the proposed methodology for estimating and comparing of accuracy of time series forecasting by two RNN configurations is described. The methodology is based on our previous study of time series models [14] and includes two separately performed parts: the formation of a set of time series and the prediction of the obtained time series by training RNN based on LSTMs and RNN based on GRUs.

Part 1. Formation of dataset of time series

Step 1. Chose the time series behaviour (trend or season components or their combination) with noise by setting the coefficients of time series model (3).

Step 2. Form and store the artificial time series at dataset *D*.

Part 2. Computation of prediction accuracy for two RNN configurations with LSTMs and with GRUs.

Step 3. Set the parameters of studied RNN.

Step 4. For each type of time series ($TS_{type}$) from the dataset *D*

Step 4.1. Divide time series into training (180 time points: $t = 1,2, \dots ,180$) and testing parts (20 time points: $t = 181,182, \dots ,200$). Define prediction horizon as 20 time points: $t = 201,202, \dots ,220$.

Step 4.2. Use RNN with LSTM units to learn temporal dependencies in train part of a time series and produce its forecasts for test part and for horizon. Evaluate the accuracy by criterion MAPE (Mean of Absolute Percent Error):

$$MAPE(U) = \frac{100\%}{m}\sum_{t=1}^{m}\frac{|\hat{x}_t - x_t|}{|x_t|} \quad (10)$$

where $\hat{x}_t$ is the predicted values of a test part of time series; $m$ is the number of time points in test part, that is 20; $x_t$ is the real values of the unknown for RNN test part of the time series, $U$ denotes the type of the unit. In this Step $U = LSTM$.

Step 4.3. Use RNN with GRU units to learn temporal dependencies in train part of a time series and produce its forecasts for test part and for horizon. Evaluate the accuracy by criterion MAPE, in this Step $U = GRU$.

Step 4.4. Output the prediction accuracy of two RNN configurations in the following form:

$$< Id, TS_{type}, MAPE\ (GRU), MAPE\ (LSTM) > \quad (11)$$

To conduct the comparative study of two configurations of RNN (with GRU or LSTM) the software tool was developed. In the software the Python 2.7 was used with the Keras library for building a neural network under the Tensorflow library for

automatic differentiation [15]. The Pandas library for data processing and the library scikit-learn mainly for time series normalization/denormalization were implemented as well. The developed software includes the set of modules grouped by features: TS generator, TS pre-processing, training RNN for time series prediction, time series post-processing, calculating the accuracy of prediction, output the results of prediction in graph and text forms. TS generator using the parameters of time series models (3-9) creates time series with different behaviour and put them in TS dataset in csv format. Pre-processing module normalizes a time series and divides it into training and testing parts. The scikit-learn and Pandas libraries are used to quickly read a file with a time series from TS dataset, for filtering, normalizing and dividing the data in order to obtain a time series that will be submitted to the network input. Two consequence configurations of RNN are created and trained to produce two forecasts for each time series at the RNN module using Keras library. The first configuration of RNN includes the LSTM units and the second one contains GRU units. In the study, the three-layer architecture of RNN was used. The input layer contained twenty and the hidden layer ten GRU or LSTM units in dependence of chosen type of RNN configurations. At the output layer a single-layer perceptron generates the forecasts of a time series. The loss function MSE, activation function ReLU and optimizer Adam were chosen for two configurations of RNN. A maximum of 200 epochs has been established for trainig each neural network. The RNN module produces forecasts of normalized time series in two forms: for the test part of the time series (test prediction) and for new time (real prediction), that is, the prediction horizon. The post-processing module performs the denormalization of time series forecasts using scikit-learn library. The MAPE criterion corresponding to expressions (10) is calculated in the MAPE module to obtain the comparative accuracy of the two RNN configurations for all time series from the TS dataset. The package NumPy for Python was used for fast and convenient work with the time series and for output the results in according to structure (11).

## 5 Comparison of LSTM and GRU Prediction Accuracy

The task of study is to compare the prediction accuracy of the LSTM and GRU units of two RNN configurations. The dataset of 30 artificial time series of equal length with three types of behavior was previously created. Using the developed technique and the software described in Section 4, the time series forecasting was conducted and prediction accuracy of two RNN configurations were evaluated by MAPE criterion (10). The first study of the prediction accuracy of LSTM and GRU blocks of two RNN configurations was performed without taking into account the type of time series behavior. A comparative study of two configurations of RNN regardless of the behavior of the time series showed that the GRU units provide less prediction accuracy on average (MAPE (GRU) = 5.729%) compared with LSTM (MAPE (LSTM) = 5.072%) for the 30 time series used in this study. Based on these averaged errors, a small superiority of LSTM in time series forecasting can be established, and at the same time they require more time for training. Experiments have shown that RNN

with LSTM units predicted time series with a smaller error, not much more often than with GRU, as a percentage of this was 60% of the analyzed time series. These results suggest that a comparison of the accuracy of RNN models for forecasting without taking into account time series behavior may not fully characterize the capabilities of LSTM and GRU.

Therefore, then we evaluated prediction accuracy of RNN (with GRU or LSTM) in respect to three types of time series behavior:

1. SR. Time series with seasonality and random component corresponding to model (5).
2. TR. Time series with trend and random component corresponding to model (4).
3. TSR. Time series with trend, seasonality and random component corresponding to model (6).

The experimental data presented in Tables 1 - 4 form the basis for obtaining new knowledge of the accuracy of the two studied RNN configurations (with GRU or LSTM) in time series prediction.

The structure of the Tables 1, 2, 3 corresponds to the form (11) and shows the minimal MAPEs for the test parts of each predicted time series from dataset for GRU and for LSTM units respectively. The minimal MAPEs were calculated for each time series in 10 experiments by varying the RNN parameter «lookback» from 1 to 10. The first column (Id) in Tables 1, 2, and 3 indicates the number of predicted time series, the second and third columns contain the values of the MAPE criterion (10) for the GRU and LSTM units, respectively.

The data from the Table 1 show prediction accuracy for time series with seasonality and random components (SR) and demonstrate, that for 5 time series LSTM overperforms GRU and for others 5 time series GRU overperforms LSTM. The experimental data in Table 1 demonstrate, that the prediction accuracy of both the studied RNN configurations for SR time series is not high and approximately is the same:
$5.141\% \leq \text{MAPE(LSTM)} \leq 9.698\%$ and $5.399\% \leq \text{MAPE(GRU)} \leq 10.365\%$.

**Table 1.** Prediction accuracy of LSTMs and GRUs for time series with seasonality and random components (SR)

| Id | MAPE(GRU)% | MAPE (LSTM)% | Unit with min MAPE |
|---|---|---|---|
| 1 | 5.654 | 5.978 | GRU |
| 2 | 5.697 | 5.612 | LSTM |
| 3 | 10.365 | 9.698 | LSTM |
| 4 | 7.965 | 8.092 | GRU |
| 5 | 6.531 | 6.933 | GRU |
| 6 | 6.655 | 7.428 | GRU |
| 7 | 6.579 | 5.453 | LSTM |
| 8 | 7.367 | 7.301 | LSTM |
| 9 | 6.205 | 5.141 | LSTM |
| 10 | 5.399 | 5.978 | GRU |

For time series with trend and random (TR) component the MAPEs in Table 2 show that the forecasts of LSTM units were more accurate for 8 time series, while the GRU units showed more accurate forecasts for two time series only. At the same time, it should be noted that the prediction accuracy of LSTM and GRU presented in Table 2 is significantly better for time series with trend and random components compared to time series having seasonality (see Table 1). Prediction errors for TR time series were in the interval $0.78\% \leq$ MAPE (LSTM) $\leq 3.326\%$ and $1.724\% \leq$ MAPE (GRU) $\leq 4.173\%$.

**Table 2.** Prediction accuracy of L

STMs and GRUs for time series with trend and random components (TR)

| Id | MAPE(GRU)% | MAPE (LSTM)% | Unit with min MAPE |
|---|---|---|---|
| 1 | 2.378 | 1.019 | LSTM |
| 2 | 3.185 | 1.123 | LSTM |
| 3 | 1.724 | 0.78 | LSTM |
| 4 | 3.564 | 2.252 | LSTM |
| 5 | 2.514 | 2.42 | LSTM |
| 6 | 4.173 | 3.173 | LSTM |
| 7 | 3.528 | 1.98 | LSTM |
| 8 | 2.591 | 1.647 | LSTM |
| 9 | 2.384 | 2.398 | GRU |
| 10 | 3.223 | 3.326 | GRU |

Both RNN configurations, regardless of the type of unit for time series with trend, seasonality and random component (TSR), have more variations of prediction errors than time series with no trend (see table 1) or without seasonality (see table 2) according to the experimental data presented in Table 3. Comparative analysis of the data in Table 3 shows that for 5 time series, LSTM is superior to GRU, and for the remaining 5 time series, GRU is superior to LSTM. Prediction errors for TSR time series varied in the range of $2.251\% \leq$ MAPE(LSTM) $\leq 17.354\%$ and $2.258\% \leq$ MAPE(GRU) $\leq 22.01\%$.

**Table 3.** Prediction accuracy of LSTMs and GRUs for time series with trend, seasonality and random components (TSR)

| Id | MAPE(GRU)% | MAPE (LSTM)% | Unit with min MAPE |
|---|---|---|---|
| 1 | 2.258 | 2.251 | LSTM |
| 2 | 3.104 | 3.172 | GRU |
| 3 | 4.719 | 4.726 | GRU |
| 4 | 4.216 | 4.572 | GRU |
| 5 | 3.964 | 3.564 | LSTM |
| 6 | 22.010 | 11.437 | LSTM |
| 7 | 12.859 | 17.354 | GRU |
| 8 | 7.969 | 6.601 | LSTM |
| 9 | 8.364 | 5.066 | LSTM |
| 10 | 4.717 | 5.698 | GRU |

At the same time, the comparison of the studied RNN configurations according to the experimental data from Table 1,2,3 allow us to conclude that their prediction accuracy for TSR time series was the worst compared to TR and SR time series.The results of a comparative study of the prediction accuracy of RNN configurations with different units on the dataset of all time series grouped by their behavior, are given in Table 4. The average, minimal and maximal values of the criterion MAPE are presented for GRU and LSTM configurations of RNN in respect to time series with different behavior. In addition, the amount of the best predictions performed by GRU or LSTM on the test parts of time series is depicted at last columns.

**Table 4.** Average, minimal and maximal of prediction accuracy of LSTMs and GRUs for different time series types

| TS type | MAPE (GRU) % | | | MAPE (LSTM)% | | | Amount | |
|---|---|---|---|---|---|---|---|---|
| | Average | Min | Max | Average | Min | Max | GRU | LSTM |
| SR | 6.842 | 5.399 | 10.365 | 6.761 | 5.141 | 9.698 | 5 | 5 |
| TR | 2.926 | 1.724 | 4.173 | 2.012 | 0.78 | 3.326 | 2 | 8 |
| TSR | 7.418 | 2.258 | 22.01 | 6.444 | 2.251 | 17.354 | 5 | 5 |

When forecasting stationary time series with seasonality (SR), it is difficult to determine the winner, because the results are about the same, and for 50% of the time series the LSTM was better than the GRU. Regarding trend time series prediction (TR), RNN with LSTM should be preferred, since it demonstrated much better average prediction accuracy and superiority in accuracy for most time series. At the same time, when the noise level in such time series was small, the results were completely opposite: the GRU forecasts were in most cases more accurate. For time series containing noisy time dependencies in the form of both trend and seasonality, on average, more accurate predictions were obtained by LSTM. However, the LSTM showed higher accuracy for only 50% of the predicted series. In general, both the studied RNN configurations predicted more accurately the time series constructed using the model (4) containing the trend and random components.

## 6 Discussions and conclusions

In the paper the study of two RNN configurations in time series forecasting was derived. The study focused on mining knowledge about differences in the prediction accuracy of RNNs on test parts of the time series. The proposed framework of the study is based on analysis of prediction accuracy of the LSTM or GRU units in RNN configurations in dependence of time series behavior and regardless of their behavior. The dataset of 30 artificial time series of equal length with three types of behavior was created previously. Using the proposed framework, the software was developed to conduct the study and to obtain experimental data in the form of prediction accuracy for further analysis LSTM and GRU. The software includes modules for time series generating with different behavior, for RNN creating and training to predict time series, for time series forecasting and computing prediction accuracy of LSTM and GRU by MAPE criterion. The study allows to conclude following findings.

In this study for the 30 artificial time series the prediction accuracy of RNNs in average was not very high: MAPE (GRU) = 5.729% compared with LSTM (MAPE (LSTM) = 5.072%. LSTM shows a slightly better accuracy (within 1 percent) on average and in most cases compared to GRU in predicting time series depending on the behavior of time series and regardless of their behavior. At the same time RNN with LSTM has demonstrated a better prediction accuracy for 60% of time series (for 18 out of 30 time series). Comparison of forecasting accuracy for time series in respect to different behaviors showed that forecasting accuracy for time series with a trend and random component was best on average for both LSTM and GRU (MAPE(GRU)= 2.9% and MAPE(LSTM)= 2.0%), while LSTM for 8 out of 10 time series showed a smaller error. As for LSTM and GRU, this accuracy exceeded by more than 2 times the prediction accuracy of time series containing a seasonal component (one or with a trend), moreover, the largest error variation was observed for time series having both a trend and seasonality and a random component. The obtained knowledge is useful in producing new techniques of meaningful predictions of time series and expand our understanding of the applicability of recurrent neural networks with GRU or LSTM in time series forecasting.

The future work will be focused on developing the technique and software for combining the positive abilities of two RNN configurations in time series forecasting.

## References


1. Haykin, S. Neural networks: a comprehensive foundation. Prentice Hall PTR. (1994).
2. Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation, 9(8):1735–1780. (1997).
3. Zachary C. Lipton, John Berkowitz, Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning. https://arxiv.org/abs/1506.00019. (2015), last accessed 2019/08/10.
4. Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent Advances in Recurrent Neural Networks. https://arxiv.org/pdf/1801.01078.pdf. (2018), last accessed 2019/08/10.
5. Slawek Smyl, Jai Ranganathan, Andrea Pasqua.: M4 Forecasting Competition: Introducing a New Hybrid ES-RNN Model, https://eng.uber.com/m4-forecasting-competition/, last accessed 2019/05/10.
6. CIF 2016, http://irafm.osu.cz/cif/main.php?c=Static&page=results, last accessed 2019/08/10.
7. Graves, A., Abdel-rahman Mohamed, Geoffrey Hinton. Speech Recognition with Deep Recurrent Neural Networks, https://arxiv.org/abs/1303.5778, (2013), last accessed 2019/05/10.
8. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, https://arxiv.org/pdf/1412.3555.pdf, (2014), last accessed 2019/05/10.
9. Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl.: Time-series extreme event forecasting with neural networks at uber. In International Conference on Machine Learning, number 34, pages 1–5, (2017).

10. Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu.: Recurrent neural networks for multivariate time series with missing values, (2018), https://arxiv.org/pdf/1606.01865.pdf, last accessed 2019/05/10.
11. K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio.: On the properties of neural machine translation: Encoder-decoder approaches, https://arxiv.org/pdf/1409.1259.pdf, (2014) last accessed 2019/05/10.
12. Bruno Goncalves. RNNs for Timeseries Analysis, (2018). https://conferences.oreilly.com/strata/strata-ny-2018/public/schedule/detail/69080, www.bgoncalves.com, last accessed 2019/05/10.
13. Makridakis, S., Wheelwright, S.C., and Hyndman, R.J. Forecasting methods and applications. John Wiley & Sons, Inc. (1998).
14. Afanasieva T., Yarushkina N., Gyskov G. The Study of Basic Fuzzy Time series Forecasing models. in World Scientific Proceedings on Computer Engineering and Information Science, V0l.10. UNCERTAINTY MODELLING IN KNOWLEDGE ENGINEERING AND DECISION MAKING. Proceedings of the 12th International FLINS CONFERENCE ENSAIT (FLINS 2016), France, pp.295-300. (2016) https://doi.org/10.1142/9789813146976_0049.
15. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras: (2016). https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/, last accessed 2019/05/10.
16. David Salinas, Valentin Flunkert, Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, (2019). https://arxiv.org/pdf/1704.04110.pdf, last accessed 2019/08/10.
17. Hyndman, R., Koehler, A. B., Ord, J. K. and Snyder, R .D. Forecasting with Exponential Smoothing: The State Space Approach. Springer Series in Statistics. Springer, (2008). ISBN 9783540719182.
18. Douglas Garcia Torres, Hongliang Qiu. Applying Recurrent Neural Networks for Multivariate Time Series Forecasting of Volatile Financial Data, (2018), https://www.researchgate.net/publication/322027012_Applying_Recurrent_Neural_Networks_for_Multivariate_Time_Series_Forecasting_of_Volatile_Financial_Data, last accessed 2019/08/10.
19. Sima Siami Namini , Neda Tavakoli, Akbar Siami Namin. A Comparison of ARIMA and LSTM in Forecasting Time Series, in Proc of 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), (2018), DOI: 10.1109/ICMLA.2018.00227.
20. G´abor Petneh´azi. Recurrent Neural Networks for Time Series Forecasting, (2019), https://arxiv.org/pdf/1901.00069, last accessed 2019/08/10.
21. Kasun Bandara, Christoph Bergmeir, Slawek Smyl. Forecasting Across Time Series Databases using Recurrent Neural Networks on Groups of Similar Series: A Clustering Approach, (2018), https://arxiv.org/pdf/1710.03222.pdf, last accessed 2019/08/10.
22. Himadri Sikhar Khargharia ; Siddhartha Shakya ; Russell Ainslie ; Sara AlShizawi ; Gilbert Owusu/ Predicting Demand in IoT Enabled Service Stations, 2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), (2019), DOI: DOI:10.1109/COGSIMA.2019.8724239, last accessed 2019/08/10.
Cristian Rodriguez Rivero, ets.Time Series Forecasting using Recurrent Neural Networks modified by Bayesian Inference in the Learning Process, IEEE Colombian Conference on Applications of Computational Intelligence ColCACI 2019, At: Barranquilla, Colombia (2019),DOI: 10.1109/ColCACI.2019.8781984, last accessed 2019/08/10.

# Optimal and Efficient Model Selection Criteria for Parametric Spectral Estimation

**Abass Ishola TAIWO***

*Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria*

(***Corresponding author's e-mail:** taiwo.abass@oouagoiwoye.edu.ng)

## Abstract

Traditional Parametric spectral estimation methods have been widely used to obtain spectral estimate, resolution and variance distribution in signals or time series data across several fields. But a problem of how to choose an optimal and efficient model order selection was encountered. In this research work, modified forms of Final Prediction Error, Akaike Information Criteria, Bayesian Information Criteria and Minimum Description Length which involved the replacement of variance error with sample autocorrelation function that has capabilities of detecting non-randomness in time series data were proposed, to choose an optimal and efficient model order selection. The results of the modified and traditional information criteria were used to choose AR(11) and AR(7) as the optimal model. The spectral estimate and resolution of AR(11) indicated the modified methods outperformed the traditional methods for analyzing Heart-beat readings. Conclusively, the proposed method outperformed the traditional Information criteria for selecting optimal and efficient model order for Heart-beat readings with lower values of parameters.

**Keywords**: Parametric Spectral estimation, Information Criteria, Heart-beat readings, Optimal model order, Autocorrelation function

### Introduction

Parametric spectral estimation methods like autoregressive, moving average and autoregressive moving average spectral estimation methods have been widely used to obtain spectral estimate, resolution and variance distribution in signals or time series data across several fields over the years, Andrew and Barry (2017), Taiwo (2017), Abdul-Majeeb *et. al (*2014), Andrea *et. al.* 2013, Batorova I. (2012) and many more. These methods were prominent due to their abilities to handle short segments of data, computationally efficient, better frequency resolution and smoother power spectra. However, the main challenge encountered when using these methods has to do with the selection of an optimal and efficient model to be estimated, Thanagasundran and Schlindwein (2006), Beheshti (2006). Intuitively, a model order which is too small will not represent the properties of the signal, whereas a model order which is too high will also represent noise and inaccuracies thus will not be a reliable representation of the true signal, Palaniappan (2010).

There are many criteria used to determine model order selection in parametric spectral estimation and the commonly used were Final Prediction Error (FPE) (Akaike, 1970), Akaike Information Criterion (AIC) (Akaike, 1974), Akaike's Bayesian Information criteria (Schwartz, 1978) and Minimum Description Length (MDL) (Kashyap, 1980 and Rissasen, 1983). Over the year, several other criteria and algorithms have been developed and these included Criterion Autoregressive Transfer (CAT), Parzen (1974) and Residual Variance (RV), Box and

Jenkins (1970), Hannan and Quinn (HQ) (Hannan and Quinn (1979), Generic algorithm (Palaniappan, 2006), Particle swarm optimization (PSO) (Bijaya, 2010) and many more. But these model selection methods are not completely outlined here since the vast amount of techniques for solving the problem of selecting model order may have not been mentioned. The main goal of this research article is to improve the method of selecting the optimal and efficient model order in parametric spectral estimation. This will be done by modifying some existing traditional information criteria (Final prediction error, Akaike, Schwartz Bayesian and Minimum Description length information criteria) for selecting optimal order. The modification will involve the replacement of variance of error with sample autocorrelation function which has the capabilities of detecting non-randomness, help in identifying an appropriate model for non-randomness time series data and instead of estimating of the error variance that required minimization of the log-likelihood function of the given model.

**Materials and Methods**

**Final Prediction Error (FPE)**

This is the first criteria proposed by Akaike (1970) and it is based on minimizing one step ahead predictor error. It is denoted by

$$FPE_{(k)} = \left(1 + \frac{k}{N}\right)\hat{\sigma}_k^2 \qquad (1)$$

where $\hat{\sigma}_k^2$ is the unbiased estimate of $\sigma_k^2$ after fitting the $k^{th}$ order model.

**Akaike Information Criteria**

This is the most well-known and most used criteria and it was proposed by Akaike (1974). It is denoted by

$$AIC(k) = Nln\hat{\sigma}_k^2 + 2k \qquad (2)$$

where $N$ is the number of observation and $\hat{\sigma}_k^2$ is the maximum likelihood estimation of the residual after fitting the $k^{th}$ order model.

**Schwartz's SBC Criteria**

Similar to Akaike's, Bayesian criteria, Schwartz (1978) suggest the Bayesian criteria defined as

$$SBIC(k) = NIn\hat{\sigma}_k^2 + MInN \qquad (3)$$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate of $\sigma_k^2$, $M$ is the number of parameters in the model and $N$ is the number of observations.

**Minimum Description Length Criteria**

Based on the work of Kashyap (1980), it was proved that the Akaike information criteria is inconsistent and it tends to overestimate the order. Rissasen (1983) proposed modified Akaike information criteria by replacing the term $2k$ by a term which increases more rapidly. This criterion was named minimum description length (MDL) and is of the form

$$MDL = Nln|\hat{\sigma}_k^2| + kln(N) \qquad (4)$$

**Modified Final Prediction Error (FPE)**

This is the first criteria proposed by Akaike (1970) and it is modified by replacing the variance of error by the sample autocorrelation function. It is denoted by

$$FPE_{(k)} = \frac{|N+k|}{N-k}|\hat{\rho}_k| \quad (5)$$

where $\hat{\rho}_k = \hat{\gamma}_k(0) + \sum_{i=1}^{k} \hat{a}_j \hat{\gamma}_{kk}(1)$ is the power of the prediction error that decreases with $k$ while the term $\frac{N+K}{N-k}$ increases with $k$.

**Modified Akaike Information Criteria**

This is the most well-known criteria and it was proposed by Akaike (1974) and was modified by replacing the variance of error by the sample autocorrelation function. It is denoted by

$$AIC(k) = Nln\,|\hat{\rho}_k| + 2k \quad (6)$$

This criterion is more general than the final prediction error and it can be applied to determine the order of the moving average part of an autoregressive moving average model.

**2.7 Schwartz's SBC Criteria**

Similar to Akaike's, Bayesian criteria, Schwartz (1978) suggest the Bayesian criteria model that is modified by replacing variance of error with sample autocorrelation function and it is defined as

$$SBC(k) = kIn|\hat{\rho}_k| + NInk \quad (7)$$

where $\hat{\rho}_k$ is an estimate of $\rho_k$, $k$ is the number of parameters in the model and $N$ is the number of observations.

**2.8 Minimum Description Length Criteria**

Based on the work of Kashyap (1980), it was proved that the Akaike information criteria is inconsistent and it tends to overestimate the order. Rissasen (1983) proposed modified Akaike information criteria by replacing the term $2k$ by a term which increases more rapidly. This criterion is named minimum description length (MDL) and is of the form

$$MDL = Nln|\hat{\rho}_k| + kln(N) \quad (8)$$

**2.9 Spectral Density function of $AR(1)$ Process**

Given an autoregressive model $AR(1)$,

$$y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \varepsilon_t \quad (9)$$

Using the backshift operator gives

$\Phi(B)y_t = \varepsilon_t$, where $\Phi(B) = (1 - \emptyset_1 B)$

since $y_t = \Phi^{-1}(B)\varepsilon_t$, then

$$\gamma_k = E(y_t y_{t-1}) = \sigma_\varepsilon^2 \left(\frac{1}{(1-\phi B)(1-\phi \mathrm{B}^{-1})}\right) \quad (10)$$

Equation (3.87) is written in spectral representation as

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{|\Phi(e^{-i\omega})|^2}\right] \quad (11)$$

Since $|1-\Phi(e^{-i\omega})|^2 = 1-\phi_1 e^{i\omega}-\phi_1 e^{-i\omega}+\phi_1^2$

$$= 1-\phi_1(e^{i\omega}+e^{-i\omega})+\phi_1^2$$

and using a standard trigonometric form given as

$$cos\omega = \frac{e^{i\omega}+e^{-i\omega}}{2}$$

$$|1-\Phi(e^{-i\omega})|^2 = 1-2\phi_1 cos\omega+\phi_1^2$$

Then, equation (11) becomes

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{1-2\phi_1 cos\omega+\phi_1^2}\right] \quad (12)$$

**2.10 Spectral Density function of $AR(2)$ Process**

The Yule-Walker process is given by

$$y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \varepsilon_t \quad (13)$$

Using the backshift operator to obtain

$$\Phi(B)y_t = \varepsilon_t,$$

where $\Phi(B) = (1-\emptyset_1 B-\phi_2 B^2)$

Since $y_t = \Phi^{-1}(B)\varepsilon_t$

$$\gamma_k = \sigma_\varepsilon^2\left(\frac{1}{(1-\emptyset_1 B-\phi_2 B^2)(1-\emptyset_1 B^{-1}-\phi_2 B^{-2})}\right) \quad (14)$$

Equation (14) is written in spectral representation as

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{|1-\phi_1(e^{-i\omega})-\phi_2(e^{-2i\omega})|^2}\right]$$

$$= \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{|1-\phi_1(\cos\omega-i\sin\omega)-\phi_2(cos2\omega-isin2\omega)|^2}\right]$$

$$= \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{(1-\phi_1 e^{-i\omega}-\phi_2 e^{-2i\omega})(1-\phi_1 e^{i\omega}-\phi_2 e^{2i\omega})}\right]$$

$$= \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{1-\phi_1 e^{i\omega}-\phi_2 e^{2i\omega}-\phi_1 e^{-i\omega}+\phi_1^2+\phi_1\phi_2 e^{i\omega}-\phi_2 e^{-2i\omega}+\phi_2\phi_1 e^{-i\omega}+\phi_2^2}\right]$$

Using a standard trigonometric form given as

$$cos\omega = \frac{e^{-i\omega}+e^{i\omega}}{2}$$

Equation (14) can be expressed as

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left[\frac{1}{1+\phi_1^2+\phi_2^2-2\phi_1 cos\omega-2\phi_2 cos2\omega+2\phi_1\phi_2 cos\omega}\right] \quad (15)$$

**Discussion and Conclusion**

The time plot in figure 1 showed the Heartbeat reading observed at equal space and time of 0.05 second. In order to identify the optimal model for the Heartbeat reading, autocorrelation and partial autocorrelation functions were used to identify tentative models AR(3), AR(5), AR(7), AR(9) and AR(11) for 50 heart-beat readings observed at equal space and time of 0.05 second. To validate these models, modified and traditional information criteria were obtained. From table 1 and figure 2, the lowest values of all the modified information criteria occurred at AR(11) while the lowest values for the traditional information criteria occurred at AR(7) in table 2. The performance of both information criteria was determined based on spectral estimate and resolution of Autoregressive spectral estimation using a modified covariance autoregressive estimator. Based on figure 3, the spectral estimate of AR(7), AR(9) and AR(11) indicated a relatively fast oscillation. This was explained by the two sinusoidal components in all autoregressive order but AR(11) is better since it has a dominant peak, better spectral estimate and resolution when compared to AR(7) and AR(9). AR(11) depicts the general oscillation better than AR(7) and AR(9). Thus the modified information criteria that is, modified final prediction error, Akaike, Schwartz Bayesian and Minimum Description length information criteria give an optimal and effective model selection as against the most frequently used traditional information criteria.

**Conclusion**

This research article was used to propose an improved method of model selection in parametric spectral estimation. This was done by modifying the traditional information criteria and from the results obtained, AR(7) and AR(11) were the optimal models selected using modified and traditional information criteria. The spectral estimate and resolution of AR(11) were better than AR(7). Conclusively, the modified information criteria outperformed the traditional information criteria when analyzing the spectral estimate of 50 heartbeat readings, with lower parameters than the traditional methods.

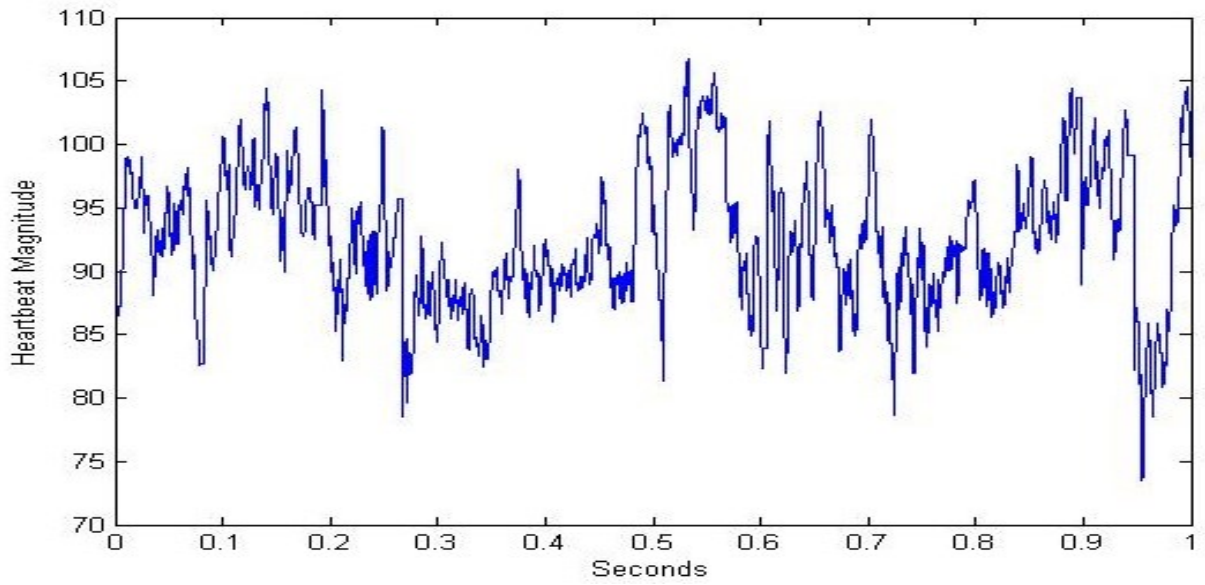



**Figure 1. Time Plot of Heartbeat Readings observed at Equal space and Time of 0.05 Second**

**Table 1. Modified Information Criterion for 50 Heartbeat Readings**

| NUMBER | FPE | AIC | BIC | MDL |
|---|---|---|---|---|
| 1 | 45.74546 | -1.45679 | -0.06914 | 0.45524 |
| 2 | 19.42644 | -4.61658 | 34.3127O | -0.79254 |
| 3 | 16.46170 | -8.77227 | 54.04428 | -3.03620 |
| 4 | 14.04178 | -12.50380 | 67.67442 | -4.85569 |
| 5 | 11.93420 | -16.31390 | 77.84051 | -6.75377 |
| 6 | 9.63539 | -22.57740 | 85.43868 | -11.10530 |
| 7 | 7.23339 | -32.35490 | 90.80581 | -18.97080 |
| 8 | 5.00028 | -46.11840 | 94.03314 | -30.82220 |
| 9 | 3.18856 | -63.76340 | 95.14381 | -46.55520 |
| 10 | 1.79256 | -87.53610 | 93.62203 | -68.41590 |
| 11 | 0.59493 | -137.46600 | 84.81228 | -116.43400 |
| 12 | 0.50280 | -140.43800 | 84.78019 | -117.49400 |
| 13 | 1.38346 | -84.13700 | 99.61185 | -59.28070 |
| 14 | 1.95248 | -60.92240 | 107.05460 | -34.15410 |
| 15 | 2.22812 | -47.98480 | 112.00710 | -19.30440 |
| 16 | 2.28275 | -40.02700 | 115.58080 | -9.43460 |
| 17 | 8.28499 | -35.61560 | 117.99140 | -3.11123 |
| 18 | 1.89294 | -34.28410 | 119.21630 | 0.13236 |
| 19 | 1.56426 | -35.18040 | 119.41340 | 1.14806 |
| 20 | 1.24642 | -36.87910 | 119.03500 | 1.36134 |

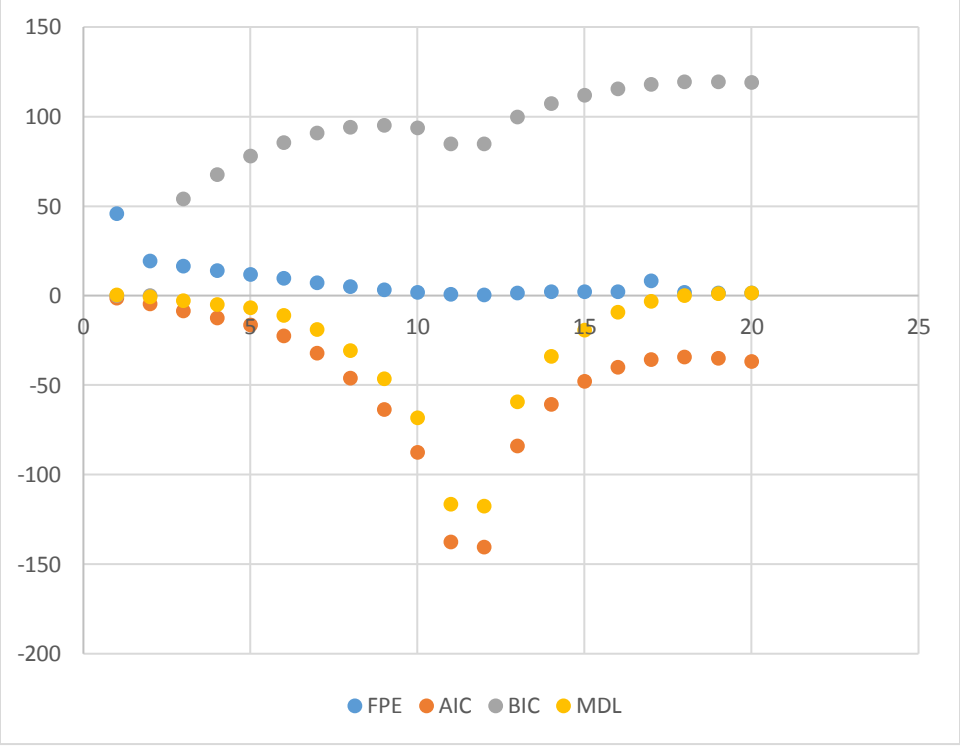


**Figure 2. Graphical representation of Modified Information Criteria**

**Table 2. Traditional Information Criterion values for 50 Heartbeat Readings**

| NUMBER | FPE | AIC | BIC | MDL |
|---|---|---|---|---|
| 3 | 1.84691 | 1.99214 | 1.99214 | 1.85246 |
| 5 | 1.43891 | 1.56250 | 1.56250 | 1.43891 |
| 7 | 1.18234 | 1.47838 | 1.47838 | 1.18234 |
| 9 | 1.34821 | 1.29929 | 1.71723 | 1.34821 |
| 11 | 1.41245 | 1.86863 | 1.86863 | 1.41245 |

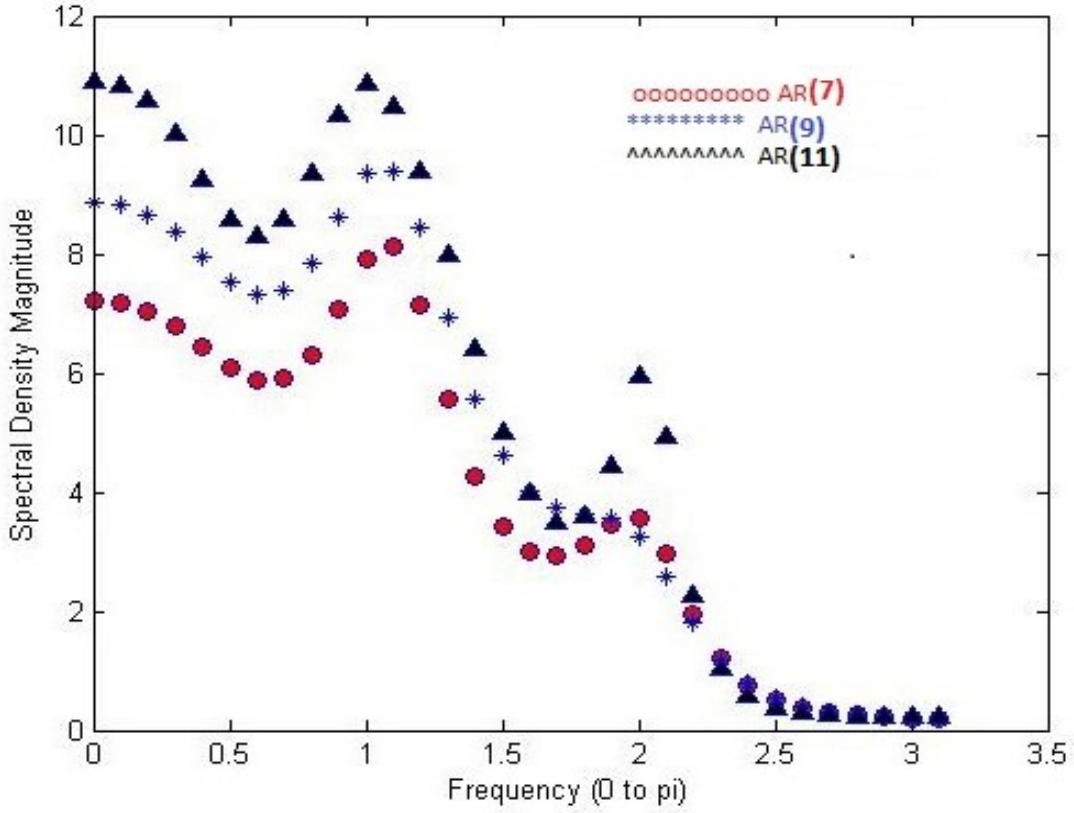



**Figure 3. Spectral Estimates for AR(7), AR(9) and AR(11)**

**References**

Abdul-Majeeb H. A., Mohammed Q. A. and Wadhah S. I. (2014). Estimating the Spectral Power density function of Non-Gaussian Second order Autoregressive model. Journal of Asian Scientific Research, 4(9), 513-521.

Akaike H. (1970). Statistical Prediction Identification. Annals of the Institute of Statistical Mathematics, 22, 200 - 217.

Akaike H. (1974). A new look at Statistical Model Identification. IEEE Trans. Automa.Contr., AC-19, 716-723.

Andrea F., Arianna P., Laura S. and Massimilland M. (2013). Fourier analysis for Demand Forecasting in a Fashion Company. International Journal of Engineering Business Management, 5(30), 1-10.

Andrew J. G. and Barry G. Q. (2017). Parametric Spectral Discrimination. Journal of Time series Analysis, 38(3), 1 - 27.

Batorova I. (2012). Spectral Techniques for Economic Time series. Cornenius University, Department of Applied Mathematics and Statistics, Bratislava, Slovakia.

Beheshti S. (2006). A New Approach to Order Selection and Parametric Spectrum Estimation Acoustics, Speech and Signal Processing, ICASSP Proceedings, Toulouse, France.

Bijaya G. (2010). Spectral Estimation of Electroencephalogram signal using ARMAX model and Particle Swarm Optimization. Texas: Lamar University.

Box G. and Jenkins G. M. (1970). Time series Analysis, Forecasting and Control. Oakland CA: Holden-Day.

Hannan, E, J. and B.G. Quinn, B. G (1979) "The determination of the order of an autoregression," J. of the Royal Statistical Soc., B41(2), 190-195.

Kashyap R. (1980). Inconsistency of the AIC rule for Estimating the order of Autoregressive model. IEEE Transaction on Acoustic Control, 25(10), 996 - 998.

Palanippan R. (2006). Towards Optimal model order Selection for Autoregressive Spectral Analysis of Mental Tasks using Genetic Algorithm, International Journal of Computer Science and Network Security, 6(1A), 153 – 161.

Parzen E. (1974). Some recent advances in Time series modeling. IEEE Trans. Automat. Contr., AC-19, 723-730.

Rissasen J. (1983). A Universal prior for the Integers and Estimation by Minimum Description Length. Annal of Statistics, 11, 417 - 431.

Schwartz G. (1978). Estimating the Dimension of a model. Annal of Statistics, 6, 461 - 464.

Taiwo, A. I. (2017). Spectral and Fourier Parameter Estimation of Periodic Autocorrelated Time Series Data. (Ph.D. Thesis), Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Wear K. A., Wagner R. F., Garra B. S., Grossman L. W., Isana M. F., Myers K. J. and Rajan S. S. (1990). Application of Parametric Spectral estimation to Medical Ultrasound and Magnetic Resonance Spectroscopy. Proc. SPIE Medical Imaging. 1231, 187-181. Newport Beach CA: SPIE.

# Forecasting Stock Market Data using a Hybrid EMD-HW Method

Ahmad M. Awajan[1,a)] and S. AL Wadi[2,b)]

[1]*Department of Mathematics, Faculty of Science, Al-Hussein Bin Talal University, Jordan*
[2]*Department of risk management, The university of Jordan., Amman, Jordan*

[a)]Corresponding author: ahmad.m.awajan@ahu.edu.jo
[b)]sadam_alwadi@yahoo.co.uk

**Abstract.** In this study, a hybrid method based on Empirical Mode Decomposition and Holt-Winter (EMD-HW) is used to forecast stock market data. First, the data are decomposed by EMD method into Intrinsic Mode Functions (IMFs) and residual components. Then, all components are forecasted by HW technique. Finally, forecasting values are sum together to get the forecasting value of stock market data. Empirical results showed that the EMD-HW outperform individual forecasting models. The daily Spain stock market time series data are applied to show the forecasting performance of the EMD-HW. The strength of this EMD-HW lies in its ability to forecast non-stationary and non-linear time series without a need to use any transformation method. Moreover, EMD-HW has a relatively high accuracy comparing with eight existing forecasting methods based on the five forecast error measures.

**Keywords**: Holt-Winter, Empirical Mode Decomposition, Forecasting

**Scope of Abstract**: Statistics (Time series)

## INTRODUCTION

In financial time series analysis, one of the primary issues is modeling and forecasting financial time serirs data specifically stock market index. Usually, the transformation of a financial time series, rather than its original scale, is taken for describing its dynamics. Proper transformation is necessary to convert original non-stationary processes to stationary processes and subsequently to utilize mathematical and statistical properties for stationary processes.

The hybrid models combine strengths of few traditional models to get a better forecasting accuracy. Recently, several hybrid models were applied EMD in the literature for time series forecasting. That by using EMD to decompose the non-stationary and non-linear time series data into Intrinsic Mode Functions (IMFs) and residual components. And then use forecasting model to forecast each component. Then all these forecasted values were aggregated to produce the final forecasted value of the original time series. Such as in [1] used a hybrid EMD with Support Vector Machine (SVM) model to forecasting the river flow data. A hybrid EMD-MA model was developed by coupling an Moving Average (MA) model with the EMD technique in [2], the EMD-MA model was applied to forecast stock market data. A hybrid EMD-LSSVR (least squares support vector regression) forecasting model has been applied on foreign exchange rate in [3].

With regard to all those literature reviews, this study attempts to apply a hybrid of EMD-HW to forecast the daily stock market data of four countries. In order to assess the performance of forecasting, and the EMD-HW method is compared with eight forecasting methods. Experimental results show that the EMD-HW method is superior to existing methods in terms of five accuracy forecasting measure. Section 2 introduces methods are used in methodology in this paper which are EMD and Holt Winter. Section 3 presented the proposed methodology with flowchart explain the steps. Section 4 analyzes the daily stock market time series data of Spain with a discussion the result showing the capability of EMD-HW. Finally, in Section 5 some concluding remarks are addressed.

## METHODS

In this section, the various steps for the implementation of the EMD-HW forecasting method are described in detail. Which are Empirical Mode Decomposition and Holt-Winter.

### Empirical mode decomposition [EMD]

EMD has been described by [4]. The main idea of EMD is decompose of nonlinear and non-stationary time series data $x(t)$ -with keeping the time domain of the signal- into $n$ of simple time series that known as intrinsic mode functions (IMF). Later, the original signal can be constructed back as the following:

$$x(t) = \sum_{i=1}^{n} IMF_i + r(t). \tag{1}$$

where $r(t)$ represents the residue of the original time series data decomposition and IMF$i$ represent the $i^{th}$ intrinsic mode function (IMF) series. In order to estimate these IMFs, the following steps should be initiated and the process is called the sifting process of time series $x(t)$ are shown below:

1. Start the first step by taking the original time series $x(t)$ for sifting process and assuming the iteration index value is $i = 1$.

2. Then, evaluate all of local extreme value of the time series $x(t)$.

3. After that, form the local maxima (local upper) envelope function $u(t)$ by connecting all local maxima values using a cubic spline line. In a similar way, form the local minimum ( local lower ) envelope function $l(t)$, and then form the mean function $m(t)$ by using this following;

$$m(t) = \frac{l(t) + u(t)}{2} \tag{2}$$

4. Next defined as a new function $h(t)$ using the mean envelope $m(t)$ and the signal $x(t)$ on this formula

$$h(t) = x(t) - m(t) \tag{3}$$

Check the function $h(t)$ is an IMF, according to IMF condition, which is

$$|Num[extreme] - Num[cross - zero]| < 1 \tag{4}$$

where Num.extreme represents the number of local extreme points (all local maxima and all local minima), also Num[cross-zero] represent the number of cross-zero points.
If the function $h(t)$ has satisfied IMF conditions continue to step 5. If not, go back to step 2 and renew the value of $x(t)$ such that became $h(t)$, repeat again steps 2 until 4.

5. In step 5, firstly saves the result of the IMF obtain from the last step. Secondly, renew the iteration index value such that became $i = i + 1$. Thirdly attain the residue function $r(t)$ using the IMF and the signal $x(t)$ on the formula.

$$IMF_i(t) = h(t) \Rightarrow r_{i+1}(t) = x(t) - IMF_{i+1}(t) \tag{5}$$

6. Finally made a decision whether the residue function r(t) that acquire from step 5 is a monotonic or constant function. Then, save the residue and all the IMFs obtained. If the residue is not monotonic or constant function, return to step 2.

The steps 1 to 6 which was discussed above allow the sifting process (EMD algorithm) to separate time-altering signal features. FIGURE 1 summarizes all the steps.

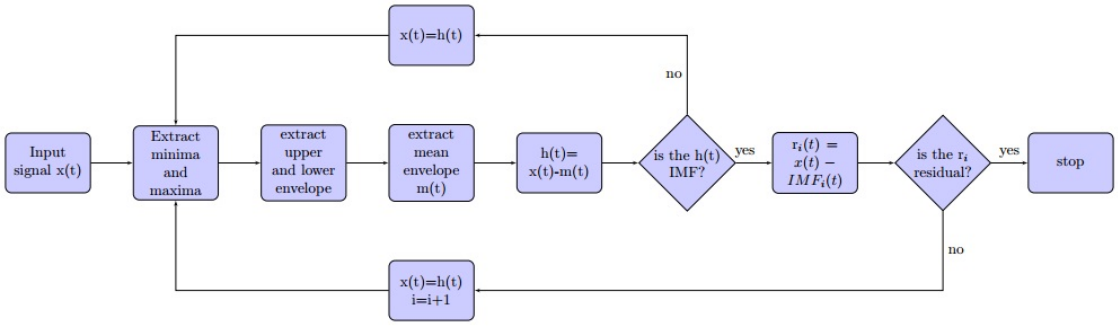



**FIGURE 1.** Flowchart of empirical mode decomposition estimation process

## Holt-Winter (HW)

More than fifty five years ago, the basic formula of the Holt-Winter model or Triple Exponential Smoothing have been presented by [5] and [6]. The Holt-Winters forecasting procedure is a variant of exponential smoothing which is simple. This method is simple, does not need high data-storage requirements, and is easily automated, is particularly suitable for producing short-term forecasts for sales or demand time-series data. And in this method the recent observations have effect more robustly than old observations in forecasting value. Two Holt-Winter models are described in this study are the Multiplicative Model and the Additive Model. Mathematically, the additive Holt-Winters forecasting function is defined by the following:

$$\hat{y}_{t+h/t} = a_t + h * b_t + s_{t-p+1+(h-1)mod(p)}, \tag{6}$$

where $a_t, b_t$ and $s_t$ are given by

$$a_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{7}$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{8}$$
$$s_t = \gamma(y_t - a_t) + (1-\gamma)s_{t-p} \tag{9}$$

And the multiplicative Holt-Winters forecasting function is defined by the following :

$$\hat{y}_{t+h/t} = (a_t + h * b_t) * s_{t-p+1+(h-1)mod(p)}, \tag{10}$$

where $a_t, b_t$ and $s_t$ are given by

$$a_t = \alpha(y_t/s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{11}$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{12}$$
$$s_t = \gamma(y_t/a_t) + (1-\gamma)s_{t-p} \tag{13}$$

such that, $a_t$ represent the level of series at time $t$, $b_t$ represent the slope (growth) at time $t$, $s_t$ represent the seasonal component of the series at time $t$, and $p$ represent the number of seasons in a year. The constants $\alpha$, $\beta$ and $\gamma$ are smoothing parameters in the [0,1]-interval, $h$ is the forecast horizon. This method uses the maximum likelihood function to estimate the starting parameters and then it may estimate iteratively all the parameters to forecasting future values of time series. The data in $x$ are required to be non-zero for a multiplicative model, but it makes most sense if they are all positive.

## METHODOLOGY AND DATA

This section contains three parts. The first part is about the data that is used to implement the proposed methodology. While the second part presents the EMD-HW methodology with detailed description.

**TABLE 1.** Basic statistics

| Country | Mean | Median | Min | Max | Standard Deviation | Skewness | Kurtosis | # of observations |
|---|---|---|---|---|---|---|---|---|
| Spain | 9537 | 9933.1 | 5956.3 | 11866.4 | 1297.48 | -0.55 | -0.67 | 1515 |

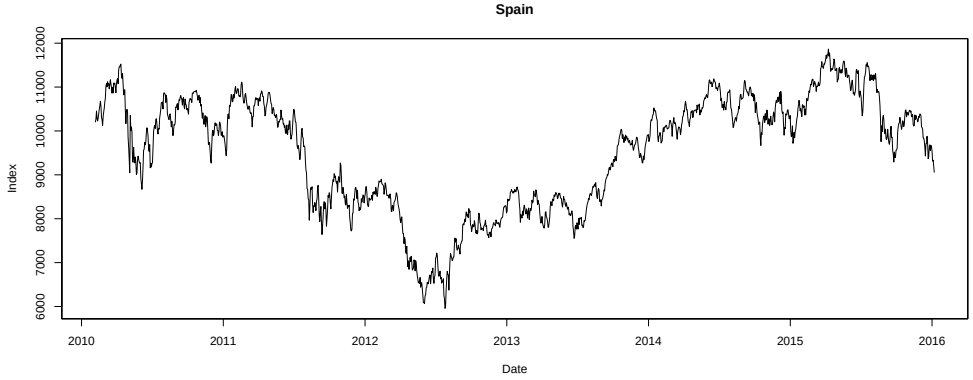



**FIGURE 2.** Time Series Plots

## Data

In this study, a nonlinear and non-stationary time series is used, which is Spain daily stock market data. Table 1 presents the Basic statistics and the number of observations. The data are extracted from the Yahoo finance website. Figure 2 shows the time series plot of these countries. The daily closing prices are used as a general measure of the stock market over the past six years. The whole data set - for each country - covers the period from 9 February 2010 to 7 January 2016. The data set is divided into two parts. The first part (n observations) is used to determine the specifications of the models and parameters. The second part, on the other hand, (h observations) is reserved for out-of-sample evaluation and comparison of performances among various forecasting models.

## Methodology

The EMD-HW method consists of three stages. Firstly, the use of empirical mode decomposition (EMD) on the daily stock market time series data. In this stage, Intrinsic Mode Functions (IMFs) and residue are obtained. Secondly, the Holt-Winter (HW) is applied on each IMFs and residue to forecast h days ahead. Finally, in the last stage all the forecasted results for IMFs and residue are added up. The EMD-HW methodology is presented as a flowchart in Figure 3.

## RESULT AND DISCUSSION

In this study, Spain stock market are used to present the forecasting accuracy of the EMD-HW method. Eight forecasting models are used in order to validate the forecasting performance of EMD-HW. Table 2 shows five error measurements with their formula. These measurements will be utilized to evaluate the forecasting accuracy for each method. Where $\widehat{y_i}$ is the forecast value of the variable $y$ at time period $i$ from knowledge of the actual series values.

Table 3 presents the RMSE, MAE, MAPE, MASE, and TheilU of EMD-HW and eight forecasting methods for forecasting at h = 1 into 10 with its average for the Spain stock market. This indicates that the forecast accuracy for EMD-HW is better than the eight traditional forecasting methods.

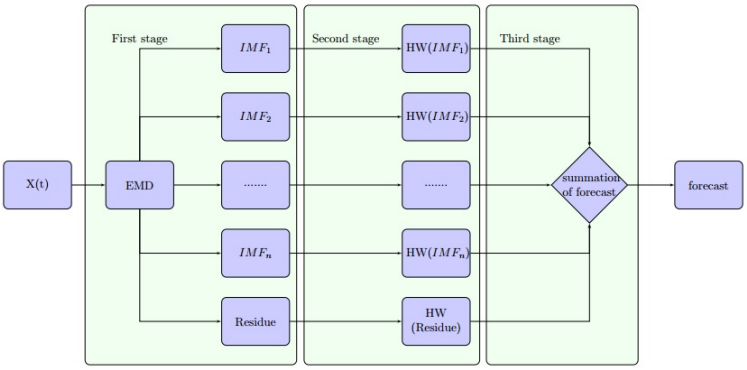



**FIGURE 3.** Flowchart of a hybrid Empirical Mode Decomposition with Holt-Winter

**TABLE 2.** Error measures are used in study

| Name of measure error | Formula of measure error |
|---|---|
| Mean Absolute Scaled Error | $MASE = \frac{1}{h}\sum_{i=1}^{h}\left(\frac{\lvert y_i-\widehat{y_i}\rvert}{\frac{1}{h-1}\sum_{i=2}^{h}\lvert y_i-y_{i-1}\rvert}\right)$ |
| Root Mean Squared Error | $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}$ |
| Mean Absolute Error | $MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\widehat{y_i}\rvert$ |
| Mean Absolute Percentage Error | $MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i-\widehat{y_i}\rvert}{y_i}.100\%$ |
| Theil's U-statistic | $TheilU = \frac{\sqrt{\sum_{i=1}^{h-1}\left(\frac{\widehat{y}_{i+1}-y_i}{y_i}-\frac{y_{i+1}-y_i}{y_i}\right)^2}}{\sqrt{\sum_{i=1}^{h-1}\left(\frac{y_{i+1}-y_i}{y_i}\right)^2}}$ |

## CONCLUSION

Time series forecasting still remains as one of the most difficult areas due to the non-stationary and non-linearity of financial time series data. In this study, we have applied a hybrid method, a composite of empirical mode decomposition with Holt-Winter model (EMD-HW) for non-stationary and nonlinear time series forecasting. EMD-HW was tested on daily stock market time series data of Spain based on the comparison of five of forecast accuracy measurements. It was found that EMD-HW is able to outperform eight forecasting methods. Thus, this paper has strengthened the idea that EMD-HW forecasting method is suitable for non-stationary and nonlinear time series.

## ACKNOWLEDGMENTS

The authors would like to thank the Al-Hussein Bin Talal University for the financial support. [7]

## REFERENCES

[1] S. Ismail and A. Shabri, "Combination model of empirical mode decomposition and svm for river flow forecasting," in *AIP Conference Proceedings*, Vol. 1830 (AIP Publishing, 2017) p. 080005.
[2] A. M. Awajan, M. T. Ismail, and S. Al Wadi, Italian Journal of Pure and Applied Mathematics **38**, 1–20.
[3] C.-S. Lin, S.-H. Chiu, and T.-Y. Lin, Economic Modelling **29**, 2583–2590 (2012).
[4] N. E. Huang, *Hilbert-Huang transform and its applications*, Vol. 16 (World Scientific, 2014).
[5] C. C. Holt, International Journal of Forecasting **20**, 5–10 (1957, reprinted 2004).
[6] P. R. Winters, Management Science **6**, 324–342 (1960).
[7] A. M. Awajan and M. T. Ismail, "A hybrid approach emd-hw for short-term forecasting of daily stock market time series data," in *AIP Conference Proceedings*, Vol. 1870 (AIP Publishing, 2017) p. 060006.

**TABLE 3.** ERROR

| MASE | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=7$ | $h=8$ | $h=9$ | $h=10$ | *Average* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HW | – | 1.38 | 0.83 | 2.97 | 2.48 | 2.35 | 1.77 | 2.12 | 1.64 | 1.78 | 1.92 |
| MA | – | 1.65 | 1.03 | 3.22 | 2.60 | 2.62 | 2 | 2.27 | 1.85 | 1.84 | 2.12 |
| ARIMA | – | 1.42 | 0.59 | 2.41 | 2.38 | 2.30 | 1.40 | 2.02 | 3.33 | 3.64 | 2.16 |
| RW | – | 1.50 | 0.95 | 3.20 | 2.66 | 2.57 | 1.99 | 2.43 | 1.89 | 1.90 | 2.12 |
| EXP | – | 1.50 | 0.95 | 3.20 | 2.66 | 2.57 | 1.99 | 2.43 | 1.89 | 1.90 | 2.12 |
| STS | – | 1.49 | 0.94 | 3.19 | 2.65 | 2.56 | 1.98 | 2.42 | 1.87 | 1.89 | 2.11 |
| EARIMA | – | 8.27 | 5.69 | 7.50 | 6.40 | 6.70 | 6.93 | 6.11 | 6.10 | 6.17 | 6.65 |
| **EMD.HW** | – | 2.16 | 1.76 | 2.40 | 1.38 | 1.19 | 1.20 | 1.10 | 1.12 | 2.73 | 1.67 |
| MAE | | | | | | | | | | | |
| HW | 125.39 | 189.98 | 115.06 | 295.17 | 327.77 | 294.36 | 193.58 | 233.85 | 185.59 | 186.94 | 214.77 |
| MA | 133.22 | 227.74 | 142.58 | 319.73 | 344.02 | 328.83 | 218.01 | 251.01 | 209.41 | 193.62 | 236.82 |
| ARIMA | 17.43 | 196.05 | 80.73 | 239.07 | 314.21 | 288.58 | 153.22 | 222.77 | 376.20 | 381.94 | 227.02 |
| RW | 138.1 | 206.85 | 130.57 | 317.93 | 352.04 | 321.87 | 217.21 | 268.64 | 212.93 | 199.66 | 236.58 |
| EXP | 138.11 | 206.85 | 130.57 | 317.93 | 352.04 | 321.86 | 217.22 | 268.63 | 212.93 | 199.66 | 236.58 |
| STS | 137.44 | 205.99 | 129.78 | 316.83 | 350.92 | 320.63 | 215.85 | 267.08 | 211.73 | 198.98 | 235.52 |
| EARIMA | 941.71 | 1,141 | 784.50 | 744.27 | 846.10 | 840.18 | 756.87 | 674.49 | 688.45 | 648.41 | 806.70 |
| **EMD.HW** | 305.09 | 298.30 | 242.18 | 238.26 | 182.60 | 148.75 | 131.57 | 121.01 | 126.82 | 286.26 | 208.08 |
| RMSE | | | | | | | | | | | |
| HW | 125.39 | 200.29 | 138.63 | 311.79 | 360.34 | 346.97 | 230.25 | 300.99 | 246.92 | 212.34 | 247.39 |
| MA | 133.22 | 238.48 | 174.53 | 338.80 | 379.82 | 381.77 | 265.52 | 323.54 | 285.88 | 211 | 273.26 |
| ARIMA | 17.43 | 203.84 | 93.89 | 254.55 | 345.04 | 325.04 | 168.54 | 292.31 | 462.22 | 392.33 | 255.52 |
| RW | 138.1 | 218.07 | 161.62 | 336.32 | 386.95 | 377.24 | 266.13 | 339.16 | 289.82 | 215.68 | 272.91 |
| EXP | 138.11 | 218.07 | 161.63 | 336.33 | 386.95 | 377.23 | 266.14 | 339.16 | 289.80 | 215.68 | 272.91 |
| STS | 137.44 | 217.16 | 160.44 | 335.14 | 385.72 | 375.87 | 264.43 | 337.48 | 287.89 | 215.02 | 271.66 |
| EARIMA | 941.71 | 1,145 | 838.63 | 791.41 | 859 | 849.81 | 790.99 | 728.31 | 726.48 | 685.97 | 835.79 |
| **EMD.HW** | 305.09 | 307.01 | 267.53 | 264.02 | 208.79 | 165.75 | 148.13 | 149.51 | 149.50 | 301 | 226.63 |
| TheilU | | | | | | | | | | | |
| HW | – | 1.83 | 1.22 | 2.97 | 2.68 | 2.71 | 1.89 | 2.51 | 2.07 | 1.70 | 2.18 |
| MA | – | 2.16 | 1.55 | 3.26 | 2.82 | 2.98 | 2.19 | 2.71 | 2.39 | 1.75 | 2.42 |
| ARIMA | – | 1.82 | 0.37 | 2.42 | 2.56 | 2.53 | 1.30 | 2.44 | 3.86 | 3.30 | 2.29 |
| RW | – | 2 | 1.44 | 3.22 | 2.87 | 2.94 | 2.21 | 2.83 | 2.43 | 1.76 | 2.41 |
| EXP | – | 2 | 1.44 | 3.22 | 2.87 | 2.94 | 2.21 | 2.83 | 2.43 | 1.76 | 2.41 |
| STS | – | 1.99 | 1.43 | 3.21 | 2.87 | 2.93 | 2.19 | 2.82 | 2.41 | 1.75 | 2.40 |
| EARIMA | – | 8.93 | 7.16 | 7.72 | 6.05 | 6.21 | 6.52 | 6.01 | 5.85 | 5.80 | 6.69 |
| **EMD.HW** | – | 2.69 | 2.32 | 2.59 | 1.54 | 1.20 | 1.08 | 1.21 | 1.14 | 2.52 | 1.81 |
| MAPE | | | | | | | | | | | |
| HW | 1.37 | 2.04 | 1.24 | 3.10 | 3.41 | 3.06 | 2.04 | 2.43 | 1.94 | 2 | 2.26 |
| MA | 1.45 | 2.43 | 1.53 | 3.35 | 3.57 | 3.40 | 2.28 | 2.60 | 2.17 | 2.05 | 2.48 |
| ARIMA | 0.19 | 2.10 | 0.88 | 2.53 | 3.27 | 3 | 1.63 | 2.32 | 3.82 | 4.21 | 2.39 |
| RW | 1.5 | 2.22 | 1.40 | 3.33 | 3.65 | 3.33 | 2.27 | 2.77 | 2.21 | 2.12 | 2.48 |
| EXP | 1.5 | 2.22 | 1.40 | 3.33 | 3.65 | 3.33 | 2.27 | 2.77 | 2.21 | 2.12 | 2.48 |
| STS | 1.49 | 2.21 | 1.39 | 3.32 | 3.64 | 3.32 | 2.26 | 2.76 | 2.20 | 2.11 | 2.47 |
| EARIMA | 9.42 | 11.1 | 7.81 | 7.42 | 8.35 | 8.25 | 7.45 | 6.66 | 6.79 | 6.41 | 7.97 |
| **EMD.HW** | 3.26 | 3.16 | 2.57 | 2.52 | 1.94 | 1.58 | 1.40 | 1.29 | 1.35 | 3.12 | 2.22 |

# The Non-Stationary INARMA(1,1) Model with Generalized Innovation.

Sunecher Yuvraj

University of Technology Mauritius
La Tour Koenig, Mauritius
{ysunecher@umail.utm.ac.mu }

**Abstract.** This paper considers modelling of a non-stationary integer-valued autoregressive moving average of order 1 (INARMA(1,1)) model by assuming that the innovation follows a Poisson and negative binomial distribution. Two simulation experiments are also conducted to assess the performance of conditional maximum likelihood (CML) and generalized quasi-likelihood (GQL) estimation methods.

**Keywords:** INARMA(1,1) model; CML; GQL; Poisson; Negative Binomial

## 1 Introduction

Time series of counts are usually modeled using parameter-driven (PD) or observation-driven (OD) approaches as illustrated in Cox [4]. In the PD approaches, the serial-correlations are induced by some correlated randomly distributed e ects [14, 17] while in the OD approach, the time correlations are induced through the previous-lagged observations [7, 9, 10]. Under both approaches, time-varying covariates may be specified [1] but the estimation of the regression parameters in the PD models are quite laborious since the specification of the likelihood function in such models is cumbersome whilst the construction of the likelihood function is more parsimonious in OD models [3, 5].
The simplest family of stationary OD integer-valued autoregressive of first order (INAR(1)) models were developed initially by McKenzie [8] and were further extended by McKenzie [10] and Al-Osh and Alzaid [11]. As for the modeling of non-stationary INAR(1) series, some authors such as Brannas [1] and Mamode Khan et al. [7] have re-formulated the classical OD process of McKenzie [8] by considering the non-stationarity through time-dependent covariates in the link predictor specification of the model. On the other hand, some researchers have also considered developing the integer-valued moving average of first order (INMA(1)) models [2, 6, 10, 12]. However, very few researchers have developed integer-valued models having both the AR and MA components (INARMA) [9, 11]. Hence, this paper proposes an INARMA(1,1) model with non-stationary generalized innovation. As for the estimation of the model parameters, the conditional maximum likelihood (CML) and the generalized quasi-likelihood (GQL)

are used [7].
The organization of the paper is as follows: In Section 2, the INARMA(1,1) model is constructed. In Section 3, the CML and GQL approaches for estimating the unknown parameters of the model are described. Section 4 compares the performance of the CML and GQL under two distributions of the innovation term. The conclusion is provided in the last Section.

## 2 Model Construction

The INARMA(1,1) model $\{[Y_t],\ t \in \mathbb{Z}\}$ is specified as

$$Y_t = \rho_1 \circ Y_{t-1} + \rho_2 \circ R_{t-1} + R_t \tag{1}$$

where $Y_t$ is the random count observation at the $t^{th}$ time point with corresponding innovation terms $R_t$. In the above model (1), $'\circ'$ indicates the binomial thinning operator such that $\rho \circ Y_{t-1} = \sum_{j=1}^{Y_{t-1}} b_j(\rho) = z_t$ where $b_j(\rho)$ is the Bernoulli random variable with probability $\rho$, where $\rho \in [0,1]$ or simply $\{\rho \circ Y_{t-1}|Y_{t-1}\} \sim \text{Binomial}(Y_{t-1}, \rho)$ [15]. As for the innovation terms, $R_t$ and $R_{t+h}$ are independent for $h \neq 0$, with $E(R_t) = \lambda_t$ and $\text{Var}(R_t) = \nu\lambda_t$, where $\lambda_t = \exp(\boldsymbol{x_t'\beta})$, with $\boldsymbol{x}_t = [x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tp}]'$ with $p$ explanatory e ects and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_p]'$.

Based on the above conditions, it is proved under non-stationary moment condition that

$$\begin{aligned} E(Y_t) &= E(\rho_1 \circ Y_{t-1} + \rho_2 \circ R_{t-1} + R_t) \\ \mu_t = E(Y_t) &= \rho_1\mu_{t-1} + \rho_2\lambda_{t-1} + \lambda_t, \end{aligned} \tag{2}$$

$$\begin{aligned} \text{Var}(Y_t) &= \text{Var}(\rho_1 \circ Y_{t-1} + \rho_2 \circ R_{t-1} + R_t) \\ &= \text{Var}(\rho_1 \circ Y_{t-1}) + \text{Var}(\rho_2 \circ R_{t-1}) + 2\text{Cov}(\rho_1 \circ Y_{t-1}, \rho_2 \circ R_{t-1}) + \text{Var}(R_t) \\ &= \rho_1(1-\rho_1)\mu_{t-1} + \rho_1^2\text{Var}(Y_{t-1}) + \left[\rho_2(1-\rho_2) + (\rho_2^2 + 2\rho_1\rho_2)\nu\right]\lambda_{t-1} + \nu\lambda_t, \end{aligned} \tag{3}$$

$$\begin{aligned} \text{Cov}(Y_t, Y_{t+h}) &= \text{Cov}[Y_t, (\rho_1 \circ Y_{t+h-1} + \rho_2 \circ R_{t+h-1} + R_{t+h})] \\ &= \rho_1\text{Cov}[Y_t, Y_{t+h-1}] \\ &= \rho_1^h\text{Var}(Y_t) + \rho_1^{h-1}\rho_2\nu\lambda_t. \end{aligned} \tag{4}$$

## 3 Estimation of Parameters

### 3.1 CML

In this subsection, the CML estimation method is constructed based on thinning and convolution properties as in Pedeli and Karlis [13]:

The conditional density of the proposed INARMA(1,1) model, can be constructed as the convolution of

$$f_1(k) = \sum_{j_1=0}^{k} \binom{y_{t-1}}{j_1} \binom{r_{t-1}}{k-j_1} \rho_1^{j_1}(1-\rho_1)^{y_{t-1}-j_1} \rho_2^{k-j_1}(1-\rho_2)^{r_{t-1}-k+j_1} \quad (5)$$

and the marginal distribution of the innovation term $f_2(r_t = y_t - k)$.

Then, the conditional density is written as

$$f((y_t|(y_{t-1}, r_{t-1}), \boldsymbol{\theta}) = \sum_{k=0}^{g_1} f_1(k) f_2(r_t), \quad (6)$$

where $\boldsymbol{\theta} = [\beta_1, \beta_2, ..., \beta_p, \rho_1, \rho_2, \nu]$ is the vector of unknown parameters, $g_1 = \min(y_t, y_{t-1})$.

Using some initial value of $\boldsymbol{y_0}$, the conditional log-likelihood function

$$\log[L(\boldsymbol{\theta}|\boldsymbol{y})] = \log\left[\sum_{t=1}^{T} f((y_t)|(y_{t-1}, r_{t-1}), \boldsymbol{\theta})\right] \quad (7)$$

ia maximized to obtain the maximum likelihood estimates of $\boldsymbol{\theta}$.

### 3.2 GQL

This section formulates two GQL equations [16] to estimate the unknown parameters of the INARMA(1,1) model:
The GQL-I estimates the regression and thinning e ects and is given by

$$\boldsymbol{D}^{'} \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = 0 \quad (8)$$

where $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_T]^{'}$ and $E(\boldsymbol{Y}) = \boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_T]^{'}$.

The derivative matrix is a block diagonal

$$D^{'} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_1} & \cdots & \frac{\partial \mu_T}{\partial \beta_1} \\ \frac{\partial \mu_1}{\partial \beta_2} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_T}{\partial \beta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_1}{\partial \beta_p} & \frac{\partial \mu_2}{\partial \beta_p} & \vdots & \frac{\partial \mu_T}{\partial \beta_p} \\ \frac{\partial \mu_1}{\partial \rho_1} & \frac{\partial \mu_2}{\partial \rho_1} & \cdots & \frac{\partial \mu_T}{\partial \rho_1} \\ \frac{\partial \mu_1}{\partial \rho_2} & \frac{\partial \mu_2}{\partial \rho_2} & \cdots & \frac{\partial \mu_T}{\partial \rho_2} \end{pmatrix}_{(p+2)\times T} \quad (9)$$

where the expressions for the above derivatives with respect to the unknown parameters are given in di erence forms:

1. $\frac{\partial \mu_t}{\partial \beta_j} = \rho_1 \frac{\partial \mu_{t-1}}{\partial \beta_j} + \rho_2 \lambda_{t-1} x_{t-1,j} + \lambda_t x_{tj}$ for t=2,...,T, with $\frac{\partial \mu_1}{\partial \beta_j} = \frac{(1+\rho_2)}{(1-\rho_1)} \lambda_t x_{tj}$.
2. $\frac{\partial \mu_t}{\partial \rho_1} = \rho_1 \frac{\partial \mu_{t-1}}{\partial \rho_1} + \mu_{t-1}$ for t=2,...,T, with $\frac{\partial \mu_1}{\partial \rho_1} = \frac{(1+\rho_2)\lambda_t}{(1-\rho_1)^2}$.
3. $\frac{\partial \mu_t}{\partial \rho_2} = \rho_1 \frac{\partial \mu_{t-1}}{\partial \rho_2} + \lambda_{t-1}$ for t=2,...,T, with $\frac{\partial \mu_1}{\partial \rho_2} = \frac{\lambda_t}{(1-\rho_1)}$.

The covariance structure $\Sigma$ comprises of the variance and covariance components, where $\mathrm{Var}(Y_t)$ and $\mathrm{Cov}(Y_t, Y_{t+h})$ are computed using equations (3) and (4).

The Newton-Raphson iterative technique is used to solve the GQL in equation (8) which yields

$$\left(\boldsymbol{\psi}_{r+1}\right) = \left(\boldsymbol{\psi}_r\right) + [\boldsymbol{D}^{'} \boldsymbol{\Sigma}^{-1} \boldsymbol{D}]_r^{-1} [\boldsymbol{D}^{'} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})]_r \tag{10}$$

where $\boldsymbol{\psi}_r = [\hat{\boldsymbol{\beta}}_r, \rho_{1r}, \rho_{2r}]$ are the estimates at the $r^{th}$ iteration and $[.]_r$ are the values of the expression at the $r^{th}$ iteration.

The GQL-II estimates the dispersion e ect and is given by

$$\boldsymbol{D}_{\nu}^{'} \boldsymbol{\Sigma}_{\nu}^{-1} (\boldsymbol{f} - \boldsymbol{\theta})_{\nu} = 0 \tag{11}$$

with $\boldsymbol{f} = [f_{1|0}, f_{2|1}, \ldots, f_{T|T-1}]^{'}$ where $f_{t|t-1} = \{Y_t^2 | Y_{t-1}, R_{t-1}\}$ and $E(\boldsymbol{f}) = \boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_T]^{'}$ where $\theta_t = E(Y_t^2 | Y_{t-1}, R_{t-1})$.

The derivative matrix is given by

$$D_{\nu}^{'} = \left( \frac{\partial \theta_1}{\partial \nu} \ \frac{\partial \theta_2}{\partial \nu} \ \ldots \ \frac{\partial \theta_T}{\partial \nu} \right)_{1 \times T} \tag{12}$$

where

$$E(Y_t^2 | Y_{t-1}, R_{t-1}) = \rho_1(1-\rho_1)Y_{t-1} + \rho_2(1-\rho_2)R_{t-1} + \nu\lambda_t + (\rho_1 Y_{t-1} + \rho_2 R_{t-1} + \lambda_t)^2 \tag{13}$$

and the entries of the derivative matrix (12) are $\frac{\partial \theta_t}{\partial \nu} = \lambda_t$.

The covariance structure $\Sigma_\nu$ is given by

$$\begin{pmatrix} Var({Y_1}^2 | Y_0, R_0) & Cov({Y_1}^2, {Y_2}^2 | Y_0, R_0, Y_1, R_1) & \ldots & Cov({Y_1}^2, {Y_T}^2 | Y_0, R_0, Y_{T-1}, R_{T-1}) \\ Cov({Y_2}^2, {Y_1}^2 | Y_0, R_0, Y_1, R_1) & Var({Y_2}^2 | Y_1, R_1) & \ldots & Cov({Y_2}^2, {Y_T}^2 | Y_1, R_1, Y_{T-1}, R_{T-1}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov({Y_T}^2, {Y_1}^2 | Y_0, R_0, Y_{T-1}, R_{T-1}) & Cov({Y_T}^2, {Y_2}^2 | Y_1, R_1, Y_{T-1}, R_{T-1}) & \ldots & Var({Y_T}^2 | Y_{T-1}, R_{T-1}) \end{pmatrix} \tag{14}$$

where the entries are computed using the exact thinning properties, where $\mathrm{Cov}({Y_t}^2, {Y_{t+h}}^2 | Y_{t-1}, R_{t-1}, Y_{t+h-1}, R_{t+h-1}) = 0$ and
$\mathrm{Var}({Y_t}^2 | Y_{t-1}, R_{t-1})$

$$\begin{aligned}
&= Var\left[(\rho_1 \circ Y_{t-1} + \rho_2 \circ R_{t-1} + R_t)^2 | Y_{t-1}, R_{t-1}\right] \\
&= Var\left[(\rho_1 \circ Y_{t-1})^2 + 2(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t) + (\rho_2 \circ R_{t-1} + R_t)^2 | Y_{t-1}, R_{t-1}\right] \\
&= Var[(\rho_1 \circ Y_{t-1})^2 | Y_{t-1}] + 4Var[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t) | Y_{t-1}, R_{t-1}] \\
&+ Var[(\rho_2 \circ R_{t-1} + R_t)^2 | R_{t-1}] + 4Cov[(\rho_1 \circ Y_{t-1})^2, (\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t) | Y_{t-1}, R_{t-1}] \\
&+ 4Cov[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t), (\rho_2 \circ R_{t-1} + R_t)^2 | Y_{t-1}, R_{t-1}]
\end{aligned} \tag{15}$$

where the components of equation (15) are computed as follows:

1.

$$Var[(\rho_1 \circ Y_{t-1})^2|Y_{t-1}] = E\left[(\rho_1 \circ Y_{t-1})^4|Y_{t-1}\right] - E\left[(\rho_1 \circ Y_{t-1})^2|Y_{t-1}\right]^2$$
$$= 4(1-\rho_1)\rho_1^3 Y_{t-1}^3 + 2(1-\rho_1)(3-5\rho_1)\rho_1^2 Y_{t-1}^2 + (1-\rho_1)(1-6\rho_1+6\rho_1^2)\rho_1 Y_{t-1}. \quad (16)$$

2.

$$4Var[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t)|Y_{t-1}, R_{t-1}]$$

$$= 4E[(\rho_1 \circ Y_{t-1})^2(\rho_2 \circ R_{t-1} + R_t)^2|Y_{t-1}, R_{t-1}] - 4E[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t)|Y_{t-1}, R_{t-1}]^2$$
$$= 4Y_{t-1}^2\left(\nu\,\lambda_t\rho_1^2 + R_{t-1}\,\rho_1^2(1-\rho_2)\rho_2\right)$$
$$+ 4Y_{t-1}\left(\lambda_t(\nu+\lambda_t)(1-\rho_1)\rho_1 + R_{t-1}(1-\rho_1)\rho_1(1+2\lambda_t-\rho_2)\rho_2 + R_{t-1}^2(1-\rho_1)\rho_1\rho_2^2\right). \quad (17)$$

3.

$$Var[(\rho_2 \circ R_{t-1} + R_t)^2|R_{t-1}]$$

$$= E[(\rho_2 \circ R_{t-1} + R_t)^4|R_{t-1}] - E[(\rho_2 \circ R_{t-1} + R_t)^2|R_{t-1}]^2$$
$$= 2\rho_2^2 R_{t-1}^2\left(-2\lambda_t^2 - 4(\lambda_t+2)\rho_2 + 4\lambda_t + 2\lambda_t(\nu+\lambda_t) + 5\rho_2^2 + 3\right) - 4(\rho_2-1)\rho_2^3 R_{t-1}^3$$
$$+ \rho_2 R_{t-1}\Big( - 4\lambda_t(\nu\lambda_t + \lambda_t^2 + \rho_2(3-2\rho_2) - 1) - 4\lambda_t(\nu+\lambda_t)(\rho_2-1)$$
$$+ 4E(R_t^3) - 6(\rho_2-2)\rho_2^2 - 7\rho_2 + 1\Big) - \lambda_t(\nu+\lambda_t)^2 + E(R_t^4). \quad (18)$$

4.

$$4Cov[(\rho_1 \circ Y_{t-1})^2, (\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t)|Y_{t-1}, R_{t-1}]$$

$$= 4E[(\rho_1 \circ Y_{t-1})^3(\rho_2 \circ R_{t-1} + R_t)|Y_{t-1}, R_{t-1}]$$
$$= 8Y_{t-1}^2\left(\lambda_t(1-\rho_1)\rho_1^2 + (1-\rho_1)\rho_1^2\rho_2 R_{t-1}\right)$$
$$+ 4Y_{t-1}\left(\lambda_t(1-\rho_1)\rho_1(1-2\rho_1) + (1-\rho_1)\rho_1(1-2\rho_1)\rho_2 R_{t-1}\right). \quad (19)$$

5.

$$4Cov[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t), (\rho_2 \circ R_{t-1} + R_t)^2|Y_{t-1}, R_{t-1}]$$

$$= 4E[(\rho_1 \circ Y_{t-1})(\rho_2 \circ R_{t-1} + R_t)^3|Y_{t-1}, R_{t-1}]$$
$$= 4Y_{t-1}\Big(\rho_1\rho_2 R_{t-1}\left(2\left(-\lambda_t^2 - \lambda_t\rho_2 + \lambda_t + \rho_2^2\right) + 2\lambda_t(\nu+\lambda_t) - 3\rho_2 + 1\right)$$
$$+ \rho_1(E(R_t^3) - \lambda_t\lambda_t(\nu+\lambda_t)) - 2\rho_1(\rho_2-1)\rho_2^2 R_{t-1}^2\Big). \quad (20)$$

Taken together, this gives

$$
\begin{aligned}
&Var({Y_t}^2|Y_{t-1}, R_{t-1}) \\
&\quad = Y_{t-1}\Big(4\rho_1\rho_2 R_{t-1}\left(2\lambda_t(\nu+2-\rho_1-\rho_2)+2\rho_1^2+\rho_1(\rho_2-4)+2(\rho_2-2)\rho_2+3\right) \\
&\quad + \rho_1\left(-4\lambda_t(\nu+\lambda_t)(\lambda_t+\rho_1-1)-4\lambda_t(\rho_1(3-2\rho_1)-1)+4E(R_t^3)-6(\rho_1-2)\rho_1^2-7\rho_1+1\right) \\
&\quad - 4\rho_1\rho_2^2 R_{t-1}^2(\rho_1+2\rho_2-3)\Big)+2\rho_2^2 R_{t-1}^2\left(4\lambda_t(1-\rho_2)+2\lambda_t\nu+5\rho_2^2-8\rho_2+3\right) \\
&\quad + Y_{t-1}^2\left(2\rho_1^2\left(4\lambda_t(1-\rho_1)+2\lambda_t\nu+5\rho_1^2-8\rho_1+3\right)-4\rho_1^2\rho_2 R_{t-1}(2\rho_1+\rho_2-3)\right) \\
&\quad + \rho_2 R_{t-1}\Big(-4\lambda_t(\lambda_t(\nu+\lambda_t)+\rho_2(3-2\rho_2)-1)-4\lambda_t(\nu+\lambda_t)(\rho_2-1) \\
&\quad + 4E(R_t^3)-6(\rho_2-2)\rho_2^2-7\rho_2+1\Big)-\lambda_t(\nu+\lambda_t)^2+E(R_t^4) \\
&\quad + 4(1-\rho_2)\rho_2^3 R_{t-1}^3+4(1-\rho_1)\rho_1^3 Y_{t-1}^3. \qquad (21)
\end{aligned}
$$

The Newton-Raphson iterative technique is used to solve the GQL-II (11) which yields

$$\left(\nu_{r+1}\right)=\left(\nu_r\right)+[\boldsymbol{D}_\nu^{'}\boldsymbol{\Sigma}_\nu^{-1}\boldsymbol{D}_\nu]_r^{-1}[\boldsymbol{D}_\nu^{'}\boldsymbol{\Sigma}_\nu^{-1}(\boldsymbol{f}-\boldsymbol{\theta})_\nu]_r \qquad (22)$$

where $\nu_r$ is the estimate at the $r^{th}$ iteration and $[.]_r$ is the value of the expression at the $r^{th}$ iteration.

The algorithm works as follows: Using the updated estimates from equation (10) and an initial value of $\nu$, we compute the dispersion index using the iterative equation (22). The updated value of $\nu$ is then used to re-calculate an updated value of $\boldsymbol{\psi}$ using the iterative equation (10), which is in turn used to calculate another updated value of $\nu$. This cycle continues until the convergence criterion $||\hat{\boldsymbol{\psi}}_{r+1}-\hat{\boldsymbol{\psi}}_r||<10^{-5}$ and $||\nu_{r+1}-\nu_r||<10^{-5}$.

Let us assume

$$g(\boldsymbol{\psi})=\boldsymbol{D}_{\boldsymbol{\psi}}{}^{'}[\boldsymbol{\Sigma}(\boldsymbol{\psi})]^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\psi})). \qquad (23)$$

Since the mean score, $\boldsymbol{\mu}(\boldsymbol{\psi})$, is correctly specified in the above, $E(g(\boldsymbol{\psi}))=0$ and the Fisher information matrix [16] is $\mathrm{Cov}(g(\boldsymbol{\psi}))=\left[\boldsymbol{D}_{\boldsymbol{\psi}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\psi}}\right]$. Hence, as $T\to\infty$,

$$\sqrt{T}[(\hat{\boldsymbol{\psi}})_{GQL}-(\boldsymbol{\psi})]\to N\left(0,[\boldsymbol{D}_{\boldsymbol{\psi}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\psi}}]^{-1}\right).$$

Similarly,

$$\sqrt{T}[(\nu)_{GQL}-(\nu)]\to N\left(0,[\boldsymbol{D}_\nu^{'}\boldsymbol{\Sigma}_\nu^{-1}\boldsymbol{D}_\nu]^{-1}\right).$$

## 4 Simulation Study

This section presents a simulation study where the INARMA(1,1) count data are generated using the observation-driven equation (1). Hence, assuming $Y_1 = R_1$ such that $Y_2 = \rho_1 \circ Y_1 + \rho_2 \circ R_1 + R_2$, and thereon subsequent values of $Y_t$, $t = 3, ..., T$ are generated, with $\lambda_t = (\beta_1 x_{t1} + \beta_2 x_{t2})$, where

$$x_{t1} = \begin{cases} 1.2 & (t = 1, \ldots, T/4), \\ 3t & (t = (T/4) + 1, \ldots, 3T/4), \\ cos(\frac{2\pi t}{6}) & (t = (3T/4) + 1, \ldots, T), \end{cases}$$

$$x_{t2} = \begin{cases} sin(\frac{3\pi t}{6}) & (t = 1, \ldots, T/4), \\ cos(\frac{\pi t}{12}) & (t = (T/4) + 1, \ldots, 3T/4), \\ sin(\frac{2\pi t}{12}) & (t = (3T/4) + 1, \ldots, T), \end{cases}$$

for $T = 60, 500, 1000$. As for the innovation term, we proposed two cases: first $R_t$ is assumed to follow the Poisson distribution ($\nu = 1$) and second, it is assumed to follow the negative binomial distribution ($\nu > 1$).

### 4.1 Poisson distribution ($v = 1$)

If $R_t$ follows the Poisson distribution, that is, $R_t \sim P(\lambda_t)$, then in Equation (6),

$$f_2(r_t = y_t - k) = \frac{e^{-\lambda_t}(\lambda_t)^{y_{t-1}-k}}{(y_{t-1} - k)!}. \tag{24}$$

As for $E(R_t^3)$ and $E(R_t^3)$ in Equation (21), the moment generating function (mgf) $M_{R_t}(q) = \exp(\lambda_t(e^q - 1))$ is di erentiated repeatedly to obtain

$$E(R_t^3) = \lambda_t + 3\lambda_t^2 + \lambda_t^3 \tag{25}$$

and

$$E(R_t^4) = \lambda_t + 7\lambda_t^2 + 6\lambda_t^3 + \lambda_t^4. \tag{26}$$

Hence, 5000 Monte Carlo replications are implemented assuming $[\rho_1, \rho_2] = ([0.2, 0.3], [0.2, 0.6], [0.7, 0.6])$ and $[\beta_1, \beta_2] = [1, 2]$ and the results are shown in Table 1:

**Table 1.** Mean estimated parameters and the corresponding estimated standard errors (SE) for the INARMA(1,1) model.

| True value | | | | $[\beta_1, \beta_2] = [1, 2]$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $\rho_2$ | T | Method | $\widehat{\beta_1}$ | SE | $\widehat{\beta_2}$ | SE | $\widehat{\rho_1}$ | SE | $\widehat{\rho_2}$ | SE |
| 0.2 | 0.3 | 60 | GQL | 0.9813 | (0.1780) | 0.9809 | (0.1850) | 0.1855 | (0.0754) | 0.2862 | (0.0679) |
| | | | CML | 0.9837 | (0.1711) | 0.9820 | (0.1794) | 0.1870 | (0.0705) | 0.2881 | (0.0630) |
| | | 500 | GQL | 0.9966 | (0.0980) | 0.9971 | (0.1012) | 0.1956 | (0.0397) | 0.3041 | (0.0308) |
| | | | CML | 0.9978 | (0.0970) | 0.9988 | (0.1001) | 0.1982 | (0.0388) | 0.3059 | (0.0295) |
| | | 1000 | GQL | 1.0014 | (0.0544) | 0.9989 | (0.0521) | 0.1978 | (0.0091) | 0.2987 | (0.0085) |
| | | | CML | 0.9990 | (0.0547) | 0.9994 | (0.0518) | 0.1980 | (0.0093) | 0.2990 | (0.0084) |
| 0.2 | 0.6 | 60 | GQL | 0.9861 | (0.1811) | 0.9831 | (0.1812) | 0.1819 | (0.0813) | 0.5833 | (0.0690) |
| | | | CML | 0.9879 | (0.1761) | 0.9839 | (0.1764) | 0.1832 | (0.0756) | 0.5851 | (0.0643) |
| | | 500 | GQL | 0.9927 | (0.0974) | 0.9932 | (0.1053) | 0.1934 | (0.0401) | 0.5948 | (0.0360) |
| | | | CML | 0.9960 | (0.0955) | 0.9953 | (0.1040) | 0.1959 | (0.0386) | 0.5957 | (0.0339) |
| | | 1000 | GQL | 0.9990 | (0.0585) | 0.9975 | (0.0416) | 0.1978 | (0.0151) | 0.5985 | (0.0129) |
| | | | CML | 0.9992 | (0.0582) | 0.9986 | (0.0412) | 0.1989 | (0.0150) | 0.5989 | (0.0126) |
| 0.7 | 0.6 | 60 | GQL | 0.9854 | (0.1734) | 0.9837 | (0.1773) | 0.6836 | (0.0765) | 0.5811 | (0.0640) |
| | | | CML | 0.9869 | (0.1682) | 0.9851 | (0.1737) | 0.6855 | (0.0712) | 0.5844 | (0.0592) |
| | | | CML | 0.9940 | (0.1540) | 0.9935 | (0.1550) | 0.6938 | (0.0528) | 0.5917 | (0.0520) |
| | | 500 | GQL | 0.9972 | (0.1046) | 0.9950 | (0.0923) | 0.7035 | (0.0331) | 0.6035 | (0.0356) |
| | | | CML | 0.9983 | (0.1029) | 0.9961 | (0.0910) | 0.6976 | (0.0330) | 0.6030 | (0.0340) |
| | | 1000 | GQL | 1.0010 | (0.0511) | 0.9979 | (0.0410) | 0.6991 | (0.0105) | 0.5980 | (0.0111) |
| | | | CML | 0.9995 | (0.0510) | 0.9990 | (0.0406) | 0.7005 | (0.0099) | 0.5991 | (0.0108) |

From Table 1, it can be deduced that both GQL and CML yield e cient estimators of the regression and dependence parameters based on the di erent combinations, with CML yielding estimators with slightly lower standard errors.

### 4.2 Negative Binomial distribution ($\nu > 1$)

Assuming that $R_t$ follows the negative binomial distribution, that is, $R_t \sim NB(\frac{\lambda_t}{\nu-1}, \frac{\nu-1}{\nu})$, then in Equation (6),

$$f_2(r_t = y_t - k) = \binom{y_{t-1} - k + \frac{\lambda_t}{\nu-1} - 1}{y_{t-1} - k} (1 - \frac{\nu-1}{\nu})^{(\frac{\lambda_t}{\nu-1})} (\frac{\nu-1}{\nu})^{y_{t-1}-k}. \quad (27)$$

As for $E(R_t^3)$ and $E(R_t^3)$ in Equation (21), we di erentiate the mgf $M_{R_t}(q) = \left[\frac{1}{\nu-(\nu-1)\exp(q)}\right]^{\frac{\lambda_t}{\nu-1}}$ repeatedly to obtain

$$E(R_t^3) = \lambda_t[3\nu - 2 + 2(\nu-1)^2] + 3\lambda_t^2\nu + \lambda_t^3 \quad (28)$$

and

$$E(R_t^4) = \lambda_t[7\nu-6+12(\nu-1)^2+6(\nu-1)^3]+\lambda_t^2[18\nu-11+11(\nu-1)^2]+6\lambda_t^3\nu+\lambda_t^4. \quad (29)$$

Based on the assumption that $R_t$ follows the negative binomial distribution, 5000 Monte Carlo replications are implemented using $[\rho_1, \rho_2] = ([0.2, 0.3], [0.2, 0.6], [0.7, 0.6])$, $[\beta_1, \beta_2] = [1, 2]$ and $\nu = 2$ and the results are shown in Table 2:

**Table 2.** Mean estimated parameters and the corresponding estimated standard errors (SE) for the INARMA(1,1) model.

| True value | | | | $[\beta_1, \beta_2] = [1, 2]$ | | | | | | | | $\nu = 2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $\rho_2$ | T | Method | $\widehat{\beta_1}$ | SE | $\widehat{\beta_2}$ | SE | $\widehat{\rho_1}$ | SE | $\widehat{\rho_2}$ | SE | $\widehat{\nu}$ | SE |
| 0.2 | 0.3 | 60 | GQL | 0.9821 | (0.1882) | 1.9858 | (0.1885) | 0.1844 | (0.0765) | 0.2811 | (0.0640) | 1.9854 | (0.1734) |
| | | | CML | 0.9843 | (0.1811) | 1.9896 | (0.1817) | 0.1855 | (0.0712) | 0.2844 | (0.0592) | 1.9869 | (0.1682) |
| | | 500 | GQL | 0.9931 | (0.1081) | 1.9953 | (0.1018) | 0.2035 | (0.0331) | 0.3035 | (0.0356) | 1.9972 | (0.1046) |
| | | | CML | 0.9960 | (0.1060) | 1.9973 | (1.9992) | 0.1976 | (0.0330) | 0.3030 | (0.0340) | 1.9983 | (0.1029) |
| | | 1000 | GQL | 0.9980 | (0.0618) | 1.9972 | (0.0656) | 0.1991 | (0.0105) | 0.2980 | (0.0111) | 2.0010 | (0.0511) |
| | | | CML | 1.0021 | (0.0615) | 2.0022 | (0.0650) | 0.2005 | (0.0099) | 0.2991 | (0.0108) | 1.9995 | (0.0510) |
| 0.2 | 0.6 | 60 | GQL | 0.9848 | (0.1785) | 1.9812 | (0.1718) | 0.1855 | (0.0754) | 0.5862 | (0.0679) | 1.9813 | (0.1780) |
| | | | CML | 0.9880 | (0.1734) | 1.9889 | (0.1650) | 0.1870 | (0.0705) | 0.5881 | (0.0630) | 1.9837 | (0.1711) |
| | | 500 | GQL | 0.9940 | (0.1019) | 1.9970 | (0.1119) | 0.1956 | (0.0397) | 0.6041 | (0.0308) | 1.9966 | (0.0980) |
| | | | CML | 0.9959 | (0.0995) | 1.9981 | (0.1088) | 0.1982 | (0.0388) | 0.6059 | (0.0295) | 1.9978 | (0.0970) |
| | | 1000 | GQL | 1.0036 | (0.0625) | 1.9979 | (0.0563) | 0.1978 | (0.0091) | 0.5987 | (0.0085) | 2.0014 | (0.0544) |
| | | | CML | 1.0013 | (0.0620) | 2.0023 | (0.0560) | 0.1980 | (0.0093) | 0.5990 | (0.0084) | 1.9990 | (0.0547) |
| 0.7 | 0.6 | 60 | GQL | 0.9856 | (0.1712) | 1.9815 | (0.1810) | 0.6819 | (0.0813) | 0.5833 | (0.0690) | 1.9861 | (0.1811) |
| | | | CML | 0.9889 | (0.1660) | 1.9875 | (0.1755) | 0.6832 | (0.0756) | 0.5851 | (0.0643) | 1.9879 | (0.1761) |
| | | 500 | GQL | 0.9957 | (0.1088) | 1.9963 | (0.1016) | 0.6934 | (0.0401) | 0.5948 | (0.0360) | 1.9927 | (0.0974) |
| | | | CML | 0.9971 | (0.1075) | 1.9977 | (0.0990) | 0.6959 | (0.0386) | 0.5957 | (0.0339) | 1.9960 | (0.0955) |
| | | 1000 | GQL | 0.9980 | (0.0628) | 2.0030 | (0.0452) | 0.6978 | (0.0151) | 0.5985 | (0.0129) | 1.9990 | (0.0585) |
| | | | CML | 1.0012 | (0.0623) | 2.0010 | (0.0450) | 0.6989 | (0.0150) | 0.5989 | (0.0126) | 1.9992 | (0.0582) |

From Table 2, we observe that for $\nu > 1$ also, both GQL and CML yield e cient estimators of the regression and dependence parameters.

## 5 Conclusion

In this paper, a novel non-stationary INARMA(1,1) model is introduced. The marginal moments and covariance expressions are derived under a generalized assumption of the innovation term. As for the inferential procedures, we consider two estimation approaches, namely CML and GQL. Furthermore, two simulation experiments are conducted to evaluate the performance of the model estimators, by assuming that the innovation term follows a Poisson and a negative binomial distribution, with the CML estimators slightly more e cient than the GQL estimators.

## References


1. Brannas, K.: Explanatory variable in the AR(1) count data model. Umea University, Department of Economics No.381, 1–21 (1995)
2. Brannas, K., Quoreshi, A.: Integer-valued moving average modelling of the number of transactions in stocks. Applied Financial Economics No.20(18), 1429–1440 (2010)
3. Chan, K., Ledolter, J.: Monte carlo EM estimation for time series models involving counts. Journal of the American Statistical Association 90(429), 242–252 (1995)
4. Cox, D.: Statistical analysis of time series: Some recent developments. Scandinavian Journal of Statistics 8(2), 93–115 (1981)
5. Durbin, J., Koopman, S.: Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. Journal of the Royal Statistical Society B62, 3–56 (2000)
6. Jowaheer, V., Sutradhar, B.: Fitting lower order nonstationary autocorrelation models to the time series of Poisson counts. Transactions on Mathematics 4, 427–434 (2005)

7. Mamode Khan, N., Jowaheer, V.: Comparing joint GQL estimation and GMM adaptive estimation in COM-Poisson longitudinal regression model. Commun Stat-Simul C. 42(4), 755–770 (2013)
8. McKenzie, E.: Some simple models for discrete variate time series. Water Resources Bulletin 21(4), 645–650 (1985)
9. McKenzie, E.: Autoregressive moving-average processes with Negative Binomial and geometric marginal distributions. Advanced Applied Probability 18, 679–705 (1986)
10. McKenzie, E.: Some ARMA models for dependent sequences of Poisson counts. Advances in Applied Probability 20, 822–835 (1988)
11. Al Osh, M., Alzaid, A.: First-order integer-valued autoregressive process. Journal of Time Series Analysis 8, 261–275 (1987)
12. Al Osh, M., Alzaid, A.: Integer-valued moving average (INMA) process. Statistical Papers 29, 281–300 (1988a)
13. Pedeli, X., Karlis, D.: Some properties of multivariate INAR(1) processes. Computational Statistics and Data Analysis 67, 213–225 (2013a)
14. Ravishanker, N., Serhiyenko, V., Willig, M.: Hierarchical dynamic models for multivariate time series of counts. Statistics and Its Interface 7, 559–570 (2014)
15. Steutel, F., Van Harn, K.: Discrete analogues of self-decomposability and stability. The Annals of Probability 7, 3893–3899 (1979)
16. Wedderburn, R.: Quasi-likelihood functions,generalized linear models,and the Gauss-Newton method. Biometrika 61(3), 439–47 (December 1974)
17. Zeger, S.: A regression model for time series of counts. Biometrika 75, 621–629 (1988)

# Numerical Study of the Conditional Time Series of the Average Daily Heat Index

Nina Kargapolova[1,2][0000-0002-1598-7675]

[1] Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, 630090, Russia
[2] Novosibirsk State University, Novosibirsk, 630090, Russia
nkargapolova@gmail.com

**Abstract.** In this paper the results of the numerical study of the statistical properties of the conditional time series of the average daily heat index (ADHI) are presented. Both point and interval conditions on the values of the ADHI were imposed. The study was carried out on a basis of the real data as well as the stochastic models of the non-stationary non-Gaussian conditional time series of the ADHI. To simulate the time series with point conditions, the conditional distributions method and the inverse distribution function method were applied. In case of interval conditions, the straightforward enumeration of the non-conditional trajectories was used. It is shown that the trajectories of the models proposed are close in their statistical properties to the real time series of the ADHI. Simulated trajectories of the ADHI are used to study statistical properties of some unfavorable (in sense of the ADHI) weather events.

**Keywords:** Average Daily Heat Index, Stochastic Simulation, Non-stationary Random Process, Non-Gaussian Random Process, Conditional Random Time Series.

## 1 Introduction

According to [1], during the last decades there arise many discussions about how to define thermal comfort and how to grade thermal stress. These efforts have resulted in various models attempting to describe thermal comfort and the resultant thermal stress. A large number of the biometeorological indices, that describe the combined thermal effect of various meteorological elements on human beings, have been proposed. In [1–3] one may find a comprehensive review of such biometeorological indices. To describe the effects of the high air temperature and relative humidity, the heat index is often used. A stochastic approach to the study and simulation of the heat index time series was first proposed in [4]. This paper presents some results of the development of this approach related to the study of the conditional time-series of the average daily heat index (ADHI). The study of the properties of the conditional time-series of the ADHI and the development of the proper stochastic models are relevant to stochastic forecasting of extreme weather events (such as heat waves) and development of the Heat–Health Warning Systems [2].

## 2 Heat Index

To describe the combined effect on a human being of several weather factors during the warm period of a year the Steadman's apparent temperature is frequently used [5–6]. The heat index is a simplified version of the apparent temperature that relies only on air temperature and air relative humidity [7]. An overview and comparison of existing approaches to the defining this index is provided in [7]. In this paper, the average daily heat index $HI$ is defined using the approach, proposed in [8]:

$$HI = T - 1.0799e^{0.03755T}\left(1 - e^{0.0801(D-14)}\right),\ D = \frac{237.3\alpha}{17.27-\alpha},\ \alpha = \frac{17.27T}{237.3+T} + lnH, \quad (1)$$

where $T$ and $H$ are the average daily air temperature (in a Celsius degree) and the average daily relative humidity (expressed as a decimal fraction), respectively, $D$ is the dew point temperature (in a Celsius degree). Unit of measurement of the heat index is supposed to be a Celsius degree. It should be noted that the heat index is not measured at weather stations, but it could be calculated using (1) with the observed values of air temperature and relative humidity.

## 3 Stochastic Models

Since the heat index is a function of air temperature and relative humidity, a natural approach to the simulation of its time series is to simulate the joint time series of these two weather elements and then to calculate, using (1), values of the heat index. Such an approach was proposed and validated in [4] for the simulation of the high-resolution non-conditional time series of the heat index at short time intervals. The model proposed therein is based on the model of the periodically correlated joint time series of air temperature and relative humidity detailed in [9]. The same approach could also be used for simulation of the conditional time series with the conditions imposed either on one or two weather elements. For simulation of the time series at long time intervals (like month-, season-, year-long) it is necessary to use an assumption that the real weather processes are nonstationary instead of the assumption about their periodically correlated structure. This makes possible to take into account both daily and seasonal variation of the real processes, but it also leads to increase in the simulation complexity.

Another approach to the simulation of the non-stationary time series $\overrightarrow{HI} = \left(HI_1, HI_2, \ldots, HI_N\right)$ of the ADHI on a $N$ – day interval is considered in [10]. In the framework of this approach, at the first step, a sample of the real time series of the ADHI is formed using the long-term observation data about the average daily temperature and relative humidity. Then the sample one-dimensional distributions of the ADHI are approximated with the mixtures $g_k(x),\ k = \overline{1,N}$ of the two Gaussian distributions (with the corresponding CDFs $G_k(x),\ k = \overline{1,N}$) and the sample $N \times N$ correlation matrix $R_{HI}$ is estimated. The last step is the simulation of trajectories of the

ADHI with the given CDFs $G_k(x)$, $k=\overline{1,N}$ and the given correlation matrix $R_{HI}$ using the method of inverse distribution function for simulation of the non-Gaussian time series [11].

To solve some applied problems of the Bioclimatology (for example, problems related to forecasting of dangerous values of the bioclimatic indices, to bioclimatic territorial zoning etc.) it is necessary to study the properties of the conditional time series. In this paper, two types of conditions are considered. Conditions of the first type are point conditions, when several values (either consecutive or inconsecutive) of the ADHI are given, namely, $HI_j = c_j$, $j \in \Omega, \Omega \subset \{1,2,\ldots,N\}$, $c_j \in \mathbb{R}$. The interval conditions $HI_j \in (a_j, b_j)$, $j \in \Omega, \Omega \subset \{1,2,\ldots,N\}$, $-\infty \le a_j, b_j \le +\infty$ form the second type of conditions. An interval $(a_j, b_j)$ could be closed, semi-open or open. Since the sample size of the real data collected at the weather stations is usually relatively small, it is not always possible to reliably estimate statistical characteristics of the nonconditional time series. Imposed conditions even more decrease the sample size. If the conditions are strict enough (for example, levels $c_j$ are high or intervals $(a_j, b_j)$ are short), the real data may not even contain the proper trajectories, but it does not mean that the fulfilment of these conditions is physically unfeasible. In such situation, one has to simulate time series that a) are close in their statistical properties to the real time series and b) satisfy the conditions, and then to study the properties of the conditional time series using the simulated trajectories.

Let us describe two stochastic models of the conditional time series of the ADHI both with the point and interval conditions.

### 3.1 Conditional model with point conditions

Let $\overrightarrow{HI^k} = (HI_1^k, HI_2^k, \ldots, HI_N^k)$, $k=\overline{1,Y}$ denote the real ADHI, calculated by the real air temperature and relative humidity data collected at a weather station during $Y$ years. The simulation algorithm of the conditional time series with the point conditions $HI_j = c_j$, $j \in \Omega, \Omega \subset \{1,2,\ldots,N\}$, $c_j \in \mathbb{R}$ could be presents as the following sequence of steps:

1. Transformation $\eta_j^k = \Phi\left(G_j^{-1}\left(HI_j^k\right)\right)$, $k=\overline{1,Y}$, $j=\overline{1,N}$. Here $\Phi(\cdot)$ is the CDF of the standard normal distribution. After such transformation conditions $HI_j = c_j$, $j \in \Omega, \Omega \subset \{1,2,\ldots,N\}$, $c_j \in \mathbb{R}$ turn into conditions $\eta_j = c_j$, $j \in \Omega$.
2. Estimation on $\eta_j^k$, $k=\overline{1,Y}$, $j=\overline{1,N}$ of the $N \times N$ correlation matrix

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

Here $R_{11}$ consists of all correlation coefficients between $\eta_j$, $j \in \Omega$, $R_{22}$ is a matrix of the correlation coefficients between the free components of the simulated vector and $R_{12}$, $R_{21}$ are corresponding cross-correlation matrixes.

3. Simulation of a conditional Gaussian vector $(\eta_1, \eta_2, \ldots, \eta_N)$ with the correlation matrix $R$ and conditions $\eta_j = c_j$, $j \in \Omega$. An effective algorithm to simulate conditional Gaussian vector is described in details in [11, 12].
4. Transformation $HI_j = G_j^{-1}\left(\Phi\left(\eta_j\right)\right)$, $j = \overline{1,N}$ of the Gaussian sequence $\eta_1, \ldots, \eta_N$ in the time series $\overrightarrow{HI} = (HI_1, HI_2, \ldots, HI_N)$ with the one dimensional distributions $G_k(x)$, $k = \overline{1,N}$ and satisfying the conditions $HI_j = c_j$, $j \in \Omega$.

Steps 3 and 4 are repeated as many times as many trajectories are required.

### 3.2 Conditional model with interval conditions

It is not possible to simulate the conditional time series with the interval conditions exactly as in case of the point conditions. The problem is that there is no effective algorithm to simulate the conditional Gaussian time series with the interval conditions. Accordingly, in is necessary to apply another approach. In this paper, a simple algorithm based in simulation of non-conditional time series (see, [10]) and following straightforward enumeration is used:

1. Calculation of the correlation matrix $R'_{HI}$ of the auxiliary Gaussian process. Each entry $r'(i,j)$ of the matrix $R'_{HI}$ is the solution of the equation

$$r(i,j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G_i^{-1}\left(\Phi(x)\right) G_j^{-1}\left(\Phi(y)\right) \varphi\left(x, y, r'(i,j)\right) dxdy,$$

where $r(i,j)$ is the corresponding entry of the matrix $R_{HI}$ and $\varphi(\cdot,\cdot,\cdot)$ is the pdf of a bivariate Gaussian vector with zero mean, variance equal to 1 and the correlation coefficient $r'(i,j)$.

2. Simulation of a Gaussian non-conditional vector $(\xi_1, \xi_2, \ldots, \xi_N)$ with the correlation matrix $R'_{HI}$.
3. Check if the vector $(\xi_1, \xi_2, \ldots, \xi_N)$ fulfills the conditions $\xi_j \in \left(a'_j, b'_j\right)$, $j \in \Omega$, where $a'_j = \Phi\left(G_j^{-1}\left(a_j\right)\right)$, $b'_j = \Phi\left(G_j^{-1}\left(b_j\right)\right)$, $j \in \Omega$. If the conditions are not satisfied, the trajectory is "forgotten", and a new one is simulated (steps 2 and 3 are repeated).

4. If all the conditions are satisfied, then the Gaussian vector is transformed in a non-Gaussian one: $HI_j = G_j^{-1}\left(\Phi\left(\xi_j\right)\right),\ j = \overline{1,N}$.

This algorithm is time-consuming (especially in case of strict conditions that are rarely met), but in seems to be the only approach to simulation of the non-Gaussian time series with such a type of conditions.

## 4 Numerical Experiments

Any stochastic model has to be verified before one starts to use simulated trajectories to study properties of a simulated process. For the model verification, it is necessary to compare the simulated and real data based on estimations of such characteristics, which, on the one hand, are reliably estimated by real data, and on the other hand are not input parameters of the model. In this chapter, several examples of such characteristics are given.

Estimations of all characteristics of the time series of the real ADHI were done on a basis of meteorological observations data collected at the weather stations, situated in different Russian cities. Although all examples in this paper are given only for the stations in the cities of Sochi (the Black Sea region, years of observation: 1966-2015) and Astrakhan (the Caspian Sea region, years of observation: 1966-2000), all the conclusions are valid for all considered weather stations.

It should be noted that it is possible to simulate as many trajectories as needed to provide a required accuracy of a characteristic estimation. In this paper, for all estimations based on the simulated data we have attained the accuracy above $10^{-4}$, so in all the tables presented, the estimations based on the simulated trajectories are given with significant digits only.

From now on, $\sigma$ is a standard deviation of the characteristic under consideration when estimating with the real data; $\mu_j$, $s_j$ are the mean value and the standard deviation of the ADHI $HI_j$; RD and SD denote the estimations based on the real and simulated data, respectively.

The first example of the characteristic that was used for verification of the model with interval conditions is the average number $AN\left(lev\right)$ of the days in a considered time-interval with the ADHI above the given level $lev$. Tab. 1 and Tab. 2 show the estimations of the $AN\left(lev\right)$, obtained on the real and simulated data under the condition $HI_1 > \mu_1, HI_2 > \mu_2$. For all the levels $lev$, all the considered weather stations and time intervals, the absolute difference of $AN\left(lev\right)$ estimated on the real and simulated data does not exceed $\sigma$. This means that this characteristic is well reproduced by the model.

The other example of the characteristics used for the model validation is the probability $p\left(l\right) = P\left(\left|HI_i - HI_{i+1}\right| > l \mid A\right)$ of a rapid change in the average daily heat index

under the condition $A$. The real and simulated data based on estimations of the $p(l)$ are shown in Tab. 3. Here the condition $A$ is $|HI_1 - HI_2| > l$. The numerical analysis shows that for all the considered weather stations, time intervals and levels $l$ deviations of the estimations based on the simulated data from the corresponding estimations based on the real data do not exceed $3\sigma$.

**Table 1.** The average number $AN(lev)$ of the days in a considered time-interval with the ADHI above given level $lev$, when $HI_1 > \mu_1$, $HI_2 > \mu_2$. Sochi. July, 1-15.

| $lev$ | RD, $AN(lev) \pm \sigma$ | SD |
|---|---|---|
| 20 | $15.000 \pm 0.210$ | 14.847 |
| 24 | $12.875 \pm 1.074$ | 12.145 |
| 28 | $5.500 \pm 1.351$ | 5.809 |
| 32 | $0.750 \pm 0.569$ | 1.059 |
| 36 | $0.000 \pm 0.079$ | 0.043 |

**Table 2.** The average number $AN(lev)$ of the days in a considered time-interval with the ADHI above given level $lev$, when $HI_1 > \mu_1$, $HI_2 > \mu_2$. Astrakhan. July, 16-30.

| $lev$ | RD, $AN(lev) \pm \sigma$ | SD |
|---|---|---|
| 20 | $15.000 \pm 0.130$ | 14.938 |
| 24 | $13.818 \pm 0.485$ | 14.095 |
| 28 | $10.455 \pm 0.905$ | 11.082 |
| 32 | $5.182 \pm 0.918$ | 4.927 |
| 36 | $2.000 \pm 0.555$ | 1.459 |
| 40 | $0.273 \pm 0.143$ | 0.150 |
| 44 | $0.000 \pm 0.021$ | 0.003 |
| 46 | $0.000 \pm 0.010$ | 0.001 |

**Table 3.** The probability $p(l)$ of a rapid change in the heat index under a condition $|HI_1 - HI_2| > l$. Astrakhan. August, 1-31.

| $l$ | RD, $p(l) \pm 3\sigma$ | SD |
|---|---|---|
| 1 | $0.683 \pm 0.048$ | 0.711 |
| 2 | $0.433 \pm 0.070$ | 0.477 |
| 3 | $0.317 \pm 0.102$ | 0.302 |
| 4 | $0.208 \pm 0.078$ | 0.194 |
| 5 | $0.067 \pm 0.178$ | 0.124 |

The verification of the model with the interval conditions has shown that this model with high accuracy reproduces many of the statistical characteristics of real ADHI time series. Accordingly, it is possible to use the model in question to study those properties of the time series that cannot be studied using the real data.

As an example of such study, let us consider how the mean value $m(j)$ of the AHDI $HI_j$, $j > i$ depends on $j$ when the conditions $HI_k > a_k$, $k = \overline{1,i}$ are imposed on $HI_k$, $k = \overline{1,i}$. Fig. 1 shows the simulated data based on estimations of $m(j)$ for $N = 15, i = 5.$ This choice of $i$ is associated with the definition of a heat wave [13]. For comparison, the mean values of the non-conditional time series of the ADHI are also shown. It could be clearly seen that high average values of the ADHI at the beginning of the time interval under consideration lead to the significant increase in the mean values during the whole time interval. This means that the heat waves are longstanding and they influence protractedly on the average heat index.

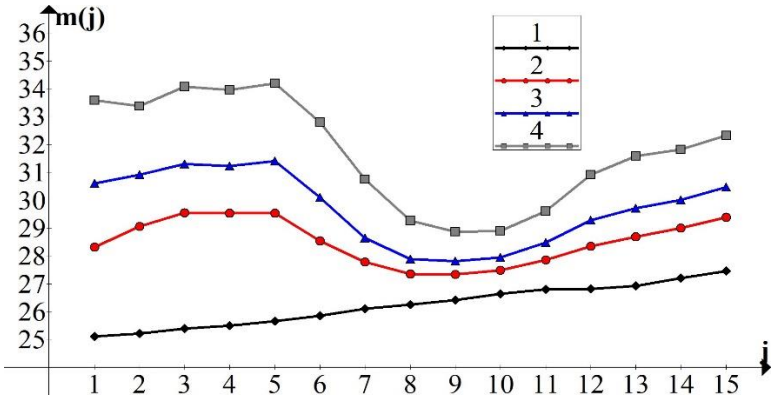



**Fig. 1.** Mean values of the ADHI. Curve 1 – non-conditional model, curves 2-4 – $a_k = \mu_k$, $a_k = \mu_k + s_k$, $a_k = \mu_k + 2s_k$, respectively. Sochi, July, 1-15.

The model with the point condition was also verified. The results of the verification have shown that the model sufficiently well reproduces the properties of the real process only for $N \le 15$. The reason why the model do not reproduce the properties of the real ADHI time series for larger values of $N$ requires further investigations. As an example of the model application in a case when the model "works" well, Fig. 2 shows the simulated trajectories based on estimations of the probabilities $pc(i,l) = P\left(HI_i > l \mid HI_j = c_j,\ j = \overline{1,7}\right)$. The values $c_j$ are the real values of the ADHI in Astrakhan in July 5-11th, 2001 and all the values $c_j > 25\,^oC$. For comparison, the unconditional probabilities $pu(i,l) = P\left(HI_i > l\right)$ are also shown. On can clearly see that relatively high values of the ADHI in the first part of the time interval in question influence on the probability of the high ADHI during the second part of the interval.

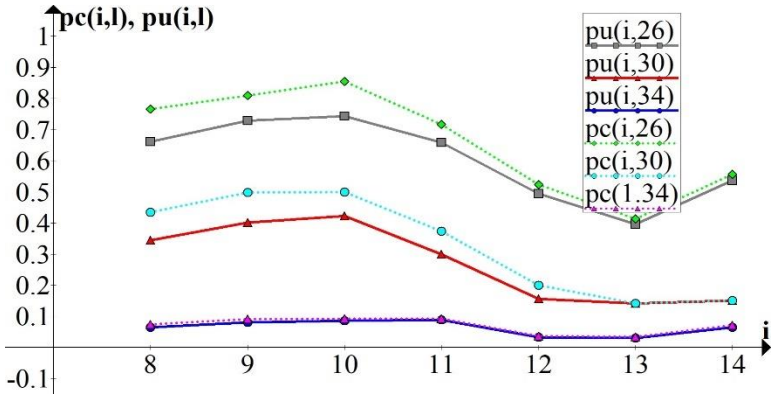



**Fig. 2.** The conditional and the unconditional probabilities $pc(i,l)$ and $pu(i,l)$, respectively. Sochi, July, 1-15.

## 5 Conclusion

In this paper, some preliminary results related to the study and simulation of the conditional time-series of the average daily heat index are presented. It is shown that simple models of the conditional time series with point or interval conditions describe well the real process.

In the future, it is intended to use the models constructed for solving a number of bioclimatological problems related to the development of the heatwaves prediction systems and to the studying the influence of a climate change on the properties of the heat index time series. The necessary conditions for the solution of these problems, are a transformation of the models into a fully parametric models, detailed study of the sensitivity of the models to small variations of their input parameters and development of the time-efficient algorithms of simulation of the conditional time series with the interval conditions.

## 6 Acknowledgements

This work was partly financially supported by the Russian Foundation for Basic Research (grant No 18-01-00149-a), Russian Foundation for Basic Research and Government of Novosibirsk region (grant No 19-41-543001-r_mol_a).

## References

1. Blazejczyk, K., Epstein, Y., Jendritzky, G., Staiger, H., Tinz, B.: Comparison of UTCI to selected thermal indices. Int. J. Biometeor. 56, 515–535 (2012).
2. Heatwaves and Health: Guidance on Warning-System Development. Ed. by G.R. McGregor. WMO, Geneva (2015).

3. Kobisheva, N.V., Stadnik, V.V., Klueva, M.V., Pigoltsina, G.B., Akentieva, E.M., Galuk, L.P., Razova, E.N., Semenov, U.A.: Guidance on specialized climatological service of the economy. Asterion, St. Petersburg (2008). [in Russian]
4. Kargapolova, N. A., Khlebnikova, E. I., Ogorodnikov, V. A.: Numerical study of properties of air heat content indicators based on the stochastic model of the meteorological processes. Russ. J. Num. Anal. Math. Modelling 34(2), 95–104 (2019).
5. Steadman, R.G.: The Assessment of Sultriness, Part I: A Temperature-Humidity Index Based on Human Physiology and Clothing Science. J. Appl. Meteor. 18, 861–873 (1979).
6. Steadman, R.G.: A universal scale of apparent temperature. J. Climate Appl. Meteorol. 23, 1674–1687 (1984).
7. Anderson, G. B., Bell, M. L., Peng, R. D.: Methods to calculate the heat index as an exposure metric in environmental health research. Env. Health Perspect. 121(10), 1111–1119 (2013).
8. Schoen, C.: A new empirical model of the temperature– humidity index. J. Appl. Meteorol. 44, 1413–1420 (2005).
9. Kargapolova, N. A., Khlebnikova, E. I., Ogorodnikov, V. A.: Monte Carlo simulation of the joint non-Gaussian periodically correlated time-series of air temperature and relative humidity. Statistical papers 59, 1471–1481 (2018).
10. Kargapolova, N. A.: Stochastic Models of Non-stationary Time Series of the Average Daily Heat Index. In: Proc. of 9th Int. Conf. on Simulation and Modeling Methodologies, Technologies and Applications "SIMULTECH-2019". SCITEPRESS, Prague, 209–215 (2019).
11. Ogorodnikov, V.A., Prigarin, S.M.: Numerical Modelling of Random Processes and Fields: Algorithms and Applications. 1st edn. VSP, Utrecht (1996).
12. Ermakov, S. M., Mikhailov, G. A.: Statistical Modeling. 2nd edn. Nauka, Moscow (1982). [in Russian]
13. Encyclopedia Britanica, https://www.britannica.com/science/heat-wave-meteorology, last accessed 2019/05/18.

# Real time prediction of irregular periodic time series data

Chi TimNg
Chonnam National University (South Korea)

## Abstract

By means of a novel time-dependent cumulated variation penalty function, a new class of real-time prediction methods is developed to improve the prediction accuracy of time series exhibiting irregular periodic patterns, in particular, the breathing motion data of the patients during the robotic radiation therapy. It is illustrated that for both simulated and empirical data involving changes in mean, trend, and amplitude, the proposed methods outperform existing forecasting methods based on support vector machines and articial neural network in terms of prediction accuracy. Moreover, the proposed methods are designed so that real-time updates can be done efficiently with O(1) computational complexity upon the arrival of a new signal without scanning the old data repeatedly.

# New test for a random walk detection based on the arcsine law

Marcin Dudziński, Konrad Furmańczyk and Arkadiusz Orłowski

Faculty of Applied Informatics and Mathematics, Waraw University of Life Sciences, Poland
marcin_dudzinski@sggw.pl
konrad_furmanczyk@sggw.pl
arkadiusz_orlowski@sggw.pl

**Abstract.** In our work, we construct a new statistical test for a random walk detection, which is based on the arcsine laws. Additionally, we consider a version of the unit root test for an autoregressive process of order 1, which is also related to the arcsine laws. Additionally, we conduct some simulation study in order to check the quality of the proposed test.

**Keywords:** random walk, arcsine law, tests for a random walk detection

## 1 Introduction

Our objective is to introduce some proposal of a new test for a random walk detection. To the best of our knowledge, the main tools that have been applied in this context so far are the two celebrated tests – an Augmented Dickey-Fuller (ADF) test ([10]) and the Runs test ([12]) - and through our work, we attempt to fill in a gap related to this field of investigations. The presented approach is a certain extention and generalization of the research conducted in [2]. We also compare the quality of the proposed tests with the efficiency of some existing tests devoted to a random walk identification. The readers who are closely interested in the field of tests devoted to the random walk identification or the existence of unit root are encouraged to refer to [5]-[9] and [11].

Our paper is organized as follows. In Section 1, we present a general idea leading to the construction of our test for a random walk identification as well as, we describe the construction of this test. In Section 2, we check the efficiency of the introduced test, whereas in Section 3 we summarize our study.

### 1.1 Random Walk

Random walk theory states that the price of financial instrument in the subsequent time point is the sum of its price in the previous time point and some random variable with a finite variance, i.e. it is modeled with the use of a stochastic process called a random walk.

We say that a stochastic process $S_0, S_1, S_2, ..., S_n$ is a random walk if the following relations hold:

$$\begin{aligned} S_0 &= s_0, \\ S_1 &= s_0 + Y_1, \\ S_2 &= s_0 + Y_1 + Y_2, \\ &\vdots \\ S_n &= s_0 + Y_1 + Y_2 + ... + Y_n, \end{aligned}$$

where $Y_1, Y_2, ..., Y_n$ form an iid sequence of symmetric r.v.'s.

In our considerations, we assume that $s_0 = 0$. Then, $\mathrm{S}_t = \sum_{i=1}^{t} Y_i,\ t = 1, 2, ..., n$.

### 1.2 Ordinary Random Walk test

Let:

$$\Pi_n = \left|1 \leq i \leq n : S_i > 0\right|. \tag{1}$$

Therefore: $\Pi_n$ - the number of those among the sums $S_1, ..., S_n$, which are positive, $\Pi_n/n$ - its frequency.

From the first arcsine law ([3], [10]), we have:

$$lim_{n\to\infty} P\left(\Pi_n < nx\right) = \int_0^x \frac{1}{\pi\sqrt{x\left(1-x\right)}} dx = \frac{2}{\pi} \arcsin\left(\sqrt{x}\right), \tag{2}$$

for all $x \in (0; 1)$,

Conclusion above may practically be used for $n \geq 20$, which means that:

$$P\left(\frac{\Pi_n}{n} \leq x\right) \approx \frac{2}{\pi} \arcsin\left(\sqrt{x}\right) \ \ for\ n \geq 20, \tag{3}$$

where obviously:

$$\frac{\Pi_n}{n} = \frac{\left|1 \leq i \leq n : S_i > 0\right|}{n}. \tag{4}$$

From Fig. 1 - depicting the density of the arcsine distribution - we observe that the values of $\Pi_n/n$ in the close neighbourhood of 0.5 are the least probable and the most probable values for $\Pi_n/n$ are close to 0 or 1 ([3]).

Thus, if we denote by $\alpha$ the significance level of the test $H_0$: $S_n$ is a random walk processes, we look for a critical area (a set of rejections) of the form:

$$K_{c(\alpha)} = \left(0.5 - c\left(\alpha\right); 0.5 + c\left(\alpha\right)\right), \tag{5}$$

where $0 < c\left(\alpha\right) < 0.5$ satisfies the condition:

$$\int_{0.5-c(\alpha)}^{0.5+c(\alpha)} \frac{1}{\pi\sqrt{x\left(1-x\right)}} dx = \left.\frac{2}{\pi} \arcsin\left(\sqrt{x}\right)\right|_{0.5-c(\alpha)}^{0.5+c(\alpha)} = \alpha. \tag{6}$$

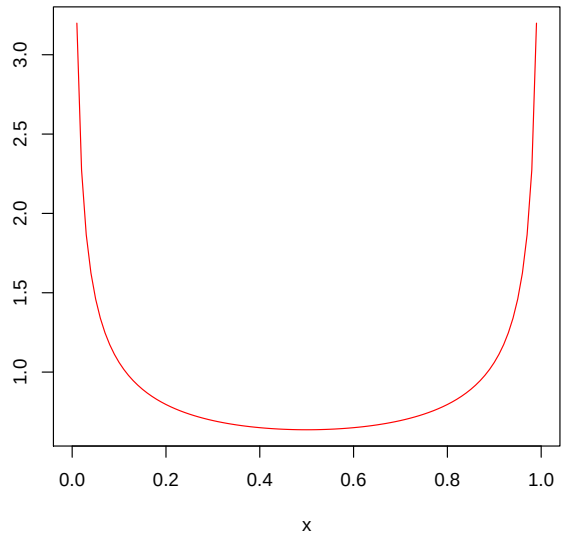



**Fig. 1.** Denisty of the arcsine distribution

Hence, for $0 < c(\alpha) < 0.5$:

$$\arcsin\left(\sqrt{0.5 + c(\alpha)}\right) - \arcsin\left(\sqrt{0.5 - c(\alpha)}\right) = \frac{\pi\alpha}{2}. \tag{7}$$

The values of $c(\alpha)$, calculated numerically for the chosen significance levels according to last formula, are collected in the following table:

**Table 1.** Values of $c(\alpha)$

| $\alpha$ | 0.01 | 0.05 | 0.1 |
|---|---|---|---|
| $c(\alpha)$ | 0.008 | 0.039 | 0.078 |

For $\alpha = 0.05$, we obtain $c(\alpha) = c(0.05) = 0.039$ and the corresponding critical area is $K_{c(0.05)} = (0.5 - 0.039; 0.5 + 0.039) = (0.461; 0.539)$.

### 1.3 Random Walk test for AR(1) process

Recall that AR(1) process is defined as follows $X_n = \rho X_{n-1} + \varepsilon_n$, where $(\varepsilon_n)$ stands for the white noise with a mean zero and variance $\sigma^2$. Observe that

assuming the starting point $x_0 = 0$, we have:

$$\begin{aligned}
&X_1 = \varepsilon_1,\\
&X_2 = \rho X_1 + \varepsilon_2 = \rho\varepsilon_1 + \varepsilon_2,\\
&X_3 = \rho X_2 + \varepsilon_3 = \rho^2\varepsilon_1 + \rho\varepsilon_2 + \varepsilon_3,\\
&\quad\vdots\\
&X_n = \rho X_{n-1} + \varepsilon_n = \rho^{n-1}\varepsilon_1 + \rho^{n-2}\varepsilon_2 + \rho^{n-3}\varepsilon_3 + ... + \rho\varepsilon_{n-1} + \varepsilon_n.
\end{aligned}$$

Therefore, the hypothesis $H_0$ is equivalent to the hypothesis that $\rho = 1$ (in this case $(X_n)$ is a RW process, since then $X_n = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + ... + \varepsilon_{n-1} + \varepsilon_n$).

## 2 Efficiency and power evaluation of the proposed test

### 2.1 Gaussian random walk

The efficiency of our test has firstly been checked for a Gaussian random walk, i.e. in the case when $Y_i \sim N(0;1)$, for $M = 1000$ generations of samples of the size $n = 1000$ or $n = 2000$:

$$\begin{aligned}
&\left(y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\right),\\
&\left(y_1^{(2)}, y_2^{(2)}, ..., y_n^{(2)}\right),\\
&\vdots\\
&\left(y_1^{(1000)}, y_2^{(1000)}, ..., y_n^{(1000)}\right).
\end{aligned}$$

For every sample above, we calculated the values of the test statistic $\Pi_n/n$:

$$\begin{aligned}
(\Pi_n/n)_{emp}^{(1)} &= \frac{\left|1 \le i \le n : y_1^{(1)} + y_2^{(1)} + ... + y_i^{(1)} > 0\right|}{n},\\
(\Pi_n/n)_{emp}^{(2)} &= \frac{\left|1 \le i \le n : y_1^{(2)} + y_2^{(2)} + ... + y_i^{(2)} > 0\right|}{n},\\
&\vdots\\
(\Pi_n/n)_{emp}^{(1000)} &= \frac{\left|1 \le i \le n : y_1^{(1000)} + y_2^{(1000)} + ... + y_i^{(1000)} > 0\right|}{n}.
\end{aligned}$$

We calculated the number of those among $\Pi_n/n$, which belonged to the critical area $K_{c(0.05)} = (0.461; 0.539)$, i.e. the number of rejections of $H_0$. We repeated this procedure 6 times and obtained the following numbers of rejections (out of 1000 possible rejections): 53, 44, 54, 33, 55, 50 (if n=1000) or 44, 48, 43, 45, 50, 56 (if n=2000). The small numbers of rejections may give an evidence about a good efficiency of the proposed test.

### 2.2 AR(1) process

We have checked the power of our test by generating $M = 1000$ samples of the size $n = 1000$ or $n = 2000$ of AR(1) processes with $\sigma = 3$:

$$\begin{aligned}
&\left(x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}\right), \\
&\left(x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)}\right), \\
&\vdots \\
&\left(x_1^{(1000)}, x_2^{(1000)}, ..., x_n^{(1000)}\right).
\end{aligned}$$

We calculated the values of the test statistic $\Pi_n/n$:

$$\begin{aligned}
(\Pi_n/n)_{emp}^{(1)} &= \frac{\left|1 \le i \le n : x_1^{(1)} + x_2^{(1)} + ... + x_i^{(1)} > 0\right|}{n}, \\
(\Pi_n/n)_{emp}^{(2)} &= \frac{\left|1 \le i \le n : x_1^{(2)} + x_2^{(2)} + ... + x_i^{(2)} > 0\right|}{n}, \\
&\vdots \\
(\Pi_n/n)_{emp}^{(1000)} &= \frac{\left|1 \le i \le n : x_1^{(1000)} + x_2^{(1000)} + ... + x_i^{(1000)} > 0\right|}{n}.
\end{aligned}$$

As previously, we calculated the numbers of those among $\Pi_n/n$, which belonged to the critical area $K_{c(0.05)} = (0.461; 0.539)$. We obtained the following results (the numbers of rejections of $H_0$ among 1000 realizations) for the chosen values of $\rho$, after the 3 repetitions of the described procedure:

**Table 2.** Numbers of rejections ($n = 1000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | 172, 170, 154 |
| 0.8 | 682, 659, 667 |
| 0.6 | 844, 841, 839 |
| 0.4 | 934, 935, 925 |
| 0.2 | 968, 953, 954 |

**Table 3.** Numbers of rejections ($n = 2000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | 232, 246, 246 |
| 0.8 | 845, 845, 813 |
| 0.6 | 951, 954, 958 |
| 0.4 | 988, 985, 985 |
| 0.2 | 997, 997, 999 |

The results are quite promising - the larger $\rho$, the smaller number of rejections of $H_0$. Next, we compare the test procedure from this subsection with the ADF and Runs tests for randomness for binary data series (we put $+1$ if the first difference is positive and $-1$ if otherwise).

**Comparison with the ADF test.** For the chosen values of $\rho$, we obtained the following numbers of rejections of $H_0$ among 1000 realizations, after the 3 repetitions of the described procedure:

**Table 4.** Numbers of rejections ($n = 1000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | 179, 148, 156 |
| 0.8 | 1000, 1000, 1000 |
| 0.6 | 1000, 1000, 1000 |
| 0.4 | 1000, 1000, 1000 |
| 0.2 | 1000, 1000, 1000 |

**Table 5.** Numbers of rejections ($n = 2000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | 546, 533, 552 |
| 0.8 | 1000, 1000, 1000 |
| 0.6 | 1000, 1000, 1000 |
| 0.4 | 1000, 1000, 1000 |
| 0.2 | 1000, 1000, 1000 |

We observe that our test has a lower power than the ADF test for $\rho = 0.8$ and $\rho = 0.6$. For the remaining cases the powers of our test and tha ADF test are comparable.

**Comparison with the Runs test.** For the chosen values of $\rho$, we obtained the following numbers of rejections of $H_0$ among 1000 realizations, after the 3 repetitions of the described procedure:

**Table 6.** Numbers of rejections ($n = 1000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | $50, 55, 53$ |
| 0.8 | $503, 482, 502$ |
| 0.6 | $991, 989, 992$ |
| 0.4 | $1000, 1000, 1000$ |
| 0.2 | $1000, 1000, 1000$ |

**Table 7.** Numbers of rejections ($n = 2000$, 3 replications)

| $\rho$ | |
|---|---|
| 0.99 | $58, 61, 44$ |
| 0.8 | $798, 840, 827$ |
| 0.6 | $1000, 1000, 1000$ |
| 0.4 | $1000, 1000, 1000$ |
| 0.2 | $1000, 1000, 1000$ |

We may see that our test has a lower power than the Runs test for $\rho = 0.6$. However, for $\rho = 0.99$ and $\rho = 0.8$ our test has a better power than the Runs test. For the remaining cases the powers of our test and the Runs test are comparable.

## 3 Conclusions

The principal goal of our study was to construct a new test for a random walk detection. The main idea of our approch was based on the first arcsine law. Appart from the construction of a new test, we examined its efficiency using 1000 replications of the Monte Carlo simulation by computing the numbers of rejections of the null hypothesis that the given process forms a random walk. The obtained results and comparisons indicate that the introduced test provides quite an effective tool leading to a random walk identification.

## References

1. Dickey, D. A., Fuller, W.A.: Distributions of the estimators for autoregressive time series with a unit root. Jour. Amer. Stat. Assoc. 74 , 427–431 (1979)

2. Dudziński, M., Furmańczyk, K., Orłowski, A.: Some proposal of the test for a random walk detection and its application in the stock market data analysis. Quantitative Methods in Economics 19, 339–346 (2018)
3. Feller, W.: An introduction to probability theory and its applications. Wiley, New York (1968)
4. Maddala, G. S.: Introduction to Econometrics. Morgan Kaufmann, Wiley, New York (2001)
5. Mankiw, N. G., Shapiro, M. D.: Trends, random walks, and tests of the permanent income hypothesis. J. Monet. Econ. 16(2), 165-174 (1985)
6. Pantula, S. G., Farias-Gonzales, G., Fuller, W. A.: A comparison of unit-root test criteria. J. Bus. Econ. Stat. 12, 449-459 (1994)
7. Perron, P., Shiller, R. J.: Testing the Random Walk Hypothesis: Power Versus Frequency of Observation. Econ. Lett. 18, 381-386 (1985)
8. Phillips, P. C. B.: Time series regression with a unit root. Econometrica 55, 277-301 (1987)
9. Phillips, P. C. B., Perron, P.: Testing for a unit root in time series regression. Biometrika 75, 335-346 (1988)
10. Qiang, L., Jiajin, L.: Arcsine laws and its simulation and application. http://individual.utoronto.ca/normand/Documents/MATH5501/Project-3/Arcsine_laws_and_simu.pdf
11. Said, E. S., Dickey, D. A.: Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71, 599-607 (1984)
12. Siegel, S., Castellan, N. J.: Nonparametric statistics for the behavioural sciences. McGraw-Hill, New York (1988).

# Analysis of non-stationary time series based on modelling stochastic dynamics considering self-organization, memory and oscillations

D. Zhukov[1], T. Khvatova[2,3], and L. Istratov[1]

[1]Russian Technological University, Russia
[2] the Great St. Petersburg Polytechnic University, Russia
[3] Lappeenranta University of Technology, Finland

zhukov_do@mirea.ru
khvatova.ty@spbstu.ru
istratov@mirea.ru

**Keywords**: non-stationary time series; stock indexes; distribution function of deviation amplitudes; stochastic dynamics; self-organization; memory; asymmetry of distribution function; oscillations; probability density oscillations

**Abstract.** Behaviors of stock and raw material market indexes are most often described using non-stationary time series. The processes observed therein show the potential for self-organization and the presence of memory of previous events having occurred within the system, both conditioned by the human factor. The present authors argue that stock exchange processes cannot be fully explained using the theory of chaos, and this should be considered when developing forecasting models. Consequently, this paper proposes a newly developed model of stochastic dynamics for forecasting stock indexes considering self-organization processes, the presence of memory and possible oscillations. To develop this model, probabilistic schemes of transition from possible states of the process (i.e. changes in stock and raw material indexes) were used, considering several previous points in time which enables the study of memory. According to the proposed approach, regarding the probability density of transitions over time, a second order non-linear differential equation was obtained. The equation contains first and second derivatives regarding time and the variable describing the change in indexes, enabling the study of self-organization processes. The results obtained are significantly different to those generated by models widely used today to describe the evolution of non-stationary distributions, i.e. models based on the theory of chaos, diffusion approaches, Liouville and Fokker-Planck equations, neural network models and etc. In order to analyze and benchmark the present model with the observed data, Dow Jones and Hang Seng index dynamics were studied and probability density histograms for the amplitudes of their deviations depending on the time of calculation, were plotted. The histograms are asymmetric with regards their maximum; the shift in time and oscillations of the probability density were observed. The research demonstrated that the present model closely fits the observed data regarding

[1] Russian Technological University (MIREA), Moscow, Russia
[2] Peter the Great St. Petersburg Polytechnic University (SPbPU), St. Petersburg, Russia
[3] Lappeenranta University of Technology (LUT), Lappeenranta, Finland

behaviors of stock index amplitudes, according to the time interval of their calculation. Depending on the parameters of the model and their relationships with one another, it is possible to describe various dynamics of stock index amplitude behaviors. The model considers the presence of asymmetry of distribution functions relative to the maximum; oscillations; the shift of the maximum value of the amplitude distribution function, depending on the time of their calculation; changes in height and width of the distribution when the amplitude calculation time changes, etc.

## 1 Introduction

Experience has shown that using traditional time series analysis approaches to modeling stock index dynamics often engenders serious errors, which can be explained by the substantial variability of their characteristics. Therefore, it is important to search for new methods with which to analyze their dynamics. The following possible avenues of research exist: using neural networks [1-4], applying fuzzy logics approaches [5,6], non-parametric models based on the theory of chaos and the method of support vector machines [7], using rule sets based on genetic algorithms [8,9], self-organizing adaptive models [10], statistical models based on using sample distribution functions [12-15], and many more.

To create models based on any given approach, information about the time series dynamics is essential. This information can be presented either explicitly (for example, as a distribution function in a statistical model) or indirectly (as behavioral patterns in neural network models). The key problem in analyzing and modeling a time series behavior is that at each moment of time there is only one process realization, upon which the forecast for the next moments of time are created.

In the existing methods of analysis, no matter which instrument is used (statistical, neural network, fuzzy logic models, etc.), the time series is split into separate parts in which it is quasi-stationary and has its own sample distribution function (SDF). Between each pair of the parts there is a fraction of the time series in which the transition process (misbalance) occurs. The length of the transition process is defined by the factors characterizing the change of the modus (i.e. misbalance occurring) and also by the size of the sample used for conducting statistical analysis [11]. The parameters of the SDF can be fixed based on analysis of the data observed during the time interval of quasi-stationarity.

If the SDF is non-stationary, then the information provided by the recognition algorithm based on past data is often inadequate. There is only one trajectory which, due to non-stationarity, does not enable the use of a large-volume sample for testing various indicators of local behaviors of the time series to assess the probability of the correct functioning of these indicators when choosing stock exchange strategy.

Options can be used as functionals: reaching a certain level of profitability can be used as a stock exchange indicator (if such criterion is fulfilled, the sale or purchase of an asset is performed). Option contracts (call-options (C(t)) and put-options (P(t))) are connected by the distribution function $\rho(x,t)$, as follows:

$C(t) = \int_{x_s}^{+\infty}(x - x_s)\cdot\rho(x,t)dx$ и $P(t) = \int_{0}^{x_s}(x_s - x)\cdot\rho(x,t)dx$,

where $x_s$ is the strike price, t is the expiry date, x is the current price, $\rho(x,t)$ – the probability density of reaching the price x at the moment of time t; $\rho(x,t)$ can be represented using a reconstructed sample distribution function (SDF) on an observed sample of data.

In most widely applied models used for analyzing and forecasting the dynamics of non-stationary time-series (their evolution), the diffusion equations are used as approximations of sample distributions. Their equations include: a) non-linear diffusion (Fuentes, 2018): $\frac{\partial \rho(x,t)}{\partial t} = D(t) \frac{\partial^2 \rho^{\frac{n-1}{n+1}}(x,t)}{\partial x^2}$, where n is a numeric parameter of the model, and $\frac{n-1}{n+1}$ is the indicator of the degree of probability density distribution function $\rho(x,t)$ (this equation accounts only for random transitions); b) the Liouville equation defines the ordered shift [12]; c) the Fokker-Planck equation [12] accounts not only for random change, but also for ordered non-stationary transitions or "non-stationary drift"; and other equations.

However, none of these models provide an underpinning for the structure and form of the equations used; they do not account for self-organization and the presence of memory.

The goal of the present research is thus to analyze stock index dynamics and, based on this, to find theoretical distribution functions for the observed data - accounting for self-organization and the presence of memory.

## 2 Selecting data for analyzing the dynamics of stock market index changes and discovering the dependencies of statistical characteristics of distributions on the time interval within which they are calculated.

In order to analyze the dynamics of stock index changes two types of markets were considered herein: developing markets and mature developed markets. This is related to the general character of the processes which occur in the stock markets. Developed markets are more stable: when fluctuations occur on these markets, their dynamics are often reflected in developing markets. If conditions are unfavorable, investors may leave the developing market, but this may not have a strong effect on mature developed markets.

In this paper only two markets are explored: Dow Jones (developed markets) and Hang Seng (developing markets). To analyze the dynamics of the Dow Jones and Hang Seng indexes, the time interval between 1st of January 2018 and 1st of January 2019 was selected. All data was extracted with an interval of one minute.

To process the data obtained and define the probability density functions of the stock index oscillation amplitudes (i.e. changes in prices during the given time interval), the following algorithm was used:

1) For every minute, the values of the researched stock index are selected for a certain time interval (a day, a week, a month, etc.).

2) The amplitudes of the stock index changes over various periods of time (one minute, two minutes, etc.) are calculated.

3) The amplitudes calculated for each of the given time intervals are arranged in ascending order (from negative to positive), and for each of the amplitude intervals, the histograms for the probability density of amplitude distributions are constructed depending on the length of the time interval of their calculations.

4) Based on the histograms obtained for each time interval of amplitude calculations, the momentums of distributions are calculated (i.e. mean value – mathematical expectation, variance, etc.).

5) The dependencies of the stock index distribution momentum oscillation amplitudes on the time intervals of their calculations are presented graphically (Fig.1).

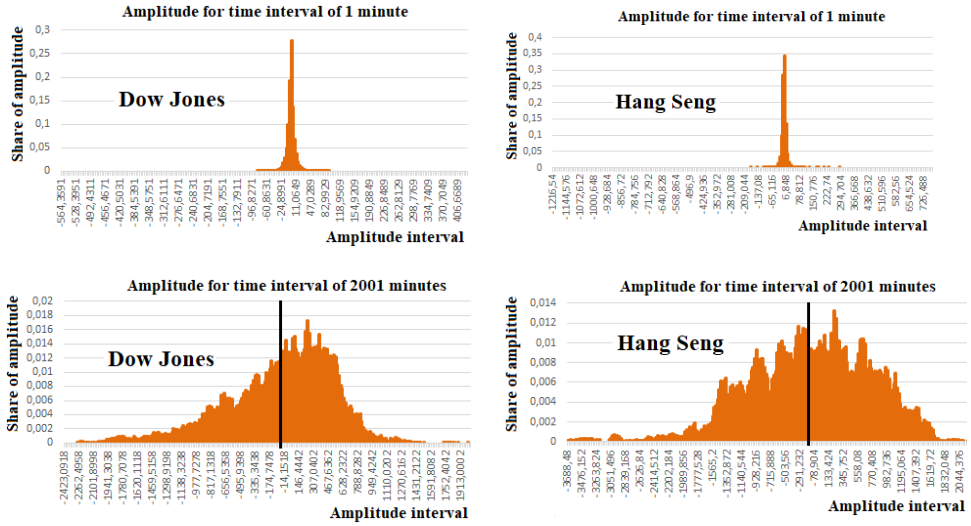



**Fig. 1.** *Histograms for dependencies of probability density for various deviation amplitudes for Dow Jones and Hang Seng stock indexes on the time interval of their calculation (for data over the period of 1st January 2018 to 1st January 2019).*

The data analysis (see Fig.1) demonstrates that for small time intervals, the histograms of stock indexes have a large central peak which is close to a value of zero. With probabilities of 0.30-0.35, small amplitudes are observed, while large amplitudes occur with small probabilities. When the time interval, for which the amplitudes are calculated, increases, the central peak decreases, the distribution width increases, the distribution symmetry is distorted, oscillations appear in the distribution spectrum (this is especially distinctive for the Hang Seng index, calculated for large time intervals). Furthermore, when the time interval for calculating oscillation amplitudes increases, all distribution moves towards positive amplitudes (to the right) and becomes asymmetric, i.e. the "tail" of the distribution from the left is plotted lower than on the right.

# 3 The stochastic dynamics model for shaping stock indexes considering self-organization processes, memory and oscillations

## 3.1 Developing probabilistic difference schemas for state-to-state transitions and deriving key equations for the model

The whole set of deviation amplitudes for stock indexes for any time interval t (a rather large interval), can be called X. We can furthermore consider that the time interval t consists of small parts $\tau$. In this case, any time interval t can be expressed as $t_h=h\,\tau$, where h – is the number of steps $\tau$ (h=0,1,2,3, …N). The value of h can be interpreted as discrete time, a unit of which equals $\tau$. The amplitude for the chosen t can be called $x_h$ ($x_h \in X$). Analysis of the observed amplitudes, represented in Fig.1, shows that $x_h$ can have positive or negative values. The probability of observing large values is considerably smaller than the probability of observing smaller values.

Let us suppose that the amplitude $x_{h,}$ after changing the discrete time h by 1, can increase by some small value $\varepsilon$, or decrease by some small value $\xi$ (in general $\varepsilon \neq \xi$ can be a rising trend or a falling trend). This enables us to represent the amplitude $x_h$ for the interval for discrete time h as sets of various $\xi$ and $\varepsilon$ signs should be considered). Furthermore, let us suppose that no $x_h$ can be permanent, i.e. it must change when the time h changes by 1. This follows on from the fact that if h changes, $x_h$ must change by $\xi$ or $\varepsilon$.

Let us find the probability P(x,h) that the deviations amplitude of stock indexes for a certain time interval h is equal to x, as follows: P(x–ε,h–1) – probability that for a certain (h–1) the amplitude is (x–ε); P(x+ξ,h–1) – probability that for a certain (h–1) the amplitude is (x+ξ); P(x,h–1) – probability that for a certain (h–1) the amplitude is equal to x.

The probability P(x,h) that the deviations amplitude for stock indexes for a discrete time interval h will be equal to x can be defined as follows :

$$P(x, h)=P(x-\varepsilon, h-1)+P(x+\xi, h-1)-P(x, h-1) \quad (1)$$

The expression (1) can be explained as follows: the probability P(x, h) that the amplitude is equal to x for the interval of discrete time h is defined by the sum of the probabilities that at (h-1) the amplitude was equal to (x-ε) (the member of equation P(x–ε, h–1)) and (x+ξ) (member of the equation P(x+ε, h–1)). Consequently, it should be noted that if h changes by 1 (if the amplitude was equal to x) then, due to non-stationarity, it will be transferred to any other state (member of the equation P(x,h-1)).

The equation (1) can be expressed as follows:

$$P(x, h+1)=P(x-\varepsilon, h)+P(x+\xi, h)-P(x, h) \quad (2)$$

To account for memory, let us define the probabilities $P(x-\varepsilon,h)$, $P(x+\xi,h)$ and $P(x,h)$ via the states on step h-1. Considering that ε and ξ are certain permanent values, the following algebraic equation for the probability of transition can be obtained:

$$P(x,h+2)=\{P(x-2\varepsilon,h)+P(x-\varepsilon+\xi,h)-P(x-\varepsilon,h)\}+\{P(x+\xi-\varepsilon,h)+ +P(x+\xi,h)-P(x+\xi,h)\}-P(x-\varepsilon,h)-P(x+\xi,h-1)+P(x,h) \quad (3)$$

Moreover, considering that $t=h\cdot\tau$, where $t$ – the time of the process, $h$ – number of the step, $\tau$ – the length of one step, we can go from h to t and perform the Taylor expansion:

$$\frac{dP(x,t)}{dt} = a\frac{d^2P(x,t)}{dx^2} - b\frac{dP(x,t)}{dx} - c\frac{d^2P(x,t)}{dt^2} \quad (4)$$

where: $a = \frac{\varepsilon^2 - \varepsilon\xi + \xi^2}{\tau};\ b = \frac{\varepsilon - \xi}{\tau};\ c = \tau$

The member of equation $\frac{\mathrm{dP(x,t)}}{\mathrm{dx}}$ describes the ordered transition either into a state where the described value is increasing ($\varepsilon > \xi$), or when it is decreasing ($\varepsilon < \xi$); the member of equation $\frac{d^2P(x,t)}{dx^2}$ describes the random change in state (*uncertainty of change*). The member of equation $\frac{dP(x,t)}{dt}$– can be identified as the speed of general change of state of the system over the course of time; the member of equation $\frac{d^2P(x,t)}{dt^2}$ describes the process wherein the states become sources of other emerging states (self-organization and acceleration of ordered ($\frac{dP(x,t)}{dx}$) and random ($\frac{d^2P(x,t)}{dx^2}$) transitions).

### 3.2 Formulating and solving the boundary value problem for the model describing stock index dynamics considering oscillations, self-organization and memory.

If we presume that $P(x,t)$ is continuous, then it is possible to shift from probability $P(x,t)$ to probability density $\rho(x,t)=dP(x,t)/dx$ and to formulate the boundary-value problem for defining the dependency between the probability density of observing deviation amplitudes for stock indexes within a certain time interval t.

The boundary conditions for possible deviation amplitudes of stock indexes will be selected according to the following propositions. Statistical data analysis shows that the probabilities of observing large deviation amplitudes of stock indexes (several percentage points increasing or decreasing) are very small for the time intervals considered. This is why it can be supposed that the probability density function for amplitudes should diminish quickly and become zero for large values. Thus, the boundary conditions can be formulated as follows:

$$\rho(x,t)_{x=\infty} = 0 \quad (a)$$

$$\rho(x,t)_{x=-\infty} = 0 \quad (b)$$

The first boundary condition will be set as delta function because for the time interval t=0, only the amplitude $x_o$=0 is possible:

$$\rho(x,t)|_{t=0} = \delta(x-0) = \begin{cases} \int \delta(x-0)dx = 1, x = 0 \\ 0, \qquad\qquad x \neq 0 \end{cases} \quad (c)$$

The second boundary condition is also required: $\left.\frac{\partial\rho(x,t)}{\partial t}\right|_{t=0}$, i.e. this is the condition which sets the speed of change in the probability density for any amplitude. On the stock market, many brokers act using various investment strategies and over various periods of time. Each of these strategies defines a certain amplitude of change within an index. The actions of brokers at the moment of the process countdown t=0 cause various speeds of change within indexes. The accumulation of various strategies may lead to certain amplitudes increasing and others decreasing. In the end this can lead to periodicity in some amplitudes, i.e. to the appearance of waves described by a periodic function. The value $\frac{\partial\rho(x,t)}{\partial t}$ can be expressed as follows:

$\left.\frac{\partial\rho(x,t)}{\partial t}\right|_{t=0} = \lim_{\Delta t\to\tau}\left.\frac{\rho(x+\Delta x,t+\Delta t)-\rho(x,t)}{\Delta t}\right|_{t=0} = \frac{\rho(x+\Delta x,0+\tau)-\rho(x,0)}{\tau} = \frac{\rho(x+\Delta x,0+\tau)-\delta(x-0)}{\tau} =$
$= \frac{1}{\tau}\cdot\psi(x)\cdot\delta(x-y)$

wherein $\psi(x)$ is a periodic function which can be defined by analyzing the observed data and verified using the model. The presence of the function $\delta(x-y)$ is due to the fact that the numerator contains $\delta(x-0)$, at the same time, it takes into account the fact that for each x its own $\psi(x)$ will exist (y means the current value of x).

Solving the boundary value problem for equation (4) using boundary and initial conditions generated the following dependency of probability density function of stock index deviation amplitudes from the time interval of their calculations:

$$\rho(x,t) = \frac{\tau\cdot\mathcal{U}(t-k)\cdot e^{\frac{(\varepsilon-\xi)x}{2(\varepsilon^2-\varepsilon\xi+\xi^2)}}\cdot e^{-\frac{t}{2\tau}}}{2\sqrt{\varepsilon^2-\varepsilon\xi+\xi^2}}\left\{\begin{array}{l}\frac{1}{\tau}\left\{\frac{1}{2}+\psi(x)\right\}\sum_{n=0}^{\infty}\frac{\left\{\omega^2\left\{t^2-k^2\right\}\right\}^n}{4^n\cdot(n!)^2}+\\ +\frac{t}{t^2-k^2}\sum_{n=0}^{\infty}\frac{2n\cdot\left\{\omega^2\left\{t^2-k^2\right\}\right\}^n}{4^n\cdot(n!)^2}\end{array}\right\} \quad (1)$$

where $\omega = \sqrt{\frac{\varepsilon\xi}{4\tau^2(\varepsilon^2-\varepsilon\xi+\xi^2)}}$; $k = \frac{|x|\tau}{\sqrt{\varepsilon^2-\varepsilon\xi+\xi^2}}$; $\psi(x) = \begin{cases}cos\left\{2\pi\frac{x}{\sqrt{\varepsilon^2-\varepsilon\xi+\xi^2}}\right\}\\ sin\left\{2\pi\frac{x}{\sqrt{\varepsilon^2-\varepsilon\xi+\xi^2}}\right\}\end{cases}$ are periodic functions; $\mathcal{U}(t-k)$ is the Heaviside step function: $\mathcal{U}(t-k) = \begin{cases}0, & if\ t<k\\ \frac{1}{2}, & if\ t=k\\ 1, & if\ t>k\end{cases}$

If there are no oscillations, $\psi(x)=0$. For the function $\rho(x,t)$, the condition of standardization is fulfilled: $\int_{-\infty}^{+\infty}\rho(x,t)dx = 1$.

### 3.3 Analysis of the model describing stochastic dynamics of stock indexes considering oscillations, self-organization and memory

Fig.2 presents the dependency of probability density of stock index oscillation amplitudes on the time of their calculation, obtained using equation (5) with various sets of parameters ξ; ε and τ, for various times t. Curve 1 is for t=1 conditional unit, curves 2 is for t=3 conditional units and curve 3 is for 7 conditional units.

Analysis of the obtained theoretical model (equation 5) demonstrates that if the interval of the amplitude calculation grows, the maximum of the distribution density shifts towards the positive amplitudes, to the right (if ξ<ε was chosen). If ξ>ε was chosen however, then the shift will be to the left. The height of the peaks decreases, the width of the distribution increases. Moreover, a larger asymmetry and oscillations of the distribution are observed. When ξ=ε, the maximum of the distribution will correspond to zero: it will become symmetrical. If the time interval increases, the position of the maximum does not change, the height decreases, the width increases, oscillations remain. If the parameter $\lambda=\varepsilon^2-\varepsilon\xi+\xi^2$, the number of oscillations in the graph for probability density of stock index oscillation amplitudes increases, the height of the distribution decreases; and vice versa, the height of the distribution increases.

Theoretical value of mathematical expectation and variance for the stock index amplitude deviations are calculated as follows:

$\mu(t) = \int_{-\infty}^{+\infty} x \cdot \rho(x,t)dx$ и $\sigma^2(t) = \int_{-\infty}^{+\infty} x^2 \cdot \rho(x,t)dx$

As the function $\rho(x,t)$ decreases fast, while calculating $\mu(t)$ and $\sigma^2(t)$, it is possible to integrate this with respect to the limited area.

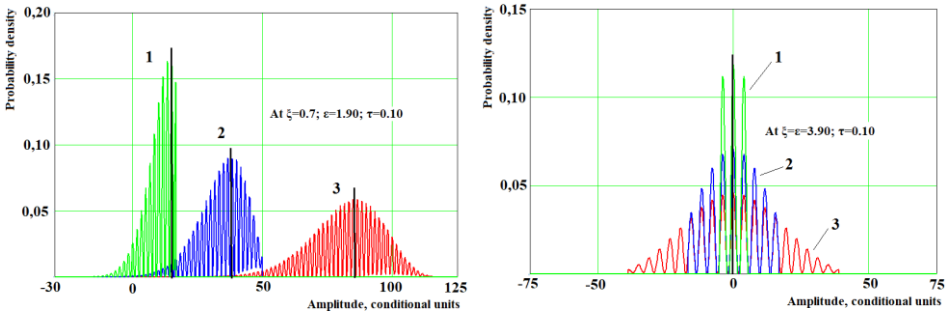



**Fig. 2.** *Dependencies of probability density of amplitudes on the time of their calculations for the model of stock index behavior considering self-organization and memory.*

Fig. 3 presents the dependency of mathematical expectation for amplitudes depending on the time of their calculations obtained using the present model (for various ε and ξ). Calculations show that if ε>ξ, then amplitudes of growth are observed, and their values are located in the positive area in Fig 3 (lines under number 2). If ε<ξ, decreasing amplitudes are observed, and their mathematical expectations are located in the negative area (lines above number 1).

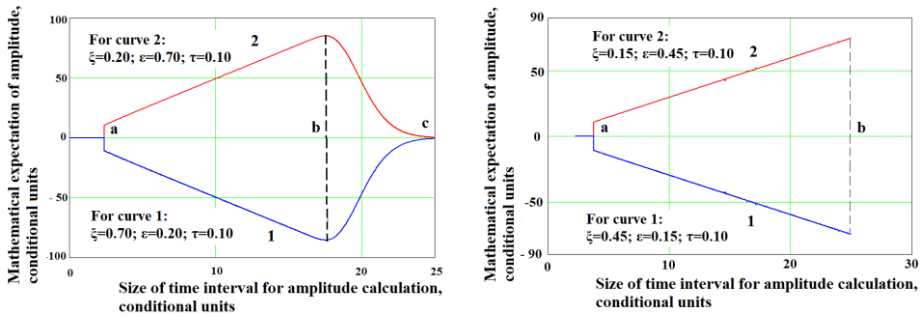



**Fig. 3.** *Dependencies of mathematical expectation for amplitudes depending on the time of their calculations for the model of stock index behavior considering self-organization and memory.*

When the time intervals are large, variance and mathematical expectation decrease to zero (Fig.4). It is important to note that for sets of parameters with inverse values of ε and ξ (they take on the values of one another) the variance behaves in the same way because it is a quadratic value and as such has no negative values.

The developed model is fundamentally different from models widely used today for describing the evolution of non-stationary distributions based on the theory of chaos, diffusion approaches, and the Liouville and the Fokker-Planck equations. Indeed, the proposed model shows that for large intervals of amplitudes calculation, their

mathematical expectation and variance become zero. This result cannot be obtained with the model describing these dynamics using the Fokker-Planck equation, which gives unlimited linear growth depending on the time; or using the diffusion model, in which the mathematical expectation either equals zero or is constant.

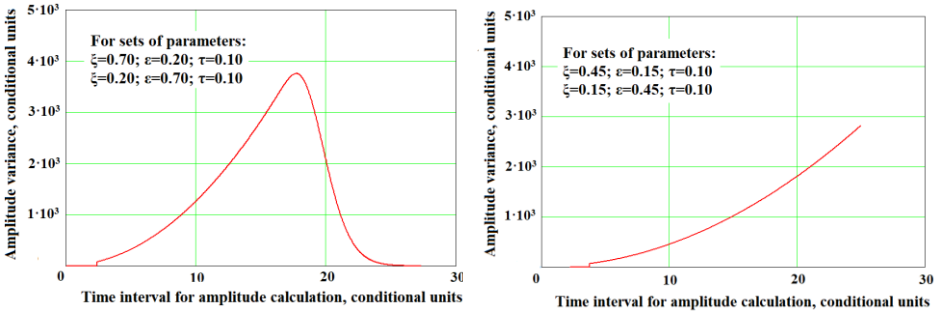



**Fig. 4.** *Dependencies of variance for amplitudes depending on the time of their calculations for the model considering self-organization and memory.*

An important feature of the developed model is that at a given interrelation of the parameters ε and ξ there is no mathematical expectation for deviations for some time intervals (see righthand side of Fig.3, the part of interval [OA] and the area to the right of point b). For very large and very small time intervals, the market in this model is completely stochastic. In the market model using the Fokker-Planck equation, when time intervals are large, an unlimited quadratic growth of the variance of stock index deviations is observed depending on the time of their calculation, which is the drawback.

While analyzing non-stationary time-series, it is important to break them into separate "smooth" areas with certain process dynamics without sharp changes in indexes. Between these "smooth" areas, the specific areas of trend changes (misbalance) are located. To verify how well the developed model fits real data, "smooth" areas from the data for the period from 1st January 2018 to 1st January 2019 must be chosen for the Dow Jones and Hang Seng indexes, for example, two "smooth" areas; then the dynamics observed in these areas can be compared with the modeled dynamics. For the Dow Jones index one of the largest "smooth" areas is the period between 01.01.2018 - 01.02.2018, and for the Hang Seng index, it is the period between 01.06.2018 - 01.07.2018.

Fig. 5 presents the behavior of the observed dependencies of mathematical expectation and variance of the stock indexes deviation amplitudes on the time intervals of their calculation on "smooth" areas. Comparing the observed data (see Fig. 5) with the results of theoretical modeling (see Fig. 3 and 4) demonstrates a rather good fit (considering a certain degree of modeling approximation). Behaviors of the observed and modeled dependencies fit each other well, showing linear character, and variances also correspond, to some extent. This as a whole confirms the adequacy of the present model, which considers self-organization and the presence of memory. The obtained results can be used for developing investment strategies.

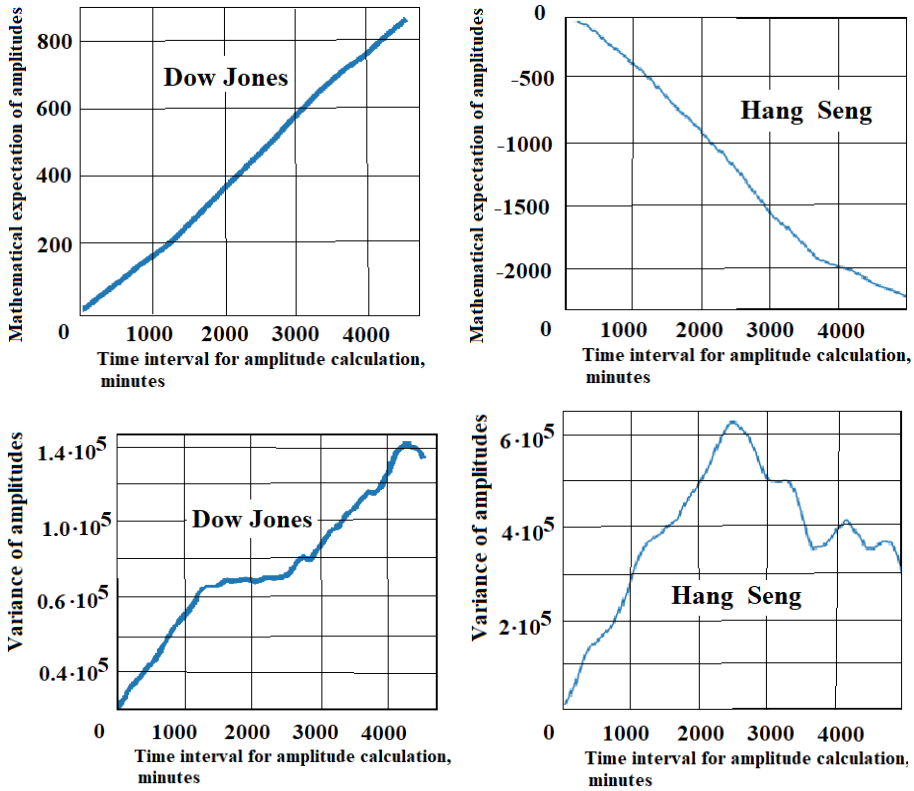



**Fig. 5.** *Dependencies of mathematical expectation and variance of Dow Jones and Hang Seng stock index amplitude deviations on the time interval of their calculation in the selected smooth areas.*

## 4 Stock exchange investment strategies

The general algorithm for developing an optimal investment strategy can consist of the following steps:

1) An array of observed stock indexes is selected. In this array, the point of the latest change of trend is identified (the area of misbalance)

2) Using the values from non-stationary time series starting from the moment of the last misbalance until the moment of the current observation, the histograms are developed. The histograms describe the dependencies of stock index amplitude deviations on the time of their calculation.

3) Using the obtained histograms and the equation (5) of the present model, we can find the parameters $\xi$; $\varepsilon$ and $\tau$ for the time interval from the last misbalance until the current moment.

4) Using the calculated parameters $\xi$; $\varepsilon$ and $\tau$ we can perform calculations of option contracts (call-options and put-options).

5) Furthermore, using, for example, the optimal profitability as the stock market indicator (i.e. the criterion which shows, if reached, that assets should be bought or sold) and the dependency of call-options C(t) or put-options P(t) on time, the stock exchange behavior can be planned.

## 5 Conclusions

In this research the dynamics of Dow Jones and Hand Seng indexes were analyzed. Based on the data obtained, the histograms of dependencies between the amplitudes of stock index deviations and the time intervals of their calculations were developed. The histograms of amplitude distributions are asymmetric in relation to the maximum value; oscillations of the probability density are observed.

A new model of stochastic dynamics of stock indexes was developed, which considers the processes of self-organization, the presence of memory and oscillations. This model describes the main characteristics of stock index behaviors. The model developed fits well with the observed data in terms of the deviations of stock index amplitudes depending on the time intervals of their calculations. The model considers a) asymmetry of the distribution functions in relation to the maximum value, b) opportunities for oscillations, c) changes in the height and width of the distribution while the time interval of the amplitude measurement varies, etc.

To develop the model, the probabilistic schemes of state-to-state transitions were considered. Based on this approach, a non-linear differential equation of the second order was derived. The boundary problem was formulated and resolved for defining the function of the probability density for the amplitude of stock index deviations depending on the time intervals of their measurements. The differential equation contains one member which is responsible for the opportunity for self-organization, and also accounts for the presence of memory. The developed model is essentially different from those models widely used today for describing the evolution of non-stationary distributions based on the theory of chaos, diffusion approaches, and the Liouville and Fokker-Planck equations. The proposed model shows that for large intervals of amplitude calculation, their mathematical expectation and variance attain zero, and at a certain combination of parameters $\varepsilon$ and $\xi$, the mathematical expectation for certain amplitudes does not exist. This result cannot be obtained within the framework of other models. The present model thus shows that for very large and very small time intervals the market is completely stochastic.

The developed stochastic model of stock indexes dynamics can be used for decision-making pertaining to investment.

**ACKNOWLEDGEMENTS**

This paper was financially supported by the Ministry of Education and Science of the Russian Federation as part of the program to improve the competitiveness of Peter the Great St. Petersburg Polytechnic University (SPbPU) among the world's leading research and education centers in the period of 2016-2020.

## 6 References


1. Khadjeh Nassirtoussi, A., Ying Wah, T., Ngo Chek Ling, D.: A novel FOREX prediction methodology based on fundamental data. African Journal of Business Management. **5**, 8322 – 8330 (2011). doi: 10.5897/AJBM11.798
2. Anastasakis, L., Mort, N.: Exchange rate forecasting using a combined parametric and non-parametric self – organising modelling approach. Expert Systems with Applications. **36**, 12001 – 12011 (2009). doi: https://doi.org/10.1016/j.eswa.2009.03.057
3. Vanstone, B., Finnie, G.: Enhancing stock market trading performance with ANNs. Expert Systems with Applications. **37**, 6602 – 6610 (2010). doi: 10.1016/j.eswa.2010.02.124
4. Vanstone, B., Finnie, G.: An empirical methodology for developing stockmarket trading systems using artificial neural networks. Expert Systems with Applications. **36**, 6668 – 6680 (2009). doi: 10.1016/j.eswa.2008.08.019
5. Sermpinis, G., Laws, J., Karathanasopoulos, A., Dunis, C. L.: Forecasting and trading the EUR/USD exchange rate with gene expression and psi sigma neural networks. Expert Systems with Applications **39**, 10, 8865 – 8877 (2012). doi: 10.1016/j.eswa.2012.02.022.
6. Drozhzhov, K., Ivanov, S.: The research and implementation of processing algorithm for a non-stationary signal with input sampled-data missing and intense impulse noise. 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 29 Jan.-1 Feb. 2018, Moscow, Russia (2018). doi: 10.1109/EIConRus.2018.8317275
7. Huang, S.-C., Chuang, P.-J., Wu, C.-F., Lai, H.-J.: Chaos-based support vector regressions for exchange rate forecasting. Expert Systems with Applications. **37**, 8590 – 8598 (2010). doi: 10.1016/j.eswa.2010.06.001
8. Premanode, B., Toumazou, C.: Improving prediction of exchange rates using differential EMD. Expert Systems with Applications, **40**, 377 – 384 (2013). doi: 10.1016/j.eswa.2012.07.048
9. Mabu, S., Hirasawa, K., Obayashi, M., Kuremoto, T.: Enhanced decision-making mechanism of rule-based genetic network programming for creating stock trading signals. Expert Systems with Applications. **40**, 6311 – 6320 (2013). doi: 10.1016/j.eswa.2013.05.037
10. Bahrepour, M., Akbarzadeh, T. M. –R., Yaghoobi, M., Naghibi, S. M. –B.: An adaptive ordered fuzzy time series with application to FOREX. Expert Systems with Applications. **38**, 475–485 (2011). doi: 10.1016/j.eswa.2010.06.087
11. Orlov, Yu.N., Shagov, D.O.: Indicative statistics for non-stationary time series. Keldysh Institute preprints. **53**, 1 – 20 (2011) (in Russian: Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В.Келдыша. 2011. № 53. 20 с. URL: http://library.keldysh.ru/preprint.asp?id=2011-53)
12. Orlov Yu.N., Fedorov, S.L.: Generation of non-stationary trajectories of a time series based on Fokker-Planck equation. MFTI Proceedings. **8**, 2, 126 – 133 (2016) (in Russian: Орлов, Ю.Н., Федоров, С.Л. Генерация нестационарных траекторий временного ряда на основе уравнения Фоккера – Планка. ТРУДЫ МФТИ. **8**, 2, 126 – 133 (2016))
13. Klochkov, Y, Gazizulina, A, Golovin, N., Glushkova, A., Selezneva, Zh.: Information model-based forecasting of technological process state. Proceedings of the International Conference on Infocom Technologies and Unmanned (ICTUS), Dubai, United Arab Emirates (2017). doi: 10.1109/ICTUS.2017.8286099
14. Didenko, N., Kulik, S.: Environmental Shocks: Modelling the Dynamics. 2018 IOP Conf. Ser.: Earth Environ. Sci. 180 012013 (2018). doi: 10.1088/1755-1315/180/1/012013
15. Fuentes, M.: Non-Linear Diffusion and Power Law Properties of Heterogeneous Systems: Application to Financial Time Series. Entropy. **20(9),** 649, 1 – 8 (2018). doi: https://doi.org/10.3390/e20090649

# From Long Memory to Oscillatory Modes – The Potentials of Detrended Fluctuation Analysis

Philipp G. Meyer and Holger Kantz

Max Planck Institute for the Physics of Complex Systems
Noethnitzer Str. 38 D 01187 Dresden Germany

**Abstract.** Detrended fluctuation analysis is a popular method for the detection of long range correlations in time series. Recently further insights were gained which allow a theoretical discussion of its results. Since then the method is no longer restricted to long memory, but can explain the whole correlation pattern of the signal and enables us to infer characteristic timescales. We show its traditional usage in a toy model for long range correlations and thereby discuss the relation to other anomalous statistical properties. Then we summarize the uncovered potentials of detrended fluctuation analysis for short range correlated systems showing examples from atmospheric science.

**Keywords:** detrending methods, Data decomposition, spectrum analysis, Atmospheric science forecasting

## 1 Introduction

Anomalous statistical behaviour has been of interest for the scientific community for decades [1] since Hurst found a scaling exponent in the time series of river Nile levels [2]. Such anomalous properties might originate from different properties of the system [3]. One prominent cause are long range correlations (LRC) or, in other words, diverging correlation times. Detrended fluctuation analysis (DFA) [4] is a popular tool for detecting LRC in data. It has been applied successfully to real world time series like temperatures [5], heart rates [6], biology [7], finance [8] and various others where anomalous exponents were found. The disadvantage of DFA and equivalent methods like detrended moving averages [9] or wavelet analysis [10] is that results are not always easy to interpret. Emergent scaling might arise from short range correlations if the dataset is too short [11], from the superposition of multiple timescales, or 'real' dynamical long range correlations [12].

Recently theoretical understanding ob DFA improved drastically. A relation between the fluctuation function of DFA and the correlation function was established [13]. This enables us to calculate theoretical fluctuation functions for well known processes [14] and even fitting of the parameters of such processes to real data [15].

The advantages of DFA are its numerical stability and the smoothness of the fluctuation function as well as the possibility to neglect trends, i.e. slow changes of the variable, and to concentrate on the short and intermediate timescale [16]. It is especially suitable for fitting because it separates characteristic timescales unlike the correlation function where oscillations of the signal also lead to oscillations in the correlation function. Due to the logarithmic scale of the fluctuation function, all timescales up to the maximum of 1/4th of the total measurement time are equally taken into account.

As stated above data analysis with DFA is not restricted to the search for LRC. DFA can also be used to generate short range correlated data models that accurately describe fluctuations on a wide range of timescales. The method can even be used to identify characteristic timescales in data. Due to the separation of timescales it is even applicable to signals with several timescales if these characteristic timescales are not too close to each other.

We want to show all three applications mentioned above. After introducing our method in section 2 we will introduce a model for long range correlations in section 3 and show that DFA works here even if the second moment of the process does not exist. For further examples we will concentrate on real world data from atmospheric science. In section 4 we use DFA for generating a suitable data model for sea level pressure in Europe. In section 5 we decompose the fluctuations of a pressure time series from Indonesia where we observe the Madden-Julian oscillation and the El Niño southern oscillation.

## 2 Methods

### 2.1 Implementation of DFA

DFA is a well known method for examining correlations of time series [4]. It is implemented as follows. Given a time series $x_t$ we first calculate the integral $y_t = \sum_{j=1}^{t} x_j$. Then we divide the time axis into $K$ non-overlapping segments of length $s$ and calculate the so-called DFA variance $f^2(\nu, s)$ in every segment $\nu$. It is defined as

$$f^2(\nu, s) = \frac{1}{s} \sum_{t=1+(\nu-1)s}^{\nu s} (y_t - p_t)^2. \tag{1}$$

Where $p_t$ is a polynomial of order q, which is fitted to $y_t$. The order $q$ is the detrending order. DFAq is able to remove polynomial trends of order $q-1$. The squared fluctuation function of DFA is defined as the average of all the DFA variances over all segments

$$F^2(s) = \frac{1}{K} \sum_{\nu=1}^{2K} f^2(\nu, s). \tag{2}$$

Traditionally people look at the asymptotic behaviour $F^2(s) \sim s^{2\alpha}$, which shows whether the process is long range correlated ($\alpha > 1/2$) or short range correlated ($\alpha = 1/2$).

### 2.2 Relation to the correlation function

The expression of the fluctuation function in terms of the autocorrelation function $C(t)$ was derived in [13]

$$F_x^2(s) = \sigma_x^2 \left( L_q(0,s) + 2\sum_{t=1}^{s-1} C_{xx}(t) L_q(t,s) \right), \tag{3}$$

where $\sigma_x^2$ is the variance of the process $x$. $L_q$ is a kernel that determined the detrending order. The kernel $L_1$ in DFA1 is

$$\begin{aligned} L_1(t,s) = \tfrac{1}{30(s^4-s^2)}[3t^5 - 5(4s^2-1)t^3 + 30(s^3-s)t^2 \\ -(15s^4 - 35s^2 + 8)t + 2(s^5 - 5s^3 + 4s)]. \end{aligned} \tag{4}$$

We exclusively use DFA1 throughout this text.

### 2.3 Fitting procedure

If the correlation function is known we can calculate the model fluctuation function using equation (3). It can be fitted to the fluctuation function of the data by minimizing the variance

$$var\left[\log\left(\frac{F_{data}(s)}{F_{model}(s)}\right)\right]. \tag{5}$$

The pre-factor $\sigma_x^2$ in equation (3) is not fitted but obtained by rescaling $F_{model}(s)$ to match $F_{data}(s)$.

## 3 Long range correlations

### 3.1 The Joseph effect

Anomalous statistical behavior describe dynamics beyond the central limit theorem. That means the distribution of a diffusive process $y$ scales like

$$P(y_t) = t^{-H} P^*\left(y_t / t^H\right), \tag{6}$$

with the Hurst exponent $H$. The premises of the central limit theorem can be violated in three ways. The Joseph exponent $J$ [17] quantifies the increment correlations. It is equivalent to the exponent $\alpha$ in detrended fluctuation analysis. While short range correlations do not change the asymptotic properties, LRC lead to anomalous scaling $J \neq 1/2$. The latent exponent $L$ quantifies the effect of fat-tails in the increment distribution (Noah effect). When $L > 1/2$, the increment distribution has "fat tails". The Moses exponent $M$ [3] quantifies the effect of non-stationarity of the increment distributions. When $M > 1/2$ the increment distribution widens with time, and for $M < 1/2$ it shrinks with time. The four exponents are related via

$$H = J + L + M - 1\,. \tag{7}$$

Detecting the Josef effect (LRC) in data is important, since it leads to slow decay of large deviations and shrinking of the effective sample size [18].

### 3.2 The Pomeau-Manneville map

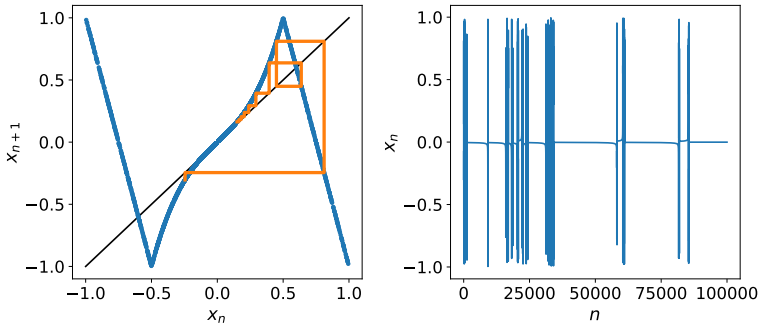



**Fig. 1.** Left: the Pomeau-Manneville map (blue) and the first iterates of a trajectory displayed in orange. Right: trajectory with $10^5$ iterations.

Consider a time series generated by a symmetric version of the Pomeau-Manneville (PM) map [19]

$$x_{t+1} = \begin{cases} -4x_t + 3 & \text{if } 0.5 < x_t \leq 1.0 \\ x_t \left(1 + |2x_t|^{z-1}\right) & \text{if } |x_t| \leq 0.5 \\ -4x_t - 3 & \text{if } -1 \leq x_t < -0.5 \end{cases}, \tag{8}$$

with $z > 1$. This map has been studied extensively in the past. It has been linked to anomalous diffusion [20], aging [21] and weak ergodicity breaking [22]. The initial increment $x_0$ is chosen randomly from a uniform distribution in the interval $[-1, 1]$.

The integral $y_t = \sum_{n=1}^{t} x_n$ exhibits anomalous diffusion caused by all three root causes $J$, $L$ and $M$ [23] for parameters $z > 2$. This is due to the intermittent dynamics of the map that leads to power law distributed waiting times close to zero as figure 1 shows. We only consider the case $z = 3$.

The existence of the Moses effect and the Noah effect mean that the variance $\langle x^2 \rangle$ does not exist which means that also the fluctuation function can no longer be defined according to equation (3). Can we still quantify the Joseph effect via DFA? In figure 2 we show that we can do so, since the scaling of the fluctuation function is pretty stable for all realizations but the ones which are trapped around zero for so long that there are almost no spikes within the observed interval. The divergence of the variance leads to a random factor to the fluctuation function that does not change the scaling. This effect was already observed for time averages of the Pomeau-Manneville map [24] and for the power spectrum [25].

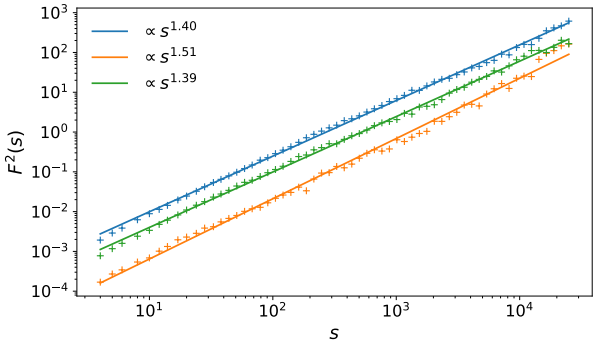



**Fig. 2.** DFA fluctuation functions for three trajectories of the Pomeau-Manneville map for $z = 3$ with different initial conditions. While the amplitude is different for all realizations, the scaling is relatively stable. The theoretical value is $2\alpha = 1.5$ [23].

## 4 Fitting fluctuation functions

### 4.1 Short memory and data models

For some systems due to high complexity, accurate modeling of the physical process is very complicated. In these cases, there is a demand for data models, that reproduce the statistics of the system without requiring physical insight in the dynamics. DFA is a very useful tool that can help us to obtain such models [15].

In the last section we saw that long memory in DFA leads to scaling $F(s) \propto s^{\alpha}$ with $\alpha > 0.5$. Short range correlations, i.e. dynamics that are described by relaxations with finite correlation time, in the long time limit scale like $F(s) \propto s^{1/2}$. These properties describe the majority of investigated time series in the real world. It is compatible with chaos theory and sensitive dependence on initial conditions described by Lyapunov exponents. The non-asymptotic shape of the fluctuation function of such processes is concave. The exact shape can be calculated from equation (3) if the correlation function is known [13].

### 4.2 The autoregressive model of order one AR(1)

The most basic example for a short range correlated process is the AR(1) model (9)

$$X_t = cX_{t-1} + \epsilon_t, \tag{9}$$

which describes a system driven by white noise $\epsilon$, that responds linearly with relaxation time $r = -1/\log(c)$, dependent on th AR-parameter $c$. The variance of the process is given by

$$\sigma_X^2 = \sigma_\epsilon^2/(1 - c^2), \tag{10}$$

where $\sigma_\eta^2$ is the variance of the noise. The correlation function of AR(1) is known to be $C(t) = c^t$. The fluctuation function for AR(1) can therefore be calculated as

$$F_c^2(s) = \sigma^2 \frac{c^s J_c(s) + K_c(s)}{15(c-1)^6(s^2-s^4)}, \tag{11}$$

with $J_c(s)$, $K_c(s)$ polynomials in $s$

$$\begin{aligned} J_c(s) &= 60[s^2(c^2-c)^2 - 3s(c^3-c) + 2(c^4+c^3+c^2)], \\ K_c(s) &= s^5(c-1)^5(c+1) + 15s^4c(c-1)^4 \\ &\quad -5s^3(c-1)^3(1-7c-7c^2+c^3) - 15s^2c(c-1)^2(1-10c+c^2) \\ &\quad +2s(c-1)^3(2-17c-17c^2+2c^3) - 120c^2(1+c+c^2). \end{aligned} \tag{12}$$

### 4.3 European sea level pressure

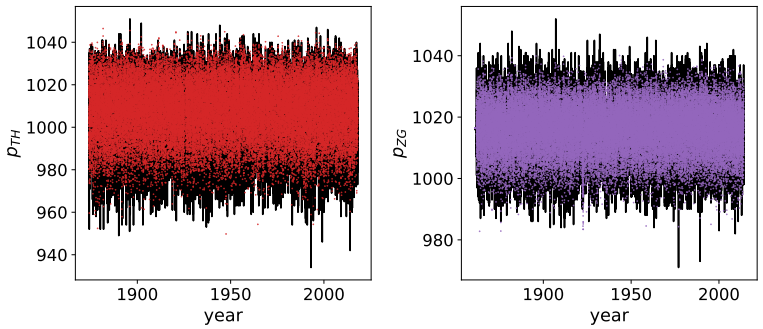



**Fig. 3.** Sea level pressure from Torhavn (left) and Zagreb-Gric (right). The black lines mark the recordings while the anomalies are plotted as colored dots.

As an example for a system for which a good data model can be obtained by fitting the fluctuation function we discuss daily mean sea level pressure from two European stations. Data is provided by the ECA&D project [26], publicly available online at `http://www.ecad.eu`. We chose Torhavn, Faroe Islands and Zagreb-Gric, Croatia, which are two of the longest time series available.

Sea level pressure in Europe exhibits seasonality that is in contrast to temperature data not mainly characterized by oscillations in the mean value [27], but rather oscillations in the amplitude of the fluctuations. Therefore we not only calculate the long time climatological average pressure for each calender day and subtract it from each day. We also divide each value by the average variance for the calender day. The resulting pressure anomalies can be described by a stationary model. We show the raw data and the anomalies for both stations in figure 3.

We model the data by an AR(1) process, but we fit its parameters by fitting equation (11) to the fluctuation function of the data [15]. The plot shows that the fit is excellent for Torhavn and works reasonably well for Zagreb where some small but systematic deviations seem to be present. The obtained relaxation times are 3.21 days for Torhavn and 3.74 days for Zagreb. Thus we obtain a data model that describes the power of the fluctuations of pressure anomalies. In figure 4 (right) we show what that means. The fluctuations of the time series, and at the same time for the 60 days mean, are well captured by the AR(1) model. Our obtained model shows no sign of predictability beyond persistence for short times.

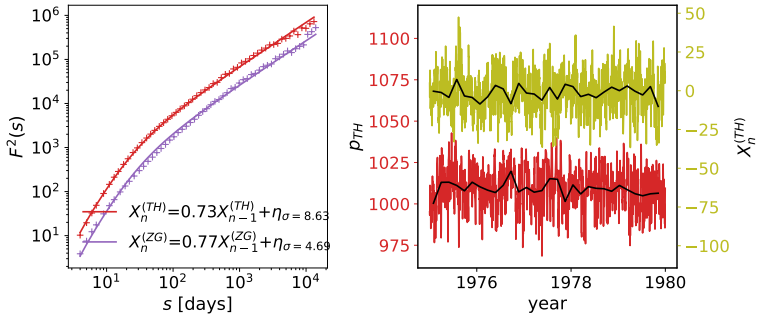



**Fig. 4.** Left: DFA of Torhavn sea level pressure (red) and Zagreb sea level pressure (purple) fitted by theoretical fluctuation functions of AR(1). Right: 5 years of the Torhavn time series (red) compared to an equally long time series of a realization of the fitted AR(1) model (yellow). The black lines indicate 60-days means.

## 5 Fitting signals with more than one characteristic timescale

### 5.1 The superposition principle of DFA

DFA highlights all characteristic timescales as local maximum values of the negative curvatures of the fluctuation function. We can formulate a model that captures several timescales in terms of the simple autoregressive models. We assume that the signal can be approximated by a superposition $X_{t_X} + Y_{t_Y}$, where $X$ and $Y$ are processes with a characteristic relaxation time or oscillatory mode. Timescales of the two process should be sufficiently well separated. Then we can make use of the superposition principle of DFA [28]

$$F^2_{X+Y}(s) = F^2_X(s) + F^2_Y(s). \tag{13}$$

We can fit both timescales individually and are thus able to use our method to systems with more than one characteristic timescale.

### 5.2 The autoregressive model of order two AR(2)

The autoregressive model of order two, AR(2), is defined by

$$Y_t = aY_{t-1} + bY_{t-2} + \eta_t, \tag{14}$$

where $\eta$ is white noise and $a$, $b \in \mathbb{R}$ are the AR-parameters. Its variance is

$$\sigma^2 = \frac{(1-b)\sigma_\eta^2}{(1+b)(1-a-b)(1+a-b)}. \tag{15}$$

The correlation function is given by $C(t) = h_1 g_1^t + h_2 g_2^t$ [29], where $g_1$ and $g_2 \in \mathbb{C}$ are the roots obtained from rewriting the definition (14) with the backshift operator $BX_t = X_{t-1}$ as $(1 - g_1 B)(1 - g_2 B)x_t = \eta_t$. The constants $h_1$ and $h_2$ are calculated from $C(1) = a/(1-b)$ and $C(0) = 1$. If the roots are complex the systems exhibits oscillatory dynamics with a period

$$\tau = \frac{2\pi}{\arctan(\mathrm{Im}(g_1)/\mathrm{Re}(g_1))}. \tag{16}$$

The fluctuation function for AR(2) can be expressed in terms of the fluctuation function of AR(1), because the correlation function is the superposition of two AR(1) correlation functions with potentially complex parameters. It reads [15]

$$F^2_{a,b}(s) = \sigma^2 \frac{[h_1 F^2_{g_1}(1-g_1^2) + h_2 F^2_{g_2}(1-g_2^2)](1-b)}{(1+b)(1-a-b)(1+a-b)}, \tag{17}$$

where $F_{g_i}$ are the fluctuation functions of the AR(1) processes with parameters $g_1$ and $g_2$.

### 5.3 Equatorial Pacific sea level pressure

Here we want to investigate the full time series of pressure recordings from Kemayoran, Indonesia, where between 1.1.1866 and 16.11.1945 only few values are missing (see figure 5). Data is available at `http://sacad.database.bmkg.go.id`, as part of the ICA&D project [30]. We calculate the anomalies as we did for the European stations and perform DFA on them. It is clear, that a simple model like AR(1) can not describe the fluctuations. The reason is that there are significant oscillatory modes.

The El Niño Southern Oscillation (ENSO) is the strongest teleconection in the atmosphere of the earth. It describes persistent weather patterns in the equatorial Pacific region. The question, weather the phenomenon is rather an oscillatory mode or a two state stochastically driven system is still under debate [31], however, people have found different estimates for an intrinsic frequency. If one averages over 11 months of a measure related to sea surface temperatures, one gets a signal which can be described by an AR(2) model with a characteristic oscillation period of 3.3 years [15].

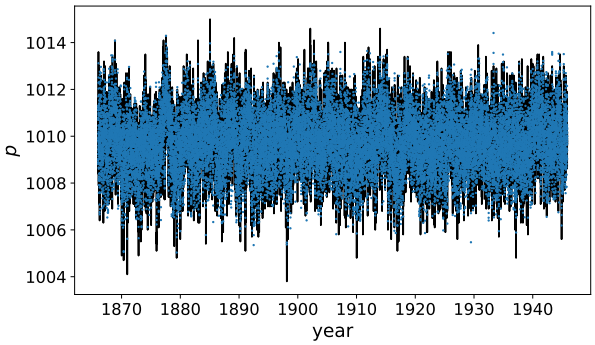



**Fig. 5.** Sea level pressure from Kemayoran. The black lines mark the recordings while the anomalies are plotted as colored dots.

There are different measures describing the phenomenon. The Southern Oscillation Index SOI is defined as the pressure difference between Tahiti and Darwin, Australia. In contrast to the pressure time series discussed in section 4 it exhibits oscillatory modes on the timescale of years. As described by [32] for the southern oscillation index, the signal does not only contain the long El Niño period, but also the shorter Madden-Julian Oscillation [33], that is typical for the equatorial area and has a period around 40-50 days.

Therefore we assume that the signal can be approximated by a superposition of three processes

$$p_t \approx X_{t_X} + Y_{t_Y} + Z_{t_Z}, \tag{18}$$

with an AR(1) process $X_t$ describing the short time dynamics and AR(2) processes $Y_t$ and $Z_t$ describing the oscillation. This is clearly a simplification and we should expect that other effects cause errors in our parameter fitting. However, the advantage of our approach is that we obtain a simple model which captures the complete power of the fluctuations and can be used as an approximation of the dynamics with qualitatively correct characteristic timescales if our fit works reasonably well.

The fitting procedure is as follows (see figure 6). We fit the theoretical fluctuation function $F_X$ of AR(1) to it for small $s$. The result is subtracted from $F_p$ according to equation (13) and we obtain $F_{p-X}$. Now due to the separation of timescales we can go to the 3-days timescale by multiplying $s_Y = 3s_X$ and dividing $F_{p-X}(s)$ by 3. Now we neglect points obtained from the daily dataset (and therefore all values $s < 4$) and fit the complete remaining $F_{p-X}$ by the theoretical fluctuation function for AR(2) $F_Y$. We rescale the resulting $F_{p-X-Y}$ by 30.417 and repeat the procedure for fitting $F_Z$ on the 3 month timescale. The free parameters are the AR-coefficients and the noise variances are calculated according to equation (5).

The result is shown in figure 6. For the relaxation time we obtain 3.63 days. The oscillation period of the AR(2) process $Y$ is 42.8 days, which is in the range

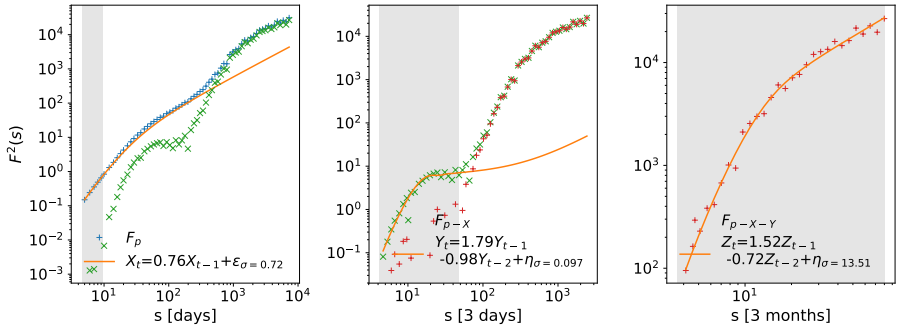



**Fig. 6.** LEFT: full fluctuation function $F_p$ from Kemayoran (blue +), fitted by AR(1) model $F_X$ (orange line) in the range highlighted in grey and the result for $F_{p-X}$ (green x). CENTER: $F_{p-X}$ fitted by $F_Y$ (orange line) in the grey area and difference $F_{p-X-Y}$ (red +) RIGHT: $F_{p-X-Y}$ (red +) and fitted AR(2) fluctuation function $F_Z$ (orange line).

of the Madden-Julian oscillation. For $Z$ it is 40.7 months, which is similar to the period obtained for the 11 months average ENSO index in [15] and compatible with other results. Note that our model even though being dominated by noise has some predictive power due to the oscillatory mode. So there is some determinism in the statistics. This explains why El Niño forecasts are not hopeless and there has been some success with different methods [34, 35].

## 6 Conclusion

Detrended fluctuation analysis is a powerful tool for time series analysis. The fluctuation function gives a lot of insights in the dynamics of a process. It can be used for detecting long range correlations, for providing data models, and for identifying characteristic timescales in data. It can even be used for signals with several characteristic timescales if these time scales are well enough separated. The obtained values are approximations that should be interpreted with care, however, immediate validation by looking at the quality of the fit is possible.

Detrended fluctuation analysis has several advantages over other methods. Unlike the correlation function or the power spectrum, the fluctuation function is an increasing function which is therefore easy to fit. It is also numerically very stable up to 1/4th of the total measurement time. Trends can easily be removed from the signal. Unlike spectral methods it measures the full power of the fluctuations. Unlike for the correlation the characteristic timescales in the fluctuation function are ordered and oscillations in the signal do not lead to long time oscillations of the fluctuation function.

## Acknowledgments

We acknowledge the data providers in the ECA&D and the SACA&D project, Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453. Data and metadata available at `http://www.ecad.eu` and `http://sacad.database.bmkg.go.id`.

## References

1. Graves, T., Gramacy, R., Watkins, N., Franzke, C.: A brief history of long memory: Hurst, Mandelbrot and the road to arfima, 1951-1980. Entropy **19(9)** (2017) 437
2. Hurst, H.E.: Long-term storage capacity of reservoirs. Am. Soc. Civil Eng. **116** (1951) 770
3. Chen, L., Bassler, K.E., McCauley, J.L., Gunaratne, G.H.: Anomalous scaling of stochastic processes and the Moses effect. Phys. Rev. E **95** (Apr 2017) 042141
4. Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L.: Mosaic organization of DNA nucleotides. Phys. Rev. E **49** (1994) 1685
5. Fraedrich, K., Blender, R.: Scaling of atmosphere and ocean temperature correlations in observations and climate models. Phys. Rev. Lett. **90** (2003) 108501
6. Bunde, A., Havlin, S., Kantelhardt, J.W., Penzel, T., Peter, J.H., Voigt, K.: Correlated and uncorrelated regions in heart-rate fluctuations during sleep. Phys. Rev. Lett. **85** (Oct 2000) 3736–3739
7. Wan, K.Y., Goldstein, R.E.: Rhythmicity, recurrence, and recovery of flagellar beating. Phys. Rev. Lett. **113** (Dec 2014) 238103
8. Baillie, R.T.: Long memory processes and fractional integration in econometrics. Journal of econometrics **73**(1) (1996) 5–59
9. Alessio, E., Carbone, A., Castelli, G., Frappietro, V.: Second-order moving average and scaling of stochastic time series. The European Physical Journal B - Condensed Matter and Complex Systems **27**(2) (May 2002) 197–200
10. Abry, P., Veitch, D.: Wavelet analysis of long-range-dependent traffic. IEEE Transactions on Information Theory **44**(1) (Jan 1998) 2–15
11. Maraun, D., Rust, H.W., Timmer, J.: Tempting long-memory – on the interpretation of dfa results. Nonlin. Process. Geophys. **11** (2004) 495
12. Meyer, P., Kantz, H.: Infinite invariant densities due to intermittency in a nonlinear oscillator. Phys. Rev. E **96** (Aug 2017) 022217
13. Hoell, M., Kantz, H.: The relationship between the detrendend fluctuation analysis and the autocorrelation function of a signal. Eur. Phys. J. B **88** (2015) 327
14. Hoell, M., Kantz, H.: The fluctuation function of the detrended fluctuation analysis – investigation on the AR(1) process. Eur. Phys. J. B **88** (2015) 126
15. Meyer, P.G., Kantz, H.: Inferring characteristic timescales from the effect of autoregressive dynamics on detrended fluctuation analysis. New Journal of Physics **21**(3) (mar 2019) 033022
16. Höll, M., Kiyono, K., Kantz, H.: Theoretical foundation of detrending methods for fluctuation analysis such as detrended fluctuation analysis and detrending moving average. Phys. Rev. E **99** (Mar 2019) 033305
17. Mandelbrot, B.: The variation of certain speculative prices. The Journal of Business **36**(4) (1963) 394–419

18. Massah, M., Nicol, M., Kantz, H.: Large-deviation probabilities for correlated Gaussian processes and intermittent dynamical systems. Physical Review E **97**(5) (2018) 052147
19. Pomeau, Y., Manneville, P.: Intermittent transition to turbulence in dissipative dynamical systems. Communications in Mathematical Physics **74**(2) (Jun 1980) 189–197
20. Geisel, T., Thomae, S.: Anomalous diffusion in intermittent chaotic systems. Phys. Rev. Lett. **52** (1984) 1936
21. Barkai, E.: Aging in subdiffusion generated by a deterministic dynamical system. Phys. Rev. Lett. **90** (Mar 2003) 104101
22. Bel, G., Barkai, E.: Ergodicity breaking in a deterministic system. Europhys. Lett. **74** (2006) 15
23. Meyer, P.G., Adlakha, V., Kantz, H., Bassler, K.E.: Anomalous diffusion and the Moses effect in an aging deterministic model. New Journal of Physics **20**(11) (nov 2018) 113033
24. Meyer, P., Barkai, E., Kantz, H.: Scale-invariant Green-Kubo relation for time-averaged diffusivity. Phys. Rev. E **96** (Dec 2017) 062122
25. Niemann, M., Kantz, H., Barkai, E.: Fluctuations of $1/f$ noise and the low-frequency cutoff paradox. Phys. Rev. Lett. **110** (Apr 2013) 140603
26. Tank, A.K., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al.: Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. International journal of climatology **22**(12) (2002) 1441–1453
27. Deng, Q., Nian, D., Fu, Z.: The impact of inter-annual variability of annual cycle on long-term persistence of surface air temperature in long historical records. Climate Dyn. **50** (2018) 1091
28. Hu, K., Ivanov, P.C., Chen, Z., Carpena, P., Stanley, H.E.: Effect of trends on detrended fluctuation analysis. Phys. Rev. E **64** (2001) 011114
29. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
30. Van Den Besselaar, E.J.M., Klein Tank, A.M.G., Van Der Schrier, G., Abass, M.S., Baddour, O., Van Engelen, A.F., Freire, A., Hechler, P., Laksono, B.I., Iqbal, Jilderda, R., Foamouhoue, A.K., Kattenberg, A., Leander, R., Gingla, R.M., Mhanda, A.S., Nieto, J.J., Sunaryo, Suwondo, A., Swarinoto, Y.S., Verver, G.: International climate assessment & dataset: Climate services across borders. Bulletin of the American Meteorological Society **96**(1) (2015) 16–21
31. Wang, C., Deser, C., Yu, J.Y., DiNezio, P., Clement, A.: El Niño and southern oscillation (enso): a review. In: Coral Reefs of the Eastern Tropical Pacific. Springer (2017) 85–106
32. Whitcher, B., Guttorp, P., Percival, D.B.: Wavelet analysis of covariance with application to atmospheric time series. Journal of Geophysical Research: Atmospheres **105**(D11) (2000) 14941–14962
33. Madden, R.A., Julian, P.R.: Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. Journal of the atmospheric sciences **28**(5) (1971) 702–708
34. Chekroun, M.D., Kondrashov, D., Ghil, M.: Predicting stochastic systems by noise sampling, and application to the el niño-southern oscillation. Proceedings of the National Academy of Sciences **108**(29) (2011) 11766–11771
35. Ludescher, J., Gozolchiani, A., Bogachev, M.I., Bunde, A., Havlin, S., Schellnhuber, H.J.: Improved el niño forecasting by cooperativity detection. Proceedings of the National Academy of Sciences **110**(29) (2013) 11742–11745

# THE CORRESPONDENCE BETWEEN STOCHASTIC LINEAR DIFFERENCE AND DIFFERENTIAL EQUATIONS

By D.S.G. POLLOCK

University of Leicester

*Email:* stephen_pollock@sigmapi.u-net.com
*Website:* http://www.le.ac.uk/users/dsgp1/

The relationship between autoregressive moving-average (ARMA) models in discrete time and the corresponding models in continuous time is examined in this paper. The linear stochastic models that are commonly regarded as the counterparts of the ARMA models are driven by a forcing function that consists of the increments of a Wiener Process. This function is unbounded in frequency.

In cases where the periodogram of the data indicates that there is a clear upper bound to its frequency content, we propose an alternative frequency-limited white-noise forcing function. Then, there is a straightforward translation from the ARMA model to a differential equation, which is based of the principle of impulse invariance.

Whenever there is no perceptible limit to the frequency content, the translation must be based on a principle of autocovariance equivalence. On the website of the author, there is a computer program that effects both of these discrete-to-continuous translations.

## 1. Introduction: The Discrete–Continuous Correspondence

Modern communications technology relies on the correspondence between continuous signals and the discrete sequences that come from sampling the signals rapidly at regular intervals. Familiar examples of the technology are the analog–digital conversions of digital radio, digital sound recordings and digital television; but the domain of this technology is much wider.

The basis of digital technology is the sampling theorem of Nyquist (1924, 1928) and of Shannon (1949), which indicates that, if a signal is sampled with sufficient rapidity, then it can be reconstituted with complete accuracy from the sampled sequence.

The theorem is a commonplace amongst electrical engineers. It ought to be equally familiar to econometricians and statisticians and, in particular, to time-series analysts, but it has been widely ignored.

This discrete–continuous equivalence began to be widely recognised at the end of the nineteenth century with the advent of the cinema. The cinema creates moving pictures from a sequence of fixed images projected in rapid succession. In the early days of the cinema, the succession of images was insufficiently rapid to convey an impression of smooth motion. The pictures tended to flicker; and, in popular parlance, we still refer to visiting the cinema as 'going to the flicks'.

There is a revealing picture by Marcel Duchamp, exhibited in the Paris Salon des Independents of 1912, which is titled *A Nude Descending a Staircase.* It exposes the paradox of the discrete–continuous correspondence; and it makes an allusion to the jerky motion of the early cinema.

Occasionally, the true nature of motion pictures is revealed by an odd quirk that occurs when the rate of sampling is insufficient to convey a convincing impression of a rapid motion. Those of a certain age will have seen a depiction of a stagecoach fleeing its pursuers. They will have noticed the blurred impression of the wagon wheels. At times, these appear to be rotating slowly in the direction of travel. At other times, they seem to be stationary, and they may even, on occasion, appear to be moving backwards.

These are instances of the so-called problem of aliasing, whereby a motion that is too rapid to be captured by the sampling process is proxied by a much slower motion.

The Shannon–Nyquist sampling theorem is an adjunct of a Fourier analysis, which depicts a temporal trajectory as a weighted combination of trigonometric functions. The theorem indicates that, if the sampled sequence is fully to capture a continuous motion, then it is necessary that at least two observations should be made in the time that it takes for the trigonometric element of highest frequency to complete a single cycle. This rate of sampling, which corresponds to a signal frequency of $\pi$ radians per sampling interval, is the so-called Nyquist relative frequency.

If the frequencies within the signal exceed the Nyquist value of $\pi$, then there will be an irremediable loss of information and it will not be possible fully to reconstitute the signal from the sampled data. Conversely, if the maximum frequency within the signal is less than the Nyquist value, then the sampling is over-rapid and other problems can arise; but these problems ought, in principle, to be remediable.

## 2. ARMA Estimation and the Effects of Over-rapid Sampling

A problem can arise in the estimation of an ARMA model when the rate of sampling exceeds the maximum frequency within the signal. The problem can be illustrated with the deseasonalised quarterly data on U.S. gross domestic product (GDP) from which a trend has been extracted with the filter of Leser (1961) and of Hodrick and Prescott (1980, 1997)—see Figure 1. The problem is revealed by examining the periodogram of the data, which is a product of its Fourier transform.

The Fourier analysis expresses the detrended data sequence $y(t) = \{y_t; t = 0, 1, \ldots, T-1\}$ as

$$y(t) = \sum_{j=0}^{[T/2]} \{\alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t)\} = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t}, \tag{1}$$

where $\omega_j = 2\pi j/T; j = 0, \ldots, [T/2]$ are the Fourier frequencies, which are placed at regular intervals running from zero up to the Nyquist frequency $\pi$, or just short of it by a half interval. Here, $[T/2]$ denotes the integer quotient of the division of $T$ by 2.

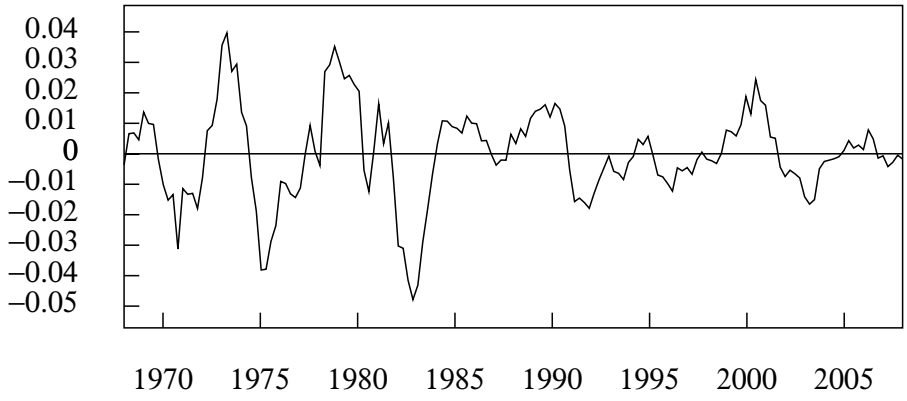



**Figure 1.** The deviations of the logarithmic quarterly index of real US GDP from an interpolated trend. The observations are from 1968 to 2007. The trend is determined by a Hodrick–Prescott (Leser) filter with a smoothing parameter of 1600.

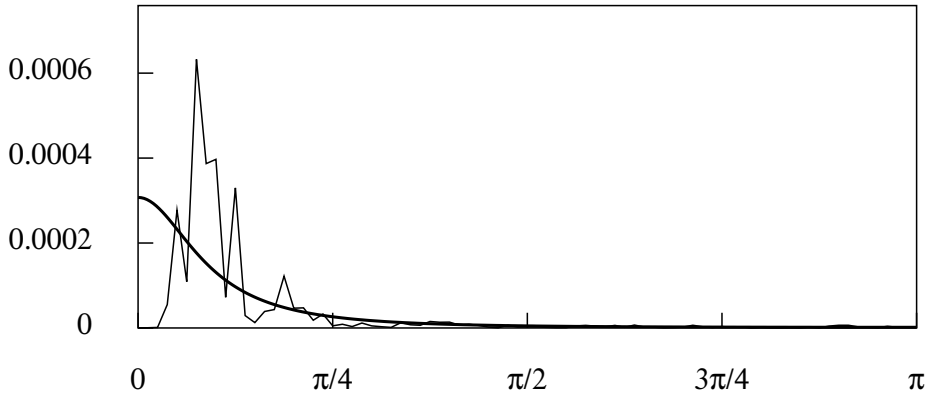



**Figure 2.** The periodogram of the data points of Figure 1 overlaid by the parametric spectral density function of an estimated regular AR(2) model.

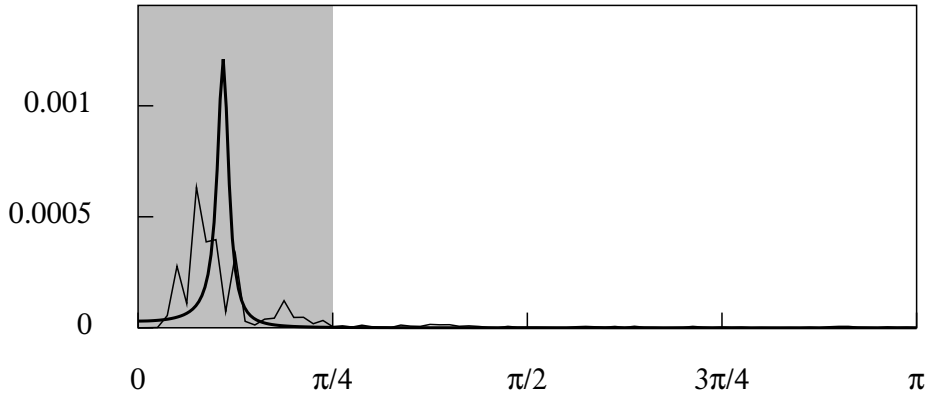



**Figure 3.** The periodogram of the data points of Figure 1 overlaid by the spectral density function of an AR(2) model estimated from frequency-limited data.

The second expression, which employs complex exponential functions, arises from Euler's equations, whereby

$$\cos(\omega_j t) = \frac{e^{\mathrm{i}\omega_j t} + e^{-\mathrm{i}\omega_j t}}{2} \quad \text{and} \quad \sin(\omega_j t) = \frac{-\mathrm{i}}{2}(e^{\mathrm{i}\omega_j t} - e^{-\mathrm{i}\omega_j t}). \tag{2}$$

Conversely, there are

$$e^{\mathrm{i}\omega_j t} = \cos(\omega_j t) + \mathrm{i}\sin(\omega_j t) \quad \text{and} \quad e^{-\mathrm{i}\omega_j t} = \cos(\omega_j t) - \mathrm{i}\sin(\omega_j t), \tag{3}$$

and it follow that $\exp\{-\mathrm{i}\omega_j t\} = \exp\{\mathrm{i}\omega_{T-j} t\}$. Also, $\xi_j = (\alpha_j - \mathrm{i}\beta_j)/2$ and $\xi_{T-j} = (\alpha_j + \mathrm{i}\beta_j)/2$ for $j = 0, 1, \ldots, [T/2]$. These results enable the two expressions of (1) to be reconciled.

The coefficients $\alpha_j, \beta_j$ are obtained by regressing the data on the ordinates of the trigonometric functions $\cos(\omega_j t), \sin(\omega_j t)$, where $t = 0, 1, \ldots, T-1$. It should be observed that, if the maximum frequency in the signal is less than $\pi$, then some of these coefficients will be zero valued.

The periodogram is the plot of the squared amplitudes $\rho_j^2 = \alpha_j^2 + \beta_j^2$, and it conveys a frequency-specific analysis of variance. That is to say

$$V(y) = \frac{1}{T}\sum_t (y_t - \bar{y})^2 = \frac{1}{2}\sum_j \{\alpha_j^2 + \beta_j^2\} = \frac{1}{2}\sum_j \rho_j^2. \tag{4}$$

The periodogram of the detrended logarithmic quarterly index of real US GDP is depicted in Figures 2 and 3.

An attempt can be made to capture the business cycle dynamics by fitting an AR(2) model to the detrended data. The expectation is that the poles of the model, i.e. its autoregressive roots, will be a conjugate complex pair. The modulus of the roots should represent the damping characteristics of the business cycle and their argument should represent an angular velocity, which would indicate the average duration of the business cycle. The parametric spectrum of the fitted ARMA model should mimic the shape of the periodogram, with its peak in roughly the same position as that of the periodogram.

When the parametric spectrum of the estimated AR(2) model is superimposed on the periodogram in Figure 2, it becomes clear that, in place of the expected complex roots, there are two real-valued roots.

In diagnosing the problem, it is recognised that there are minor elements of noise affecting the data throughout the frequency interval running for the cut-off point of the spectral signature of the business cycle at $\omega_c = \pi/4$ up to the Nyquist frequency of $\pi$. This noise is making a significant contribution to the variance of the data without greatly affecting the autocovariances at positive lags. As a result, the initial values, which determine the estimates of the autoregressive parameters, show an exaggerated rate of decline, or damping, which gives rise to the real-valued poles.

The appropriate recourse would seem to be to remove the noise from the data by suppressing the associated periodogram ordinates in the interval $(\pi/4, \pi]$. When this is done, the estimation does deliver a pair of conjugate complex poles. However, in this case, the parametric spectrum in Figure 3 misrepresents the periodogram in another way. The poles are too close to the unit circle, i.e. their modulus is close to unity. The effect is to exaggerate the prominence of the spectral spike and to underestimate of the rate of damping. Also, it can be seen that the peak is displaced to the right, implying that the argument is an overestimate, which exaggerates of the frequency of the cycles.

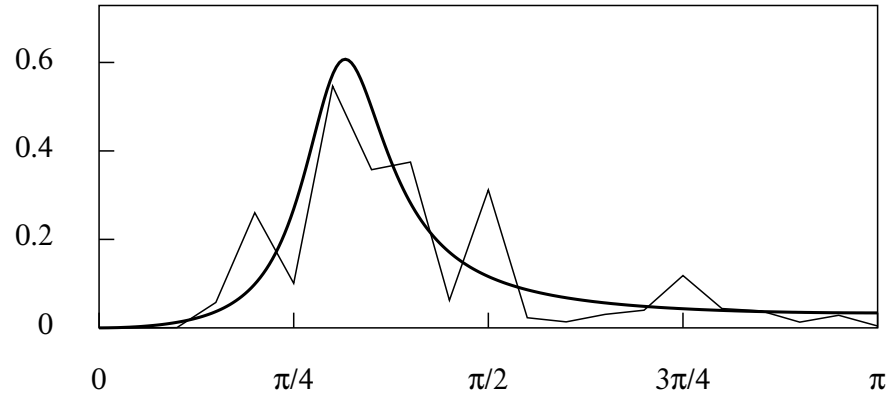



**Figure 4.** The periodogram of the de-noised data that have been filtered and subsampled at the rate of 1 observation in 4, overlaid by the parametric spectrum of an estimated ARMA(2, 1) model.

The spectral support of an ARMA process is the full Nyquist frequency interval $[0, \pi]$. Therefore, it is appropriate to dilate the spectral signature of the business cycle so that it fills the entire interval. This entails associating to each of the periodogram ordinates a higher frequency value. The frequencies are measured relative to the sampling interval. Therefore, they can be increased by increasing the length of the sampling interval.

In order to resample the data, it is necessary to reconstitute the underlying continuous trajectory. This can be achieved by a method of Fourier synthesis based on a version of equation (1) in which the coefficients associated with noisy elements, with frequencies in excess of the upper limit of the business cycle, have been set to zero.

The discrete temporal index, which is $t = 0, 1, \ldots, T-1$, can be replaced within equation (1) by a continuous variable $t \in [0, T)$ to create the continuous trajectory. This can be resampled at intervals of $\pi/\omega_c$ units of time. In the present example, wherein $\omega_c = \pi/4$, the appropriate sampling interval is 4 units, which implies that only one in 4 of the points from the de-noised data is required; and there is no need to reconstitute the continuous trajectory in order to resample it. The effect of estimating an ARMA model with the de-noised and resampled data is shown in Figure 4.

## 3. Sinc Function Interpolation and Fourier Interpolation

The procedure for resampling the data has implicitly defined a continuous ARMA process powered by a continuous frequency-limited white-noise process. The stochastic differential equations that are commonly supposed to be the continuous-time analogues of the ARMA models are driven by the increments of a Wiener process. The latter is an accumulation of a continuous steam of infinitesimal impulses. Such impulses are unbounded in frequency. The Wiener process has the characteristic that, whatever the rate of sampling, the accumulations that occur within the sampling intervals will constitute a discrete-time white-noise process.

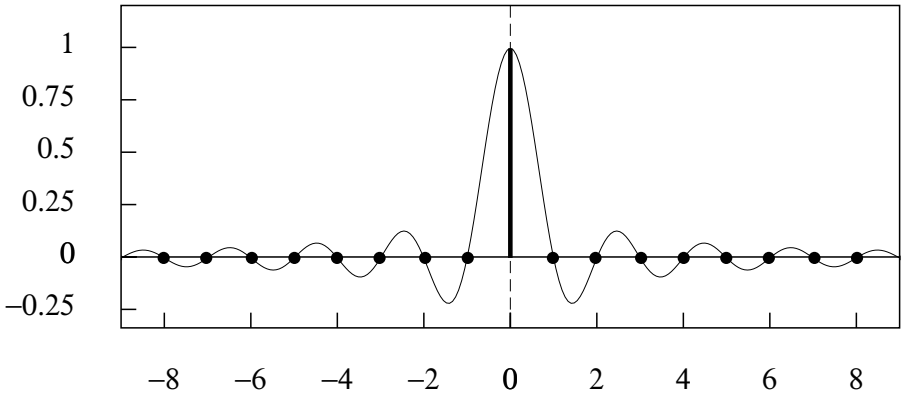



**Figure 5.** The sinc function wave-packet $\varphi(t) = \sin(\pi t)/\pi t$ comprising frequencies in the interval $[0, \pi]$.

In proposing a frequency-limited white noise, we resort to the sampling theorem. The theorem is commonly defined for square-integrable functions of time, defined of the the real line $\mathcal{R} = (-\infty, \infty)$, and limited in frequency to the interval $[-\pi, \pi]$.

The Fourier integral transform has the following expression in the time domain and the frequency domain:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \xi(\omega)e^{\mathrm{i}\omega t}d\omega \longleftrightarrow \xi(\omega) = \int_{-\infty}^{\infty} x(t)e^{-\mathrm{i}\omega t}dt. \tag{5}$$

However, with the frequency limitation, this becomes

$$x(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \xi_S(\omega)e^{\mathrm{i}\omega t}d\omega \longleftrightarrow \xi_S(\omega) = \sum_{k=-\infty}^{\infty} x_k e^{-\mathrm{i}k\omega}, \tag{6}$$

where $\{x_k; k = 0, \pm 1, \pm 2, \ldots\}$ is sampled at unit intervals from $x(t)$. Putting the RHS of (6) into the LHS and interchanging the order of integration and summation gives

$$x(t) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} x_k \left\{\int_{-\pi}^{\pi} e^{\mathrm{i}\omega(t-k)}\right\} d\omega = \sum_{k=-\infty}^{\infty} x_k \varphi(t-k), \tag{7}$$

where

$$\varphi(t-k) = \frac{\sin\{\pi(t-k)\}}{\pi(t-k)} \tag{8}$$

is the so-call sinc function. The RHS of equation (7) defines a sinc function interpolation.

The sinc function centred on $k = 0$, which is illustrated in Figure 5, is formed by applying a bi-directional hyperbolic taper to an ordinary sine function. The succession of displaced sinc functions provides an orthonormal basis for the set of continuous functions that are limited in frequency to the Nyquist interval $[-\pi, \pi]$.

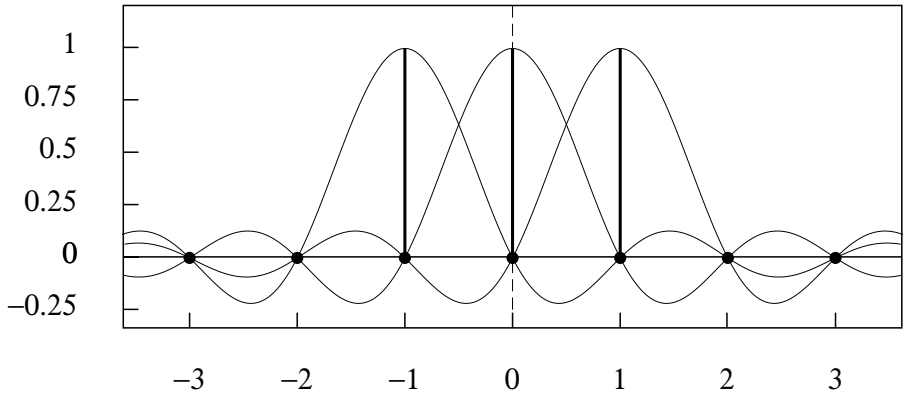



**Figure 6.** The wave packets $\varphi(t-k)$, which are bounded in frequency by $\pi$, suffer no mutual interference when $k \in \{0, \pm 1, \pm 2, \pm 3, \ldots\}$.

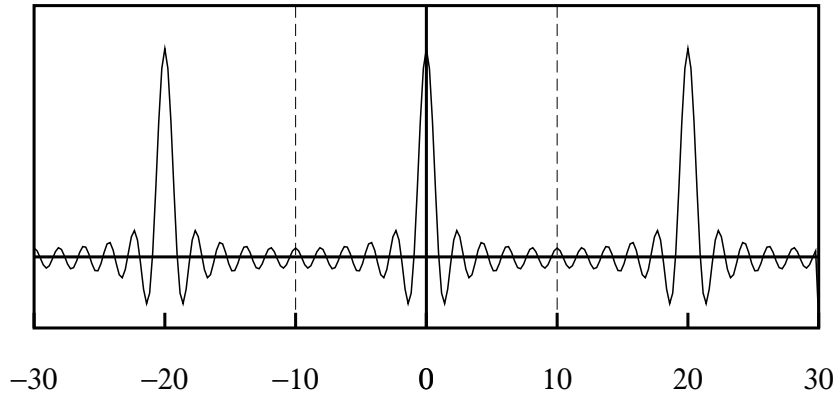



**Figure 7.** The Dirichlet function $\sin(\pi t)/\sin(2\pi t/M)$ obtained from the inverse Fourier transform of a frequency-domain rectangle sampled at $M = 21$ points.

Equation (7) implies a simple prescription for converting a data sequence into a continuous function that is limited in frequency to the Nyquist interval. Sinc function kernels are attached to each of the discrete-time ordinates, and the sum is taken of the scaled kernels. The values at the integer points are those of their associated kernels; and these values are not affected by the kernels at the other integer points. This feature is illustrated by Figure 6.

A continuous-time white-noise forcing function can be formed by replacing the impulses of a discrete-time white-noise process by sinc functions scaled by the values of those impulses. If $\varepsilon_t = \varepsilon(t)$ and $\varepsilon_s = \varepsilon(s)$ are elements sampled at arbitrary points from the continuous frequency-limited white-noise forcing function, then their covariance is the sinc function

$$C(\varepsilon_t, \varepsilon_s) = \sigma_\varepsilon^2 \varphi(t-s) = \sigma_\varepsilon^2 \varphi(\tau), \quad \tau = t - s, \tag{9}$$

where $\sigma_\varepsilon^2$ is the variance parmeter. This result follows from recognising that $\varepsilon_s = \varepsilon(\tau)\varepsilon_t + \eta$, where $\eta$ is uncorrelated with $\varepsilon_t$, and from the fact that $\varphi^2(\tau) = \varphi(\tau)$.

The practicality of a sinc function synthesis is prejudiced by the fact that the support of the kernel functions is the entire real line $\mathcal{R} = (-\infty, \infty)$. A practical

synthesis replaces the sinc function by the so-called Dirichlet kernel, which is a periodic or circular function formed by wrapping the sinc function around a circle of a circumference $T$, equal to the number of data points, and by adding the overlying ordinates. In this context, the data points to which the kernels are to be fixed are also to be regarded as a periodic or circular sequence.

Consider the discrete Fourier transform expressed as follows:

$$x_t = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t} \longleftrightarrow \xi_j = \frac{1}{T}\sum_{t=0}^{T-1} x_t e^{-\mathrm{i}\omega_j t} \quad \text{with} \quad \omega_j = \frac{2\pi j}{T}. \tag{10}$$

By putting the RHS of the latter into the LHS and commuting the two summations and allowing $t \in [0, T)$ to vary continuously, we get

$$x(t) = \frac{1}{T}\sum_{k=0}^{T-1} x_k \left\{ \sum_{j=0}^{T-1} e^{\mathrm{i}\omega_j (t-k)} \right\} = \sum_{k=0}^{T-1} x_k \varphi^\circ(t-k). \tag{11}$$

where

$$\varphi^\circ(t) = \frac{1}{T}\sum_{j=-n}^{n} e^{\mathrm{i}\omega_j t} = \frac{\sin(\{T/2\}\omega_1 t)}{T\sin(\omega_1 t/2)} \quad \text{with} \quad n = \frac{T-1}{2}. \tag{12}$$

is the periodic Dirichlet Kernel. An example is provided by Figure 7.

Equation (11) implies that a sinc function interpolation of a finite data sequence that employs a sequence of Dirichelet kernels is equivalent to an interpolation based on a Fourier synthesis.

## 4. Discrete-time and Continuous-time ARMA Processes

Whereas it is straightforward to derive a continuous version of an ARMA process (i.e. a CARMA process) by sinc function interpolation, we also require to represent it via a linear stochastic differential equation (an LSDE). The correspondence between difference equations and differential equations can be established by focusing, initially, on the first-order equations.

(In this paper, the acronym CARMA is reserved for the continuous-time linear stochastic differential equations that have the same frequency limitation as their corresponding discrete-time ARMA models. This is in spite of the common use of the acronym to denote continuous processes of unlimited frequency that are derived from ARMA models.)

The first-order autoregressive difference equation takes the form of

$$y(t) = \mu y(t-1) + \varepsilon(t) \qquad \text{or} \qquad y(t) = \frac{\varepsilon(t)}{1-\mu L} = \sum_{\tau=0}^{\infty} \mu^\tau \varepsilon(t-\tau). \tag{13}$$

Here, $y(t) = \{y_t; t = 0 \pm 1, \pm 2 \ldots\}$ denotes a sequence, and $L$ is the lag operator such that $Ly(t) = y(t-1)$. (However, $y(t)$ will be used, equally, to denote a function of a continuous-time index.) Also, the forcing function $\varepsilon(t)$ is a white-noise sequence of independent and identically distributed random elements.

The corresponding first-order stochastic differential equation is denoted by

$$\begin{aligned} \frac{dy}{dt} &= \kappa y(t) + \zeta(t) \quad \text{or} \\ y(t) = \frac{\zeta(t)}{D-\kappa} &= \int_0^\infty e^{\kappa\tau}\zeta(t-\tau)d\tau = \int_{-\infty}^t e^{\kappa(t-\tau)}\zeta(\tau)d\tau, \end{aligned} \tag{14}$$

where $D$ is the derivative operator such that $Dx(t) = dx/dt$. Here, the forcing function $\zeta(t)$ is a continuous frequency-limited white-noise process, formed by associating sinc functions to the elements of a discrete white-noise sequence. It can be seen that $\mu^\tau$ and $e^{\kappa\tau}$ play the same role in the two equations, which is to diminish or to 'dampen' the effect of the impulses of the forcing functions as time elapses.

To convert the differential equation of (14) to the difference equation of (13), the integral on the interval $(-\infty, t]$ may be separated into two parts, which are the integrals over $(-\infty, t-1]$ and $(t-1, t]$:

$$\begin{aligned} y(t) &= e^\kappa \int_{-\infty}^{t-1} e^{\kappa(t-1-\tau)}\zeta(\tau)d\tau + \int_{t-1}^t e^{\kappa(t-\tau)}\zeta(\tau)d\tau \\ &= \mu y(t-1) + \varepsilon(t). \end{aligned} \tag{15}$$

We are interested, of course, in equations of higher orders. The ARMA$(p, q)$ equation is denoted by

$$\begin{aligned} (1 + \alpha_1 L + \cdots + \alpha_p L^p)y(t) &= (\beta_0 + \beta_1 L + \cdots + \beta_p L^q)\varepsilon(t) \\ \text{or} \qquad \alpha(L)y(t) &= \beta(L)\varepsilon(t). \end{aligned} \tag{16}$$

Given that $p > q$ and that there are no repeated roots of $\alpha(z) = 0$, the rational function $\beta(z)/\alpha(z)$ is amenable to a partial-fraction decomposition, which gives rise to the equation

$$\begin{aligned} y(t) = \frac{\beta(L)}{\alpha(L)}\varepsilon(t) &= \left\{ \frac{d_1}{1-\mu_1 L} + \frac{d_2}{1-\mu_2 L} + \cdots + \frac{d_p}{1-\mu_p L} \right\} \varepsilon(t) \\ &= \sum_{\tau=0}^{\infty} \left\{ d_1\mu_1^\tau + d_2\mu_2^\tau + \cdots + d_p\mu_p^\tau \right\} \varepsilon(t-\tau). \end{aligned} \tag{17}$$

The linear stochastic differential equation of orders $p$ and $q < p$, denoted by LSDE$(p, q)$, is specified by the equation

$$\begin{aligned} (\phi_0 D^p + \phi_1 D^{p-1} + \cdots + \phi_p)y(t) &= (\theta_0 D^q + \theta_1 D^{q-1} + \cdots + \theta_q)\zeta(t) \\ \text{or} \qquad \phi(D)y(t) &= \theta(D)\zeta(t). \end{aligned} \tag{18}$$

On the assumption that there are no repeated roots, it has the following partial-fraction decomposition:

$$\begin{aligned} y(t) = \frac{\theta(D)}{\phi(D)}\zeta(t) &= \left\{ \frac{c_1}{D-\kappa_1} + \frac{c_2}{D-\kappa_2} + \cdots + \frac{c_p}{D-\kappa_p} \right\} \zeta(t) \\ &= \int_0^\infty \left\{ c_1 e^{\kappa_1\tau} + c_2 e^{\kappa_2\tau} + \cdots + c_p e^{\kappa_p\tau} \right\} \zeta(t-\tau)d\tau. \end{aligned} \tag{19}$$

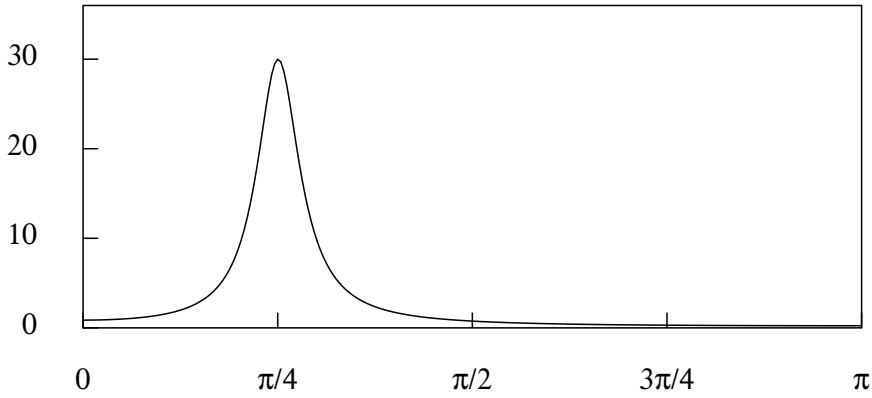



**Figure 8.** The spectrum of the ARMA(2, 1) process $(1.0 - 1.273L + 0.81L^2)y(t) = (1 - 0.5L)e(t)$.

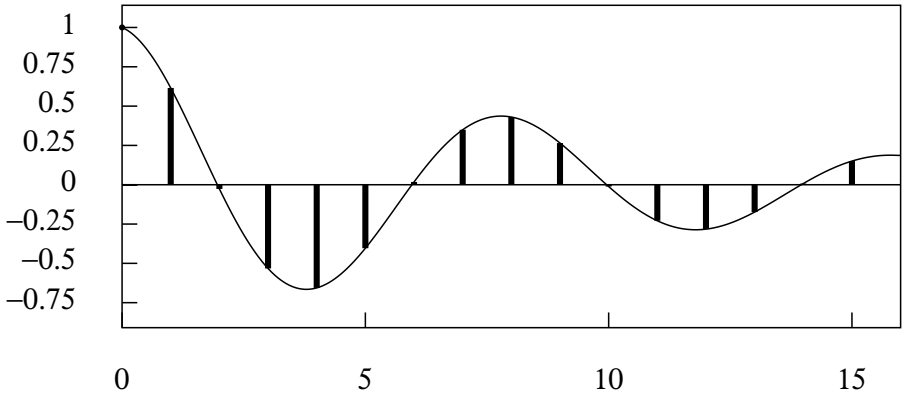



**Figure 9.** The discrete autocovariance sequence of the ARMA(2, 1) process and the continuous autocovariance function of the corresponding CARMA(2, 1) process.

A correspondence can be established between the discrete and continuous systems by invoking the principle of impulse invariance. This indicates that a sequence sampled at unit intervals from the impulse response function of the continuous system should be equal to the impulse response of the discrete-time system. This is only possible if the continuous system has the same frequency limitation as the discrete system, which is the present assumption.

Thus, at the integer values of $\tau$, the functions

$$\psi(t) = c_1 e^{\kappa_1 \tau} + c_2 e^{\kappa_2 \tau} + \cdots + c_p e^{\kappa_p \tau} \tag{20}$$

and

$$\phi(\tau) = d_1 \mu_1^\tau + d_2 \mu_2^\tau + \cdots + d_p \mu_2^\tau \tag{21}$$

should be equal. The equality can be achieved by setting $e^{\kappa_j} = \mu_j$ and $c_j = d_j$, for all $j$. The discrete-time ARMA model is driven by a white-noise process that is limited in frequency by the Nyquist value of $\pi$ radians per sample interval. Its direct continuous-time counterpart is a CARMA model, driven by a continuous frequency-limited white-noise process.

It is appropriate to adopt a CARMA model when there is clear evidence that the spectral density of the process is limited in frequency by the Nyquist value of

$\pi$ radians per sample interval, at which point the function should be zero-valued. The evidence will be provided by the periodogram of the data. In cases where the limiting frequency of the process is less the $\pi$, the resampling procedures outlined in section 2 should be pursued before estimating the ARMA model.

**Example 1.** To illustrate the mapping from the discrete-time ARMA model to a continuous frequency-limited CARMA model, an ARMA(2, 1) model is chosen with conjugate complex poles $\alpha \pm \mathrm{i}\beta = \rho \exp\{\pm \mathrm{i}\theta\}$, where $\rho = \sqrt{\alpha^2 + \beta^2} = 0.9$ and $\theta = \tan^{-1}(\beta/\alpha) = \pi/4 = 45^\circ$. The moving-average component has a zero of 0.5. The ARMA process generates prominent cycles of an average duration of roughly 8 periods.

The parameters of the resulting continuous-time CARMA model are displayed below, beside those of the ARMA model:

| ARMA | CARMA |
|---|---|
| $\alpha_0 = 1.0$ | $\phi_0 = 1.0$ |
| $\alpha_1 = -1.2728$ | $\phi_1 = 0.2107$ |
| $\alpha_2 = 0.8100$ | $\phi_2 = 0.6280$ |
| | |
| $\beta_0 = 1.0$ | $\theta_0 = 1.0$ |
| $\beta_1 = -0.5$ | $\theta_1 = 0.2737$ |

The spectral density function of the ARMA process is illustrated in Figure 8. Here, it will be observed that the function is virtually zero at the limiting Nyquist frequency of $\pi$. Therefore, it is reasonable to propose that the corresponding continuous-time model should be driven by a white-noise forcing function that is bounded by the Nyquist frequency.

The spectral density function of the CARMA process is the integral Fourier transform of the continuous autocovariance function, whereas the spectral density function of the ARMA process is the discrete Fourier transform of the autocovariance sequence. The frequency limitation of the CARMA process means that there is no aliasing in the sampling process. Therefore, the two spectra are identical.

In Figure 9, the discrete autocovariance function of the ARMA process is superimposed on the continuous autocovariance function of the CARMA process. The former has been generated by a recursive procedure. The latter has been generated by an analytic equation, to be presented below as equation (22), wherein the index $\tau$ of the lags varies continuously.

## 5. Stochastic Differential Equations Driven by Wiener Processes

The white-noise forcing function of a conventional linear stochastic differential equation (LSDE) is the derivative of a Wiener process. The latter process consists of a continuous steam of infinitesimal impulses. Since a pure impulse is unbounded in frequency, so too is the forcing function.

The concept of a pure impulse is problematic from a physical point of view, since it implies a discrete and instantaneous change in momentum. The problem of unbounded frequencies can be mitigated, if not completely overcome, in the

context of an LSDE, since its transfer function may impose a sufficient attenuation on the higher frequencies for the effect to be a virtual frequency limitation.

Whenever the spectral density function of an ARMA model has a significant value at the Nyquist frequency of $\pi$, there can be a reasonable supposition that the underlying continuous process has a frequency range that extends beyond the Nyquist limit. Therefore, it may be appropriate to adopt an LSDE with an unbounded forcing function as the continuous-time counterpart of the ARMA model.

In translating from an ARMA model to such an LSDE, it is no longer appropriate to invoke the principle of impulse invariance. Instead, the principle of autocovariance equivalence that was enunciated by Bartlett (1946) must be adopted. The principle asserts that the parameters of the LSDE should be chosen so that its autocovariance function matches the autocovariance function of the ARMA model at the integer lags.

The autocovariance function of an ARMA model can be derived from its impulse response function, represented by equation (21). It takes the form of

$$\begin{aligned}\gamma^d(\tau) &= \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \left( \sum_{k=1}^{p} d_k \mu_k^j \right) \left( \sum_{k=1}^{p} d_k \mu_k^{j+\tau} \right) \\ &= \sigma_\varepsilon^2 \sum_{i=1}^{p} \left\{ \sum_{j=1}^{p} \frac{d_i d_j}{1 - \mu_i \mu_j} \right\} \mu_i^\tau .\end{aligned} \tag{22}$$

The autocovariance function $\gamma^c(\tau)$ of the continuous-time LSDE process is also found via its impulse response function. It is assumed that the autocovariance of the white-noise forcing function at lag $\tau$ is

$$E\{\zeta(t)\zeta(t-\tau)\} = \delta(\tau)\sigma_\zeta^2, \tag{23}$$

where $\delta(\tau)$ is Dirac's delta function. Then,

$$\begin{aligned}\gamma^c(\tau) &= E\{y(t)y(t-\tau)\} \\ &= E\left\{ \int_0^\infty \psi(u)\zeta(t-u)du \int_0^\infty \psi(v)\zeta(t-\tau-v)dv \right\} \\ &= \sigma_\zeta^2 \int_0^\infty \psi(v)\psi(v+\tau)dv.\end{aligned} \tag{24}$$

Substituting the expression of (20) for the continuous-time impulse response function $\psi(t)$ into equation (24) gives

$$\begin{aligned}\gamma^c(\tau) = \sigma_\zeta^2 \int_0^\infty \psi(t)\psi(t+\tau)dt &= \sigma_\zeta^2 \sum_i \sum_j \left\{ c_i c_j \int_0^\infty e^{(\kappa_i+\kappa_j)t+\kappa_i\tau} dt \right\} \\ &= \sigma_\zeta^2 \sum_i \left\{ \sum_j c_i c_j \frac{-e^{\kappa_i \tau}}{\kappa_i + \kappa_j} \right\}.\end{aligned} \tag{25}$$

This expression, which is liable to contain complex-valued terms, may be rendered in real terms by coupling the various conjugate complex terms.

In translating from the ARMA model to the LSDE, we may exploit the one-to-one correspondence between the poles of the two systems. If a complex pole of the ARMA model takes the form of

$$\mu = \alpha + \mathrm{i}\beta = \rho\{\cos(\omega) + \mathrm{i}\sin(\omega)\} = \rho e^{\mathrm{i}\omega}, \tag{26}$$

with

$$\rho = \sqrt{\alpha^2 + \beta^2} \quad \text{and} \quad \omega = \tan^{-1}\left(\frac{\beta}{\alpha}\right), \tag{27}$$

then the corresponding pole of the LSDE and of the CARMA differential equation is

$$\kappa = \ln(\mu) = \ln(\rho) + \mathrm{i}\omega = \delta + \mathrm{i}\omega, \tag{28}$$

with $\delta \in (-\infty, 0)$, which puts it in the left half of the $s$-plane, as is necessary for the stability of the system.

The principle of autocovariance equivalence can be expressed via the equation

$$\gamma_\tau^c\{\kappa(\mu), c\} = \gamma_\tau^d(\mu, d) \quad \text{for} \quad \tau \in \{0, \pm 1, \pm 2, \ldots\}. \tag{29}$$

Then, the parameters of the LSDE can be derived once a value of $c = [c_1, c_2, \ldots, c_p]$ of the vector of the numerator parameters of (19) has been found that satisfies this equation. The value of $c$ can be found by using an optimisation procedure to find the zeros of the function

$$z(c) = \sum_{\tau=0}^{p}\{\gamma_\tau^c(c) - \gamma_\tau^d\}^2. \tag{30}$$

As Söderström (1990, 1991), and others have noted, there are ARMA models for which there are no corresponding LSDE's. The present procedure for translating from an ARMA model to an LSDE reveals such cases by its failure to find a zero-valued minimum of the criterion function. However, it can be relied upon to find the LSDE most closely related to the ARMA model.

The principle of autocovariance equivalence also indicates a way in which an ARMA model can be found to correspond to a given LSDE. The ARMA model is commonly described as the exact or equivalent discrete linear model (EDLM).

The autocovariance generating function of an ARMA model is

$$\gamma^d(z) = \sigma_\varepsilon^2 \frac{\beta(z)\beta(z^{-1})}{\alpha(z)\alpha(z^{-1})}, \tag{31}$$

whereas the $z$-transform of the elements $\gamma_\tau^c; \tau \in \{0, \pm 1, \pm 2, \ldots\}$ sampled from the autocovariance function of the LSDE may be denoted by $\gamma^c(z)$. Putting the latter in place of $\gamma^d(z)$ and rearranged the equation gives

$$\sigma_\varepsilon^2 \beta(z)\beta(z^{-1}) = \alpha(z)\gamma^c(z)\alpha(z^{-1}). \tag{32}$$

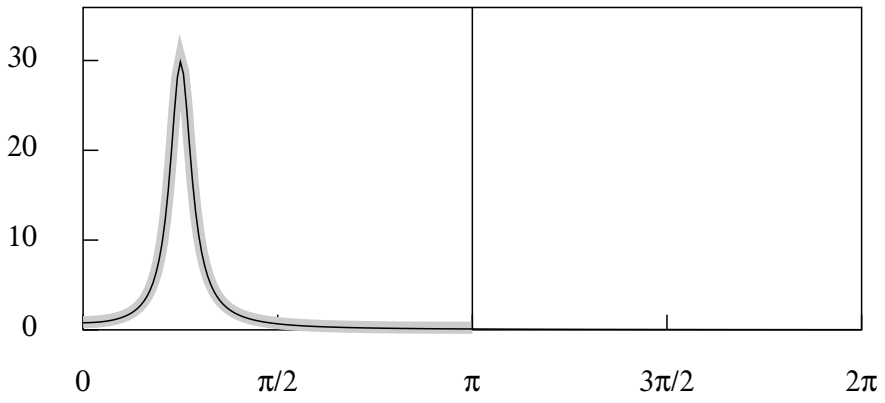



**Figure 10.** The spectrum of the LSDE(2, 1) corresponding to the ARMA(2, 1) model of Example 1 plotted on top of the spectrum of that model, represented by the thick grey line. The two spectra virtually coincide over the interval $[0, \pi]$.

Since the discrete-time autoregressive parameters within $\alpha(z)$ can be inferred from those of the LSDE, only the moving-average parameters within $\beta(z)$ and the variance $\sigma_\varepsilon^2$ need to be derived from equation (32). They can be obtained via a Cramér–Wold factorisation of the LHS.

**Example 2.** The mapping from the discrete-time ARMA model to a continuous-time LSDE model can be illustrated, in the first instance, with the ARMA(2, 1) model of Example 1.

The parameters of the corresponding LSDE(2, 1) model are obtained by using the procedure of Nelder and Mead (1965) to find the minimum of the criterion function of (30), where it is assumed that the variance of the forcing function is $\sigma_\zeta^2 = 1$ . The minimands $a, b$ of the criterion function are from the numerator coefficients $c, c^* = a \pm \mathrm{i}b$ of the partial-fraction decomposition of the LSDE (2, 1) transfer function.

There are four points that correspond to zero-valued minima, where the ordinates of the discrete and continuous autocovariance functions coincide at the integer lags. These points, together with the corresponding moving-average parameters, are as follows:

| | $a$ | $b$ | $\theta_0$ | $\theta_1$ |
|---|---|---|---|---|
| (1) | −0.4544 | 0.2956 | −0.9088 | 0.5601 |
| (2) | 0.4544 | 0.4175 | 0.9088 | 0.5601 |
| (3) | −0.4544 | −0.4174 | −0.9088 | −0.5601 |
| (4) | 0.4544 | −0.2956 | 0.9088 | −0.5601 |

Here, the parameter values of (1) and (4) are equivalent, as are those of (2) and (3). Their difference is a change of sign, which can be eliminated by normalising $\theta_0$ at unity and by adjusting variance of the forcing function accordingly.

The miniphase condition, which corresponds to the invertibility condition of a discrete-time model, requires the zeros to be in the left half of the $s$-plane. Therefore, (2) and (3) on the *NE–SW* axis are the chosen pair.

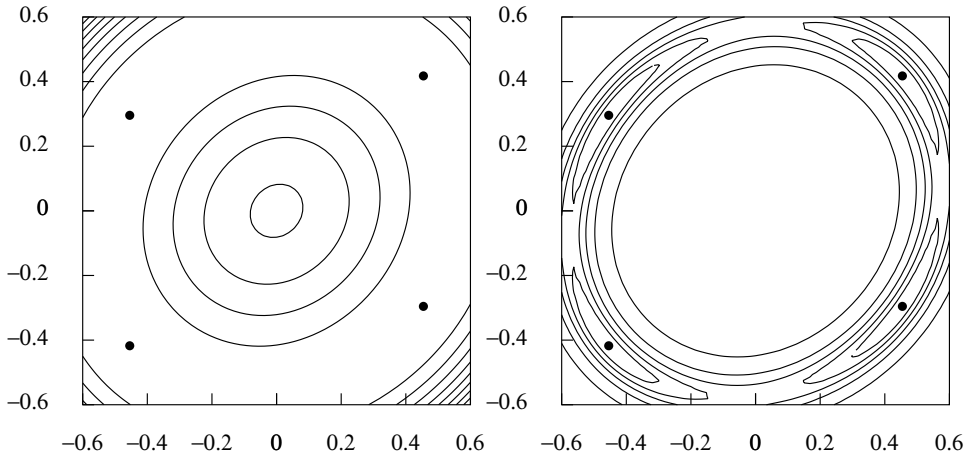



**Figure 11.** *Left* The contours of the criterion function $z = z(a, b)$ together with the minimising values, marked by black dots. *Right* The contours of the function $q = 1/(z + d)$.

These estimates of the LSDE(2, 1) are juxtaposed below with those of the CARMA(2, 1) model derived from the same ARMA model:

| CARMA | LSDE |
|---|---|
| $\phi_0 = 1.0$ | $\phi_0 = 1.0$ |
| $\phi_1 = 0.2107$ | $\phi_1 = 0.2107$ |
| $\phi_2 = 0.6280$ | $\phi_2 = 0.6280$ |
| | |
| $\theta_0 = 1.0$ | $\theta_0 = 0.9088$ |
| $\theta_1 = 0.2737$ | $\theta_1 = 0.5601$ |

The autoregressive parameters of the CARMA model and of the LSDE model are, of course, identical. However, there is a surprising disparity between the two sets of moving-average parameters. Nevertheless, when they are superimposed on the same diagram—which is Figure 10—the spectra of the two models are seen virtually to coincide. Moreover, the parameters of the ARMA model can be recovered exactly from those of the LSDE by an inverse transformation.

The explanation for this outcome is to be found in the remarkable flatness of the criterion function in the vicinity of the minimising points, which are marked on both sides of Figure 11 by black dots. The flatness implies that a wide spectrum of the parameter values of the LSDE will give rise to almost identical autocovariance functions and spectra.

The left side of Figure 11 shows some equally-spaced contours of the $z$-surface of the criterion function, which are rising from an annulus that contains the minima. The minima resemble small indentations in the broad brim of a hat.

The right side of Figure 11, which is intended to provide more evidence of the nature of the criterion function in the vicinity of the minima, shows the contours of the function $q = 1/(z + d)$, where $d$ is a small positive number that prevents a division by zero. We set $d = (X - RM)/(R - 1)$, where $M = \min(z)$, $X = \max(z)$ and where $R = \max(q)/\min(q) = 60$. The extended lenticular contours

surrounding the minimising points of the criterion function, which have become maxima in this diagram, are a testimony to the virtual equivalence of a wide spectrum of parameter values.

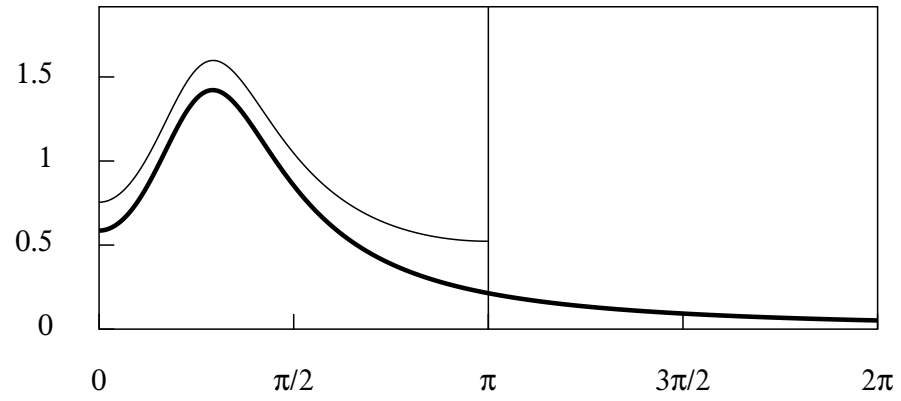



**Figure 12.** The spectrum of the revised ARMA model superimposed on the spectrum of the derived LSDE, described by the heavier line.

**Example 3.** A variant to the ARMA(2, 1) model is one that has a pair of complex conjugate poles $\rho \exp\{\pm i\theta\}$ with the same argument as before, which is $\theta = \pi/4 = 45°$, and with a modulus that has been reduced to $\rho = 0.5$. The model retains the zero of 0.5. The ARMA parameters and those of the corresponding LSDE are as follows:

| ARMA | LSDE |
|---|---|
| $\alpha_0 = 1.0$ | $\phi_0 = 1.0$ |
| $\alpha_1 = -0.7071$ | $\phi_1 = 1.3863$ |
| $\alpha_2 = 0.2500$ | $\phi_2 = 1.0973$ |
| | |
| $\beta_0 = 1.0$ | $\theta_0 = 1.5012$ |
| $\beta_1 = -0.5$ | $\theta_1 = 0.8905$ |

Figure 12 shows the spectral density functions of the LSDE and of the ARMA model superimposed on same diagram. The spectrum of the LSDE extends far beyond the Nyquist frequency of $\pi$, which is the limiting ARMA frequency.

The ARMA process, which is to be regarded as a sampled version of the LSDE, is seen to suffer from a high degree of aliasing, whereby the spectral power of the LSDE that lies beyond the Nyquist frequency is mapped into the Nyquist interval $[-\pi, \pi]$, with the effect that the profile of the ARMA spectrum is raised considerably. On this basis, it can be asserted that the ARMA model significantly misrepresents the underlying continuous-time process.

## 6. Summary and Conclusions

The intention of this paper has been to clarify the relationship between unconditional linear stochastic models in discrete and continuous time, and to provide secure means of computing the continuous models. The importance of an awareness

of the frequency-domain characteristics of the forcing functions has been emphasised.

Example 1 has demonstrated a straightforward way of deriving a frequency-limited stochastic differential equation that correspond to a discrete-time ARMA model. This has been described as a continuous-time CARMA model.

This model is a valid representation of the underlying process only if the maximum frequency of that process corresponds to the limiting frequency of the ARMA model, which is $\pi$ radians per sampling interval. To ensure that this is the case, it may be necessary to reconstitute the continuous trajectory and to resample it at a reduced rate.

The forcing function of a conventional linear stochastic differential equation, or LSDE, which consists of the increments of a Wiener process, is unbounded in frequency. This seems to be inappropriate to a model of a frequency-limited process. Nevertheless, the transfer function of the LSDE may impose a radical attenuation on the higher frequencies that implies a virtual frequency limitation. Example 2 has illustrated such a case.

Example 3 has shown the aliasing effects that occur when the forcing function has no frequency limit and when the ARMA transfer function imposes only a weak attenuation on the high-frequency elements. This provides a ready justification for adopting an LSDE as the continuous-time counterpart of the ARMA model.

The spectral density function of the ARMA model will be formed by wrapping the spectrum of the LSDE around a circle of circumference $2\pi$ and by adding the overlying ordinates. In this way, the spectral component of frequencies in excess of the Nyquist value are mapped into the interval $[-\pi, \pi]$ to produce a discrete-time spectrum that may depart significantly from the continuous-time parent spectrum, as represented by the derived LSDE. This is seen in Figure 12.

The methods for translating from an ARMA model to a continuous-time model, which may be a frequency-limited CARMA model or an LSDE model that is without an ostensible frequency restriction, have been realised in the computer program CONCRETE.PAS, which available at the the author's website, where both the compiled program and its code can be found.

An associated program CONTEXT.PAS, which plots the contour map of the surface of the criterion function that is employed in matching the autocovariance function of the LSDE(2, 1) to that of the ARMA(2, 1) model, is also available.

**References.**

Bartlett, M.S., (1946), On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series, *Supplement to the Journal of the Royal Statistical Society,* 8, 27–41.

Hodrick, R.J., and E.C. Prescott, (1980), *Postwar U.S. Business Cycles: An Empirical Investigation,* Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania.

Hodrick R.J., and E.C. Prescott, (1997), Postwar U.S. Business Cycles: An Empirical Investigation, *Journal of Money, Credit and Banking,* 29, 1–16.

Leser, C.E.V., (1961), A Simple Method of Trend Construction, *Journal of the Royal Statistical Society, Series B,* 23, 91–107.

Nelder, J.A., and R. Mead, (1965), A Simplex Method for Function Minimization, *Computer Journal,* 7, 308–313.

Nyquist, H., (1924), Certain Factors Affecting Telegraph Speed, *Bell System Technical Journal* 3, 324–346.

Nyquist, H., (1928), Certain Topics in Telegraph Transmission Theory, *Transactions of the AIEE,* 47, 617–644. Reprinted in 2002, *Proceedings of the IEEE,* 90, 280–305.

Pollock D.S.G., (2017), Linear Stochastic Models in Discrete and Continuous Time, *Unpublished paper.*

Shannon, C.E., (1949), Communication in the Presence of Noise, *Proceedings of the Institute of Radio Engineers,* 37, 10–21. Reprinted in 1998, *Proceedings of the IEEE,* 86, 447–457.

Söderström, T., (1990), On Zero Locations for Sampled Stochastic Systems, *IEEE Transactions on Automatic Control,* 35, 1249–1253.

Söderström, T., (1991), Computing Stochastic Continuous-Time Models from ARMA Models, *International Journal of Control,* 53, 1311–1326.

# Multifractal Detrended Fluctuation Analysis combined with Singular Spectrum Analysis

Anton Karmatskii

Ural Federal University, 620026 Yekaterinburg, Russia
e-mail: karmatsky.anton@urfu.ru

**Abstract.** This paper addresses the problem of finding appropriate method for analysis of non-stationary time series with complex trends and possibly with distinct periodic components in power spectrum. A well established Multifractal Detrended Fluctuation Analysis (MDFA) combined with Singular Spectrum Analysis (SSA) used on several artificial examples to demonstrate the capabilities of the method. In this combined method SSA is used for nonparametric periodic components extraction and for adaptive automatic nonparametric trend extraction. As usual, the local fractal characteristics of the signal are studied by utilizing the MFDFA algorithm. It is shown that combined method is capable of accurately extracting fractal features when signal is contaminated with broadband and narrowband noise. The main task that the author focused on when creating the method was the analysis of currents through ion channels in cell membrane.

**Keywords:** SSA • DFA •nonparametric

## 1 Introduction

The traditional fractal methods, such as fluctuation analysis (FA) [1], are developed as statistical tools to evaluate the fractal characteristics of the time series, most commonly the scaling exponent. However, FA requires stationarity of time series. To overcome this limitation, a new scale exponent calculation method named detrended fluctuation analysis (DFA) was introduced in [2]. DFA is extensively used to detect the long-range correlation and power-law properties in nonlinear and nonstationary time series, and it is suitable for extraction of precise intrinsic statistical features from the time series by removing external polynomial trends of different orders. Furthermore, detrending helps to avoid the spurious detection of correlations which are artifacts of nonstationary time series.

However DFA deals only with so called monofractals. Monofractal is mainly a description of the overall average of the object of study, with the local characteristics of the signal being insufficiently characterized [3]. Multifractal analysis provides the ability to describe the local characteristics of the signal in detail. By combining multifractal and detrended fluctuation analysis, a multifractal detrended fluctuation analysis (MDFA) algorithm [4] was introduced. Since then, MDFA has been quickly

applied in a number of research areas, including turbulence, temperature, stock market time series etc. However, despite its widespread use, it is known that the MDFA has several limitations [3]:

1. When calculating a trend the order of approximating polynomials is not always known a priori and may vary for different lengths of time series. So, several computations with different trends approximations are recommended.
2. The distinct peaks in the power spectrum of the time series (for example, annual harmonic) have a strong influence on the results of the algorithm, so they should be removed by means of appropriate filter.

To solve the above-mentioned problems in present work singular spectrum analysis (SSA) [5, 6] is used. It is a relatively new method of time series analysis that emerged in statistics, as well as in nonlinear dynamics and in classical spectral methods. The method is nonparametric, i.e. does not require an explicit specification of the time series model. SSA can be used for a wide range of tasks: trend or quasi-periodic component detection and extraction, denoising, forecasting, change-point detection.

There are several modifications of original MDFA in terms of trend extraction [3], such as moving average or empirical mode decomposition (EMD), but approach with SSA has several advantages [6]:

- In SSA the type of the extracted trend is determined by the data itself, which reduces the influence of trend features on scaling exponent.
- Ability to control extracted frequencies. Ideally, when dealing with trend extraction in MDFA frequencies $f \leq 1/N$ should be extracted from time series, where N is the number of data points, f is frequency in arbitrary units [3]. In trend extraction by means of SSA such control is part of the algorithm and does not require additional efforts.
- The mathematical theory of SSA guarantees the asymptotic separability of the trend and other components of the time series, such as harmonics or broadband noise.
- For the task of periodic components extraction SSA allow amplitude modulation and, partially, frequency modulation.
- Procedure of trend and periodic components extraction in SSA can be automated [6-8].

Combined algorithm consists of the following steps: first by means of SSA remove narrowband (distinct harmonics) and broadband noise from time series by means of periodic and trend extraction respectively. Then apply MDFA to resulted time series where parametric trend extraction replaced by nonparametric. All required code was written in C++ language.

The remainder of this paper is organized as follows: the principle of SSA, MDFA and combined method are described in Section 2. Section 3 describes test experiments and their results. Finally, conclusions and future directions are given in Section 4.

## 2 Theory Descriptions

### 2.1 Singular Spectrum Analysis (SSA)

Suppose we have a time series

$$F = (f_0, f_1, \dots, f_{N-1}) \tag{1}$$

of lenght N, it is assumed that the measurements are carried out at equal time intervals. The first step of the SSA algorithm is the so-called embedding, when a one-dimensional time series is converted to a two-dimensional matrix. The main parameter of embedding is window length 1 < m < N, which is a number of rows in constructed matrix X:

$$\mathrm{X} = \begin{pmatrix} f_0 & f_1 & \dots & f_{N-m} \\ f_1 & f_2 & & f_{N-m+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m-2} & f_{m-1} & \dots & f_{N-2} \\ f_{m-1} & f_m & & f_{N-1} \end{pmatrix}. \tag{2}$$

Resulted matrix is Hankel matrix [6].

The next step is the Singular Value Decomposition [9] of the matrix X:

$$X = X_1 + X_2 + \cdots + X_d, \tag{3}$$

where d is the rank of the matrix X,

$$X_k = \frac{1}{\sigma_k} \cdot u_k \cdot v_k^T$$

is the k-th component of singular value decomposition, $\sigma_k$ is the k-th singular value, which is always nonnegative, $u_k$ is the k-th left singular vector, $v_k^T$ is the k-th transposed right singular vector. Set of values $(\sigma_k, u_k, v_k)$ sometimes called eigentriple. Next, one group different components and reconstruct the time series according to the following rule: average all elements in the sub diagonals:

$$\{x_{pq}, p + q = const\},$$

each such average corresponds to the value of the time series. Grouping of different components of singular value decomposition depends on the problem to solved. For example, for trend extraction all eigentriples in which "low frequencies" predominate (in the left or right singular vectors) are groped to produce trend. For distinct harmonic extraction situation is different: one groups all eigentriples with similar singular values and similar predominated frequencies (in the left or right singular vectors).

SSA algorithm can be automated [7, 8], and automated version is used in this work. An effective SSA algorithm, having practically linear complexity along the length of the time series was made in accordance with papers [10, 11].

### 2.2 Multifractal Detrended Fluctuation Analysis (MDFA)

MDFA is suitable for nonlinear and nonstationary time series and consists of several steps. Initially, one need to find the cumulative function of the time series defined in (1):

$$Y_i = \sum_{k=0}^{i}(f_k - \overline{f}), \tag{4}$$

$$\overline{f} = \frac{1}{\mathrm{N}}\sum_{j=0}^{N-1} f_j.$$

The next step is to divide the time series into disjoint segments of length s, their number is $N_s = \frac{N}{s}$. Often the length of the segments s does not divide the full length N without a remainder, so, in order not to discard the remainder of the time series, the procedure of division into segments is repeated from the other end of the data. There are $2N_s$ segments in total.

On each of the $2N_s$ segments a local trend is computed and then the variation is calculated:

$$\Psi^2(s,\nu) = \frac{1}{s}\sum_{i=0}^{s-1}(Y_{(\nu-1)\cdot s+i} - y_i^{\nu})^2, \tag{5}$$

$$\nu = 1, \dots, N_s,$$

$$\Psi^2(s,\nu) = \frac{1}{s}\sum_{i=0}^{s-1}\left(Y_{N-(\nu-N_s)\cdot s+i} - y_i^{\nu}\right)^2, \tag{6}$$

$$\nu = N_s + 1, \dots, 2N_s,$$

where $y_i^{\nu}$ is the local trend on a segment ν of length s, which is often calculated by the least squares method, with possible variants including: linear, quadratic, etc.

Further, the resulting variation on each segment is averaged and a fluctuation function is found:

$$\Psi_q(s) = \sqrt[q]{\frac{1}{2N_s}\sum_{\nu=1}^{2N_s}\Psi^2(s,\nu)^{\frac{q}{2}}}, \tag{7}$$

where the index q can take any values that are not equal to zero.

Following [3], we redefine the fluctuation function (we raise it to the power)

$$\Phi_q(s) = (\Psi_q(s))^{\mathrm{q}}. \tag{8}$$

Obviously, the function (8) increases with increasing length and depends on the order of the extracted trend.

In the final step one tests the hypothesis of fractal properties (scaling properties)

$$\Phi_q(s) \approx \mathrm{s}^{\mathrm{h(q)}}. \tag{9}$$

To test the hypothesis (9), a graph is plotted in the double logarithmic scale of the function $\Phi_q(s)$ for each value of q from a certain range. If condition (9) is satisfied, then the graph will have a straight line, the slope of which will give h(q) for a given q.

The function h(q) in this case is called the generalized Hurst exponent. After the corresponding value of h(q) was found for each q, a graph of the dependence of h(q) on q plotted. If the time series is a monofractal, then the graph will have a straight line, with

$$h(q) = H \cdot (1 + q), \tag{10}$$

where H is the ordinary Hurst exponent. For h(q) to be non-linear (so-called multifractality), it is necessary that small and large fluctuations differ in their statistical properties [4].

### 2.3 Combined MDFA with SSA algorithm

In this paper, it is proposed to replace the parametric method of detrending procedure in the mDFA by nonparametric using the automatic trend extraction by means of SSA. It is also used to automatically extract all periodic components from time series as well as broadband noise before applying MDFA. A full description of the algorithm can be found in [7, 8]. We will briefly describe the algorithm of trend extraction.

We first define disrete Fourier transform and the periodogram of time series (1):

$$\Phi_k = \sum_{n=0}^{N-1} e^{-i \cdot 2 \cdot \pi \cdot n \cdot k/N} f_n,$$

$$I_N\left(\frac{k}{N}\right) = \frac{1}{N}\begin{cases} 2|\Phi_k|, if\ 0 < k < N/2 \\ |\Phi_k|, if\ k = 0\ or\ k = N/2' \end{cases} \tag{10}$$

$$\sum_{n=0}^{N-1} f_n^2 = \sum_{k=0}^{N/2} I_N\left(\frac{k}{N}\right).$$

Then one introduce a measure showing the relative contribution of all frequencies less than the specified one in the periodogram of time series:

$$P_N(\omega) = \sum_{0<\frac{k}{N}<\omega} I_N\left(\frac{k}{N}\right),$$

$$C(\omega_0) = \frac{P_N(\omega)}{P_N(0.5)}, \tag{11}$$

where frequency ω is in arbitrary units and belong to the interval [0; 0.5]. Then all eigentriples $(\sigma_k, u_k, v_k)$ for which the condition on the left singular vectors is satisfied

$$C(\omega_0) \geq \mathrm{C}_0 \tag{12}$$

will be grouped in the trend. In order to separate colored noise from white noise for $\omega_0$ we use the expression [8]:

$$\omega_0 = \max_{\frac{k}{N}, 0 \leq \frac{k}{N} \leq 0.5}\{\frac{k}{N} : I_N(0), I_N\left(\frac{1}{\mathrm{N}}\right), \ldots, I_N\left(\frac{\mathrm{k}}{\mathrm{N}}\right) < M_N\}, \tag{13}$$

where $M_N$ is median of the periodogram. The reasoning behind (13) the following: the median of values of the timeseries periodogram gives an estimation of the median of the values of the noise periodogram. So all frequencies $\omega > \omega_0$ correspond to a broadband noise.

The criterion for selecting threshold values for $\omega_0$ and $C_0$ in condition (12) is described in detail in [7, 8].

## 3 Experiment

To test the technique, a fractional Brownian motion (with several Hurst exponent H) was generated with total of $10^5$ points each. White noise was added to the generated time series with a standard deviation that was two times smaller than for the fractional Brownian motion. In order to simulate narrowband noise two exponentially modulated harmonics with frequencies: 0.05 and 0.2 (in arbitrary units) was also added to the time series.

The resulted time series with H = 0.65 and its periodogram calculated by (10) are presented in Fig. 1. The presence of harmonics and white noise is quite difficult to detect in the time series itself, however, additional patterns are clearly visible on the periodogram: two distinct peaks at frequencies of 0.05 and 0.2, as well as uniform broadband noise at high frequencies.

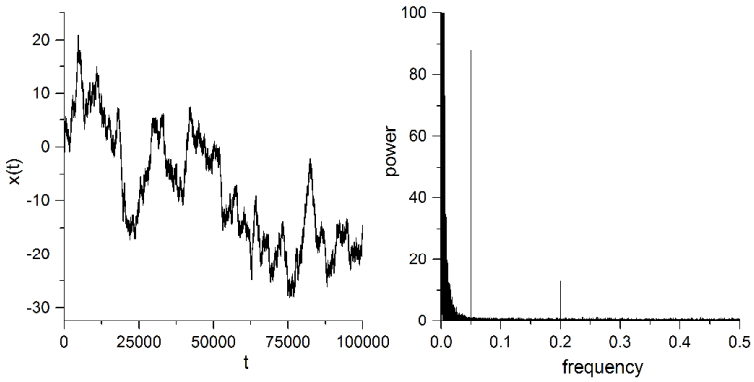



**Fig. 1** Sample of fractional Brownian motion with H = 0.65 (left) and its periodogram (right)

We will first apply SSA to extract the periodic components. The result is shown in Fig. 2. It can be seen that despite the amplitude modulation, the extracted modulated harmonics are in good agreement with their true values.

After narrow band noise have been removed, trend extraction with criterion (13) was applied to remove broadband noise. Fig.3 illustrate the influence of $\omega_0$ to trend extraction procedure in SSA. It is clear that when increasing the low frequencies threshold $\omega_0$ in (12), the trend becomes more detailed.

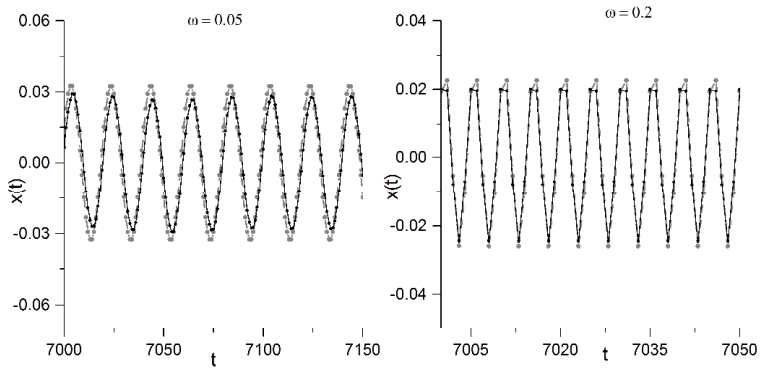



**Fig. 2.** Extracted periodic components (black) and their real values (grey) with frequencies 0.05 (left) and 0.2 (right).

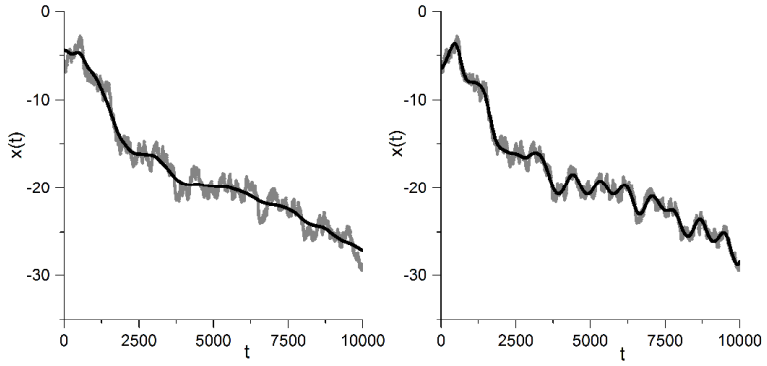



**Fig. 3.** Example of trend extraction (black line) from Brownian motion (grey line) with $\omega_0 = 10^{-5}$ (left) and $\omega_0 = 10^{-3}$ (right), $\omega_0$ in arbitrary units

When both narrowband and broadband noise have been removed MDFA is applied to the resulted time series. To calculate the fluctuation function and to test the hypothesis (9), the values of q are taken from the interval [-10; 10] with step 0.33. The calculation results are presented in Fig. 4 for several values of q in double logarithmic scale. It can be seen that the points are well approximated by a straight line for all values of the partition lengths. However, for large values of q, the deviation of points from a straight line increases.

For the time series shown in Fig.1 Hurst exponent H was calculated using MDFA and was found to be $0.64 \pm 0.02$. Similar results were obtained for the case of generated time series with H = 0.25 and H = 0.5, the computed Hurst exponent is equal to $0.24 \pm 0.01$ and $0.48 \pm 0.02$ respectively.

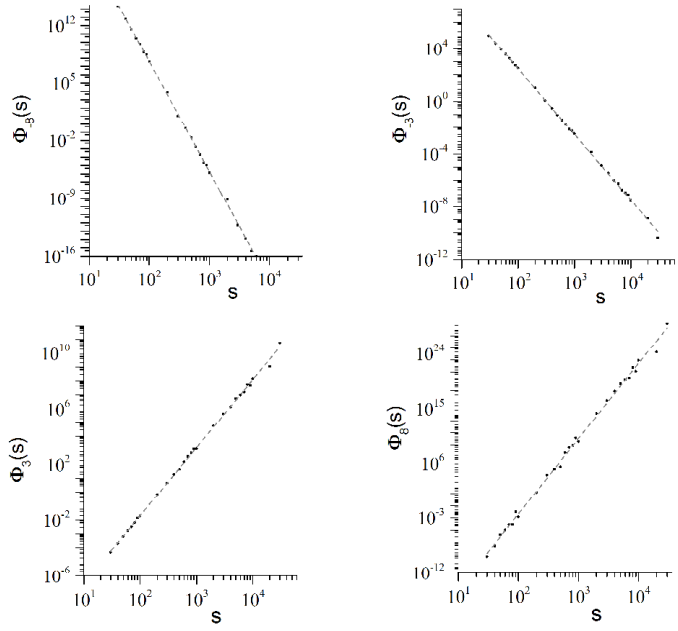



**Fig. 4.** Fluctuation function for different q and straight line approximation (dashed grey). From left to right: -8, -3 (top); 3, 8 (bottom)

## 4 Conclusion

A variant of mDFA with an automatic non-parametric trend extraction using SSA has been tested. The capabilities of the method were shown on generated fractional Brownian motion and the method is capable of extracting fractal characteristics in the presence of wideband and narrowband noise. The main task that the author focused on when choosing a technique was the analysis of time series obtained in experiments on ion channels in cells [12].

## References

1. Hergarten, S.: Self-organized criticality in Earth systems. Springer (2002). doi: 10.1007/978-3-662-04390-5
2. Peng, C. K., Buldyrev, S. V., Goldberger, A. L. et al.: Long-range correlations in nucleotide sequences. Nature, vol. 356, no. 6363, pp. 168–170 (1992). doi: 10.1038/356168a0
3. Schmitt, F.G., Huang, Y.: Stochastic analysis of scaling time series. Cambridge University Press (2015). doi: 10.1017/CBO9781107705548
4. Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde E. et al.: Multifractal detrended fluctuation analysis of nonstationary time series. Phisica A, V. 316, Issues 1-4, P. 87–114 (2002). doi: 10.1016/S0378-4371(02)01383-3

5. Vautard, R., Ghil, M.: Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Phys. D. V. 35, P.395–424 (1989). doi: 10.1016/0167-2789(89)90077-8
6. Golyandina, N., Nekrutkin, V., Zhigljavsky A.: Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC (2001). doi: 10.1201/9781420035841
7. Alexandrov, T.: Software package for automatic extraction and forecast of additive components of time series in the framework of the Caterpillar-SSA approach: PhD thesis. St.Petersburg State University, 152 p (2006). (In Russian)
8. Alexandrov T.: A Method of Trend Extraction Using Singular Spectrum Analysis. RevStat, V 7, P. 1–22 (2009).
9. Teukolsky, A., Vetterling, W.T., Flannery, B.P., Metcalf M.: Numerical Recipes in Fortran 90 - The Art of Parallel Scientific Computing. Cambridge University Press (1996).
10. Korobeynikov, A.: Computation- and Space-Efficient Implementation of SSA. Stat. Interface, V. 3, P. 357–368 (2010). doi: 10.4310/SII.2010.v3.n3.a9
11. Baglama, J., Reichel, L.: Augmented implicitly restarted Lanczos bidiagonalization method. Sci. Comput. ,V. 27, P. 19–42 (2005). doi: 10.1137/04060593X
12. Neher, E, Sakmann, B.: Single-channel currents recorded from membrane of denervated frog muscle fibres. Nature , 260, P. 799–802 (1976). doi: 10.1038/260799a0

# Metamodeling Based Approach for District Heat Network Aggregation

Nihad Aghbalou[1], Mouhamed Tahar Mabrouk[2], Pierrick Haurant[2], Mireille Batton-Hubert[1], and Bruno Lacarriere[2]

[1] Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France nihad.aghbalou@emse.com
[2] IMT Atlantique, Department of Energy Systems and Environment, GEPEA, F-44307, Nantes, France

**Abstract.** To deal with the problem of DHN modelling, this manuscript introduce a refinement approach to develop a novel method for district heat network aggregation based on soft computing techniques (Neural Networks and meta-heuristics). Reducing size and complexity of DHN by preserving the dynamic properties of district heating without physical description (network topology, pipe diameter and insulation, pump characteristics etc.) is a main challenge for simulating various operational strategies. Data from reel DHN system was used to investigate the ability of the proposed approach. This approach embed the time of flow transportation in the inputs of the FFNN model. The time of flow transportation between production plant and consumers substations is determined using the wavelet analysis. The case of study has provided encouraging results about the proposed model. However, this can be more enhanced by considering for instance the quality of the measured data and then finds application for DHN modelling.

**Keywords:** District heat network (DHN) aggregation · Feed Forward Neural Network (FFNN) · Meta-heuristics based population · Cross-Wavelet Transform .

## 1 Introduction

Nowadays, multi-energy system constitute a promising asset towards a smart grid deployment. Benefits from exploiting synergies between the different energies are manifold such controlling energy flows supplied from alternative sources, enhancing security and reliability energy supply and decreasing costs and energy losses. For instance, combined Heat and Power (CHP) units, electric boilers and heat pumps connected to a district heating system with energy storage could contribute to more efficient utilisation of distributed energy[1]. But to study the different coupling technologies in a such integrated system, there doesn't exist models enabling to compute loads and temperature at different locations in the distributed network.
The DHN is a complex and large structure comprising one or more heat producer, dozens of distribution insulated pipes which distribute the heating fluid

from the plants towards the users, and back to the plants and substations represented by the associated heat exchanger representing one or group of users [2]. Modelling this network requires checking physics laws and resolving sets of non-linear equations derived from the thermal and fluid dynamic theories. In addition, detailed model cannot be directly simulated because of their computational cost. They are usually simulated simultaneously with other models of demand, storage and multiple technology of energy supply.
Reducing size and complexity by aggregating the network into a smaller and simpler representation stay an attractive alternative for operational optimization, assessment of new connection or of introducing geothermal or solar thermal producers to the network. However, reducing a network topology to a suitable size without losing accuracy and physical significance is important to avoid aberrant between measured and simulated results and then the cost effectiveness [3]. Aggregation techniques addressed in literature are scarce and of two groups ; physical method including or excluding heat and pressure losses and statistical methods based on data measurements[4][5][6]. Physical models (i.e. node methods or function method [7]) impose the hypothesis that pipes share physical properties and geometric properties and the bends an fittings in layout of the pipes distribution. Gabrielaitiene and al. [8] show that the conventional node method and the commercial software TERMIS results in discrepancies between the predicted and measured temperatures for consumers located at from a far. Statistical methods concern for illustrative purposes the first-order autoregressive models presented in a pioneer thesis [9], the conditional finite impulse response method introduced by Pinson et al. [10]. The main drawback of those methods is the requirement of estimating the models parameters and moreover the determination of the characteristic time in the network which is a a delicate task if the temperature in the network changes abruptly. This characteristic it is often considered constant for the sake of simplicity.
In view of this context, this work aims to develop a Feed Forwarded Artificial Neural Network (FFNN) based method for aggregating a DHN without physical details considering the varying characteristic time in the network. Artificial neural networks (ANN) have been applied for some modelling DHN issues. Kato et al. [11] proposed a recurrent ANN to correspond the dynamical variation of heat load and improve the precision of predicting the heat load by reconsidering characteristics of heat load data. Strusnik et Avsec [12] used an ANN model as sub-model of DHN model to compute the consumer heat quality based on the outdoor temperature input. S. Deng et Y. Hwang [13] used the continuous-time analogue Hopfield neural network to compute the temperature distribution in both time-varying 1D and 2D forward heat conduction problems. Laakkonen et al. [14] have applied a neural network to predict future heat load and the consumer return temperature as it does take disturbance from delays and heat losses in DHN. In this paper, a new meta-model based method is proposed for forecasting the fluid temperature supplied at one or more consumer based on the temperature supplied in the inlet of the branch, the total mass flow as variable control and the consumption at the consumer as a local foretasted load profile.

This method is generic and has the particularity of taking into account the time characteristic of fluid transporting. The problem formulation and mathematical material are detailed in section 2 before the case study and results

## 2 Methodology

### 2.1 Problem Formulation

Based on the fluid temperature, a DHN is slitted in two sub-networks: The primary in the side of the district heating plants, transports heat from a boiler to delivery points at consumers and the secondary network is internal distribution in the building and it dynamics depend on occupant behaviours. The heat from a hot fluid (in the primary) is transferred to a cold fluid (in the secondary) by means of an exchanger. The characteristics involved in a DH system concerns temperature, pressure and flow mass along the supply and return pipes. Figure 1-a) illustrate a branch of six substations. The load $P_{sst}$ at each substation in the primary side of the network (delivered power) varies since the supply temperature propagate through the supply line with the flowing water. It can be written as:

$$P_{sst}(t) = c_p m(t)(T_s(t) - T_r(t)) \tag{1}$$

where $c_p$ (J/kg) is the specific heat capacity of water, $m(t)$(kg/s) is the mass flow of the water, $T_s$ and $T_r$ $(C)$ are successively the supply and the return temperature at district heat source. The supply temperature to the DHN propagate

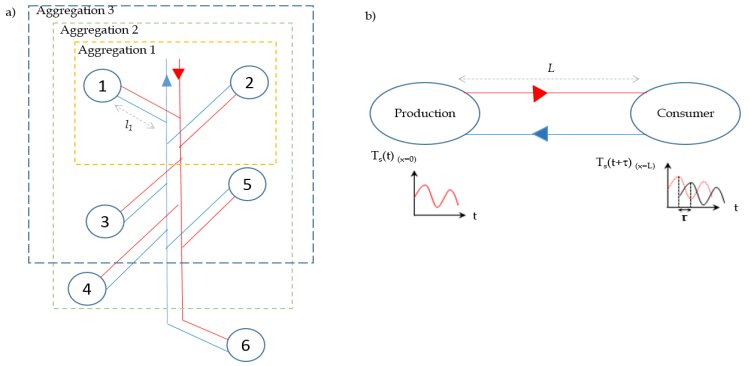



Fig. 1: a)Studied topologies of branch. b) Problem formulation : The production node represents the node 1 and the consumer node represents node 3 for aggregation 1 or node 6 for aggregation 2 in a). The red color refers to the primary side of the network and the blue to the second one.

from the plant with the flowing water through the supply pipe. Assuming that the distance $l_i$ between the consumer $i$ and the main pipe is negligible (then

loses), one can suppose that the inlet temperature to the branch equals to the temperature supplied to consumer 1. Following, the same reasoning, and based on the one-dimensional energy balance of a pipe of length $L$ supposed connecting consumer 1 to consumer $i$ (see figure 1-b)), the temperature measured at this consumer can be considered equals to the temperature delivered by the production plant (i.e. measured at consumer 1) shifted and attenuated and can be written as follow :

$$T_s(t)_{(x=L)} = T_{outdoor} + (T_s(t-\tau)_{(x=0)} - T_{outdoor})e^{(-\tau kS/(c_p \rho \pi R^2))} \quad (2)$$

where $T_{outdoor}$ refers to the ambient temperature, $L$ is the equivalent pipe length between two consumers, $\tau$ the time-varying delay, $k$ the heat conductivity of the surrounding, $R$ is radius of pipe and $S$ is shape factor for the pipe.
This paper shows that it's possible to predict $T_{s-sst_1}$ with a machine learning model to aggregate this branch using a set of historical data of upstream $T_{s-sst_1}$ and a downstream $T_{s-sst_{i(i=3:6)}}$ without physical details.
Based on equation(2), the time-varying delay is a key parameter to select the inputs of the this machine learning model. It depend on the mass flow and the distance between the two nodes. Generally, it is a key factor for system design, operation planning and optimal control [15]. More over, because of the changes in the consumers' habits and preferences, non-stationary or drifting phenomena in the time series of measured sittings (i.g.temperature) may be questionable. Hence, we propose to use wavelet analysis to determine this time characteristic.

### 2.2 Wavelet transform and cross wavelet analysis [16]

The wavelet transform is a technique of adaptive windowing signal analysis or variable bandwidth filtering. It consists in using an analysis function called "mother wavelet" (also called an atom) which acts on the signal by part. A translated and dilated mother wavelet generates a wavelet basis allowing to describe the signal variation in time (each translation of the mother wavelet) and frequency ( inverse of the mother wavelet scale approximately). Two main classes of wavelet transforms; the Continuous Wavelet Transform (CWT) and its discrete counterpart (DWT). The DWT is a compact representation of the data and is particularly useful for noise reduction and data compression whereas the CWT is better for feature extraction purposes with the aid of the wavelet power spectrum, variance fluctuations of non-stationary time series at different times and frequencies can be identified. Wavelet analysis is particularly useful to deal with data that exhibit inhomogeneities.
The CWT of a time series $x_{n'}, (n' = 1 : N)$ (which is our case) at a scale $s$ is defined as a sum of $N$ product of convolution of $x_{n'}$ with a translated wavelet function :

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^*[\frac{(n'-n)\delta t}{s}] \quad (3)$$

where $n$ is the shifting parameter, $\delta t$ is sampling time and $\psi^*(.)$ is the complex conjugate of the wavelet filter. Once transformed using the wavelet filters,

variance $|W_n(s)|^2$ as a function of time or frequencies allows analysing features. Moreover, it allows a framework for comparing similarities between signals and studying their cross-correlation in time and frequency using the squared wavelet coherence coefficient defined by :

$$R^2_{xy}(\tau, s) = \frac{|S(s^{-1}W_{xy}(\tau, s))|^2}{S(s^{-1}|W_x(\tau, s)|^2)S(s^{-1}|W_y(\tau, s)|^2)} \tag{4}$$

where $W_{xy}(\tau, s)$ is the cross wavelet power defined as product of the continuous wavelet transform of $x(t)$ and the complex conjugate of the continuous wavelet transform of $y(t)$, and $S$ is a smoothing operator. The squared wavelet coherence measures the local linear correlation between two time series at each scale and localise the strong co-movements. Then, the corresponding lead–lag relationships between the two signals is deduced as follows

$$\varphi_{xy} = tan^{-1}(\frac{\Im S(s^{-1}(W_x(\tau, s))}{\Re S(s^{-1}(W_x(\tau, s))}) \tag{5}$$

### 2.3 ARTIFICIAL NEURAL NETWORKS

The artificial neural network is the most popular paradigm in machine learning. They are models of highly connected assemblies of computational units referred to as "formal neurons", organized in layers and form computational models capable of solving very complex problems: classification; shape recognition; prediction of time series; etc. Each neuron performs a weighted sum of modulated input signals by a function called an activation function or a transfer function (a non-linear function) and generates an output that will be applied to the other neurons via weighted connections. The number of neurons and the type of the transfer function of the neurons in each layer, as well as the nature of connection between the different neurons (weighted unidirectional interlayer connection, infinite connection impulse response connection finite impulse response) define the topology of the neural network. In practice, there are two main categories of ANN models: Static and dynamic. The mathematical formulation of a static neural network with $N_e$ inputs $N_c$ nodes in the hidden layer with sigmoid activation function and is given as follows:

$$y_k = (1 + \exp(-(\sum_{j=1}^{N_c} W_{jk} \frac{1}{1 + \exp(\sum_{i=0}^{N_e} -W_{ij}x_i)} + b_{(N_c+1)})))^{-1} \tag{6}$$

where $y_k$ is the expected value of $y_k$, $W_{ij}$ is the magnitude of the weight connecting $i^{th}$ node in the input layer to the $k^{th}$ node in the hidden layer and $W_{jk}$ is the magnitude of the weight connecting $k^{th}$ node in the hidden layer to the $k^{th}$ output. In general, the number of the neurons in the hidden layer is selected based on the trial and error tests. The machine learning involves two stages: The learning consists of finding, hidden relations and rules connecting

input variables to output variables. Thus, determining optimal weights $W_{..}$ that optimize an objective function. The second stage consist on testing the model. An objective function can be a simple quadratic cost on the output layer, cross entropy, Hellinger distance, Information Divergence or other. The generalization is the most important performance, it concern how well the model generalizes to new data without excessive number of adjustable parameters (weights). Recently, many meta-heuristic techniques have proved to be very efficient in the FFNN learning as they are population based stochastic algorithm. Bat Algorithm, Cuckoo search and Particle Swarm Optimization are selected to be experimented in this study as they differs in their strategies in exploration and exploitation of the search space.

**Algorithm (BA)** The BA algorithm, proposed by Yang [17], is inspired from the echolocation behaviour of bats and their capacity of finding their prey and discriminate different types of insects even in complete darkness. X-B. Meng et al. [18] considered the bat's self-adaptive compensation for Doppler Effect in echoes and individual's difference in the compensation rate to enhance the basic BA. The BA algorithm update the local solution that minimize the objective function as follows :

$$x_{i,j}^{(k+1)} = x_{i,j}^{(k)} + v_{i,j}^{(k)} \tag{7}$$

$$v_{i,j}^{(k+1)} = w * v_{i,j}^{(k)} + (g_j^{(k)} - x_{i,j}^{(k)}) * f_{i,j}^{'} \tag{8}$$

$$f_{i,j}^{'} = \frac{(c + v_{i,j}^{(k)})}{(c + v_{g,j}^{(k)})} * (1 + C_i * (\frac{(g_j^{(k)} - x_{i,j}^{(k)})}{|g_j^{(k)} - x_{i,j}^{(k)}| + \epsilon}) * f_{i,j} \tag{9}$$

where $w$ is the inertia weight $\in$ [0,1]. $\epsilon$ a smallest constant to avoid zero-division-error. $C_i$ is the compensation rate $\in$ [0,1]. $c$ is the speed in the air. $v_{g,j}^{(k)})$ is the speed corresponding to the global best position $g_j^{(k)}$ in the bats swarm. $f_{i,j}$ is the frequency randomly assigned from $f_{min}$ and $f_{max}$.

**Particle swarm optimization(PSO)** The PSO algorithm based population proposed by Kennedy and Eberhar[19]. Inspired from the social behaviour of some animals like the bird flocking. Each particle is equipped with a memory that allows it to memorize the best position *pbest* by which it has already passed and tends to return to that position. For each iteration, the velocity $v_p^{(i)}$ and the position $x_p^{(i)}$ of the $p^{th}$ particle are updated according to this following equations :

$$v_p^{(i)} = w(i) \times v_t^{(i-1)} + c_1 \otimes r_1 \times (x_{p_{best}} - x_p^{(i-1)}) + c_2 \otimes r_2 \times (x_{P_{best}} - x_p^{(i-1)}) \tag{10}$$

$$x_p^{(i)} = x_p^{(i-1)} + v_p^{(i)} \tag{11}$$

$$w(i) = w_{max} - \big((w_{max} - w_{min})\frac{i-1}{i_{max}}\big) \tag{12}$$

where $w(i)$ is called the inertia weight, by which the impact of previous velocity of particle on its current one can be controlled, $c_1$ and $c_2$ are positive constant parameters called acceleration coefficients which control the maximum step size,

$r_1$ and $r_2$ are independently uniformly distributed random variables with range (0,1), $g_{best}$ is the best-known position within its neighbourhood

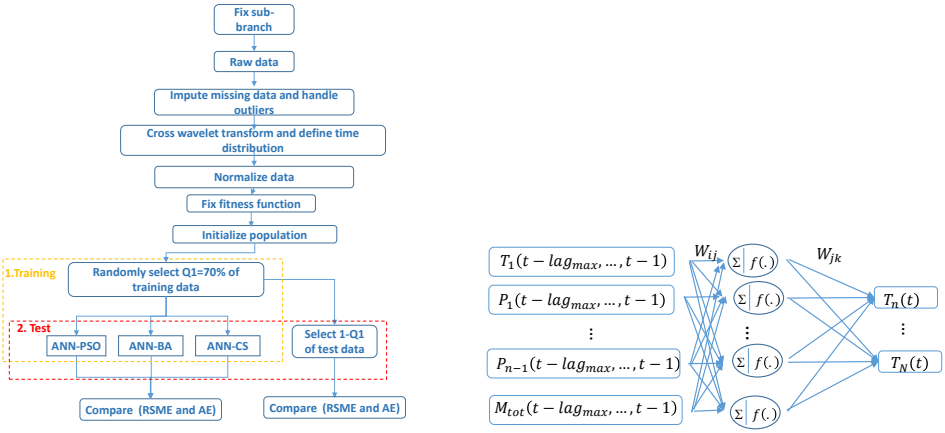
**Cuckoo search optimization** The Cuckoo search (CS) is a global search algorithm based population developed by Yang and Deb [20]. Inspired from the reproduction strategy and particularly the brood parasitism behaviour of certain cuckoo species with combination with the Levy flight behaviour. The updated generated solution $x_i^{(t+1)}$ of $i^{th}$ cuckoo is performed according to the following equation:

$$x_j^{(i)} = x_j^{(i-1)} + \alpha \oplus Levy(\lambda) \tag{13}$$

$\alpha$ is the step size which should be related to the scales of the problem. The random steps drawn from a Levy distribution for large steps govern the exploration and exploitation of the space search:

$$Lvy(\lambda) = \mathrm{l}^{-\lambda}; (1 < \lambda < 3) \tag{14}$$

### 2.4 Overview of the proposed study




(a) (b)

Fig. 2: a) Flowchart of the proposed study. b)ANN model $T_i$, $P_i$ are supply temperature and consumption at substation $i$ $M_{tot}$ is the total flow rate at the branch, $lag_{min}$ and $lag_{max}$ are minimum et maximum time delay determined by the wavelet analysis.

The study carried out in this paper is based on data measured in a subnetwork of six substations in the DHN of a city located in France as shown in figure 1-a. The purpose is to build an off-line aggregated model of this branch that can be used for dynamic modelling. The overview of the proposed study is shown in figure 2-a. After defining the different lags, consumer consumption and mass flow data in the branch are also considered in the model inputs as shown

in figure 2.b). The idea through defining the lag time is assessing the temporal dependencies and selecting the most relevant variables in the historical data. Three different population based meta-heuristics that are distinguished by their strategies of exploitation and exploration of the solution space are experienced to train the neural models. Similar initial population was affected to each comparison step of the three algorithms with same size and number of maximum iteration. All training samples are used as one batch. Two fitness function are considered : the standard deviation $fit_1$ which enables the developed model to keep stability of generalization and the mean absolute error $fit_2$ that focus particularly on the error and accuracy of estimation.

$$fit_1 = \frac{1}{N}\sqrt{[(y_k - y_k)^2]} \tag{15}$$

$$fit_2 = \frac{1}{N}(\sum_{i=1}^{N} abs(\frac{(y_k - y_k)}{y_k})) \tag{16}$$

$N$ is the size of the training data, $y_k$ is the measured value and $y_k$ is the forecasted one. The number of hidden layer was fixed to one and the number of hidden nodes to 7, with hyperbolic tangent sigmoid activation function. The gaussian noise assumption in the studied data is kept since they are collected from sensor distributed at the thermal and hydraulic circuit at different substations. A gaussian regularizer was added to the fitness functions to avoid the over-fitting [21] :

$$obj_{i(1,2)} = \alpha fit_{i(1,2)} + \beta \sum_{i}^{M} \frac{1}{2M}[w_i]^2 \tag{17}$$

$M$ is the number of ANN parameters. Here $\beta = 1 - \alpha$ as fixed in [21].

## 3 Results and Discussion

### 3.1 Transport delays of heat propagation

Database of this study covered the cold and the hot period from $22^{th}$ June 2017 to the $26^{th}$ of April 2018. Data measurement have 10 min resolution, this corresponds to one lag unit in this work. The rate of the missing values is about 20%. The gaps in these time series are mainly due to sensor values transmission problems since they have been found simultaneous. For the sake of simplicity, missing values are replaced by zero for the wavelet analysis and completely not considered in the learning phase. To conduct the wavelet analysis, the complex Morlet mother wavelet is used as it is quite well localized in both time and frequency space and allows defining the lead-lag between time series. Figure 3 depicts the wavelet power spectrum of the supplied temperature to the studied substations. Common features in the wavelet power spectrum of the six time series are obviously observed and no regular periodicity is observed. Similar peaks in the 64-256 h band in periods around 02-Nov-2017, 10-Feb-2018 and 24-Mar-2018 are also observed. Figure 4 presents coherence spectra between temperature

T1/T3, T1/T4 and T1/T6. Phases (equ 5) are represented by means of arrows and converted to time unit. The min and max lags, calculated for hight frequency bands, between temperature in substations sst1 and sst3, sst1 and sst4, sst1 and sst5 and sst1 and sst6 are [1;4],[1;5], [2;6] and [2;10] respectively.

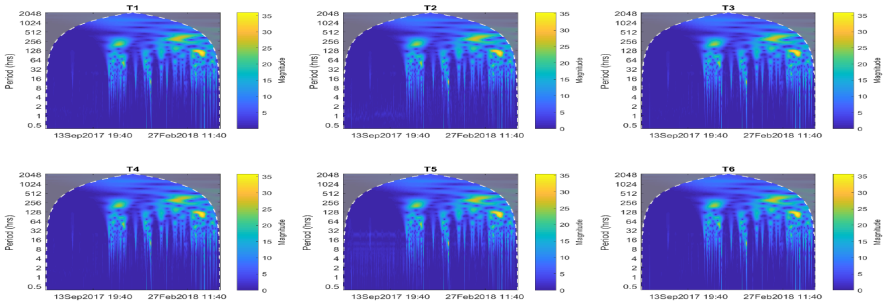



Fig. 3: Wavelet power spectrum of temperature supply at substations.

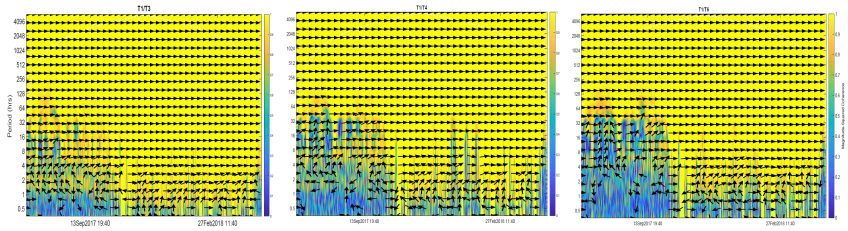
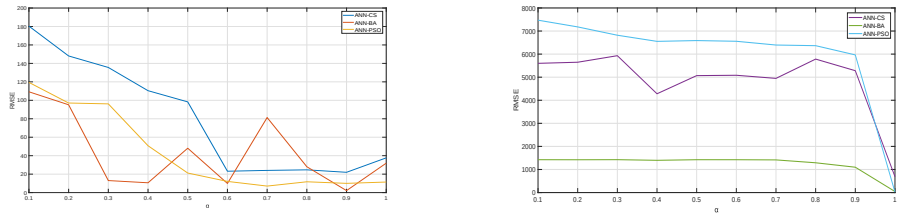


Fig. 4: Coherence spectra plots between temperature T1/T3, T1/T4 and T1/T6 at substations.

### 3.2 Neural models comparison

Three different architecture have been studied as shown in figure 1 and results are presented hereafter:

– **Aggregation 1** The first case consist of aggregating two substation sst1 and sst2. This case study a substation which is in the middle of the branch. Results depicted in figure 5 indicate the $\alpha$ values and corresponding $RMSE$ of training neural models using the different algorithms. For the two objective functions, the differences of $RMSE$ values are very important with respect to $\alpha$, particularly using $obj_2$ where all algorithms converge at $\alpha = 1$ for what the regularization term is completely eliminated. However, using $obj_1$ algorithms converge to a small $RMSE$ : 7.07 at $\alpha = 0.7$ for $ANN - PSO$,

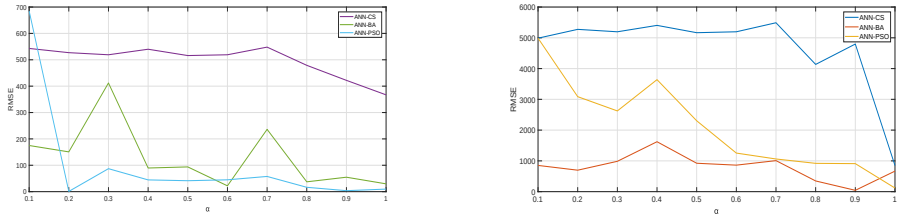



(a) (b)

Fig. 5: Aggregation 1 : RMSE of training models with different hyper-parameter $\alpha$ using a) $obj_1$ b) $obj_2$.

2.1622 for $ANN-BA$ and 22 for $ANN-CS$ at $\alpha = 0.9$. Box plots[3] in figure 8 a),b),c) and d) illustrate the variation of the absolute error for retrained models. It is evidently seen that ANN-BA and ANN-PSO outperform the ANN-CS.

– **Aggregation 2** In this second case, the purpose is to study the aggregation




(a) (b)

Fig. 6: Aggregation 2 : RMSE of training models with different hyperparameter $\alpha$ using a) $obj_1$ b) $obj_2$.

up to the end of the branch. Results are similar to the previous case as far as the performance of the algorithms using the different algorithms. Using $obj_1$ algorithms converge to a small $RMSE$ : 0.84 at $\alpha = 0.2$ for $ANN - PSO$, 22.06 for $ANN-BA$ at $\alpha = 0.6$ and 372 for $ANN-CS$ at $\alpha = 1$. Moreover, in figure 8 e),f),g) and h), the dispersion of the absolute error is reduced and shows again the performance of the $ANN - PSO$.

– **Aggregation 3** This example treats the case when more than one bifurcation are concerned (sst2, sst3 and sst5) and two specific points of the network are considered (sst4 and sst6). This ANN model is with two outputs. The

[3] On each box, the central mark indicates the median, and the lower and upper edges of the box indicate respectively $25^e$ and $75^e$ percentiles. Whiskers extend to the most extreme data points (minimum and maximum) that not were considered outliers. The outliers are plotted individually using the " + " symbol.

average RMSE is of training models is depicted in figure 7. The performance of the $ANN-BA$ and $ANN-PSO$ are approximately similar. Based on the performance of models of aggregation 2 and 3 for forecaseting $T6$ (box-plots in figure 8 e), f),k) and l)) the accuracy of the proposed model decrease when more than one reference in the network is considered for the output of the aggregation.

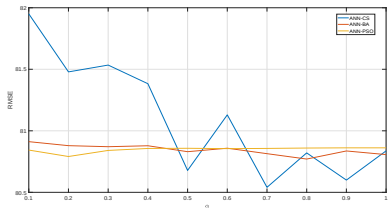



Fig. 7: Aggregation 3 : Average RMSE of training models with different hyperparameter $\alpha$ using $obj_1$ .

## 4 Conclusion

For a particular application to dynamic modelling, this paper introduces a methodology to reduce the complexity of a DHN. The attention was more paid to the supplied temperature. The dynamic characteristic was preserved through the inclusion of the characteristic time of heat transport between two nodes (i.e. substations) of a DHN. The consumption profile and the total mass flow, as an operational characteristic, were also considered by the model. The preliminary results indicate that an ANN based meta-model enables aggregating substations. However, some enhancements are still necessary to increase the performance of the model. The hypothesis of the non-homogeneity and low quality of data detected by the wavelet analysis is possible. To check this, it is necessary to run the model with other real or simulated data set. Moreover, to enhance the stability and the accuracy of the ANN based model, a possible multi-criteria optimization using different fitness function can be investigated.

## References

1. Liu, Xuezhi, et al. "Combined analysis of electricity and heat networks." Applied Energy 162 (2016): 1238-1250.
2. Bordin, Chiara, Angelo Gordini, and Daniele Vigo. "An optimization approach for district heating strategic network design." European Journal of Operational Research 252.1 (2016): 296-307.

Fig. 8: Box-plot of absolute error of forecasting T3 for models ANN-CS, ANN-BA et ANN-PSO optimizing $obj_1$ with retrained value of $\alpha$ a) Training. b)Test and optimizing $obj_2$ with retrained value of $\alpha$ c) Training. d)Test. Box-plot of absolute error of forecasting T6 for models ANN-CS, ANN-BA et ANN-PSO optimizing $obj_1$ with retrained value of $\alpha$ e) Training. f)Test and optimizing $obj_2$ with retrained value of $\alpha$ g) Training. h)Test. Box-plot of absolute error of forecasting T4 for models ANN-CS, ANN-BA et ANN-PSO optimizing $obj_1$ with retrained value of $\alpha$ i) Training. j)Test. Box-plot of absolute error of forecasting T6 for models ANN-CS, ANN-BA et ANN-PSO optimizing $obj_1$ with retrained value of $\alpha$ k) Training. l)Test.

3. Velut, Stéphane, et al. "Short-term production planning for district heating networks with JModelica. org." Proceedings of the 10 th International Modelica Conference; March 10-12; 2014; Lund; Sweden. No. 096. Linköping University Electronic Press, 2014.
4. Larsson, Henning. "District Heating Network Models for Production Planning." (2015).
5. Larsen, Helge V., et al. "Equivalent models for district heating systems." Proceedings of the 7th International Symposium on District Heating and Cooling. Lund, Sweden: S. Frederiksen. 1999.
6. Larsen, Helge V., et al. "Aggregated dynamic simulation model of district heating networks." Energy conversion and management 43.8 (2002): 995-1019.
7. Zheng, Jinfu, et al. "Function method for dynamic temperature simulation of district heating network." Applied Thermal Engineering 123 (2017): 682-688.
8. Gabrielaitiene, Irina, Benny Bøhm, and Bengt Sunden. "Modelling temperature dynamics of a district heating system in Naestved, Denmark—A case study." Energy conversion and management 48.1 (2007): 78-86.
9. Søgaard, H.T., 1993. Stochastic systems with embedded parameter variations – Applications to district heating. Ph.D. Dissertation. Technical University of Denmark, Institute of Mathematical Statistics and Operations Research.
10. Pinson, Pierre, et al. "Temperature prediction at critical points in district heating systems." European Journal of Operational Research 194.1 (2009): 163-176.
11. Kato, Kosuke, et al. "Heat load prediction through recurrent neural network in district heating and cooling systems." 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2008.
12. Strušnik, Dušan, and Jurij Avsec. "Artificial neural networking model of energy and exergy district heating mony flows." Energy and Buildings 86 (2015): 366-375.
13. Deng, S., and Y. Hwang. "Applying neural networks to the solution of forward and inverse heat conduction problems." International Journal of Heat and Mass Transfer 49.25-26 (2006): 4732-4750.
14. Laakkonen, Leo, et al. "Predictive Supply Temperature Optimization of District Heating Networks Using Delay Distributions." Energy Procedia 116 (2017): 297-309.
15. Schwarz, M. B., Mabrouk, M. T., Silva, C. S., Haurant, P., Lacarrière, B. (2019). Modified Finite Volumes Method for the Simulation of Dynamic District Heating Networks. Energy.
16. Torrence, C., Compo, G. P. (1998). A practical guide to wavelet analysis. Bulletin of the American Meteorological society, 79(1), 61-78.
17. Yang, Xin-She, and Suash Deb. "Cuckoo search via Lévy flights." 2009 World Congress on Nature Biologically Inspired Computing (NaBIC). IEEE, 2009.
18. Meng, Xian-Bing, et al. "A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization." Expert Systems with Applications 42.17-18 (2015): 6350-6364.
19. Kennedy, James, and Russell C. Eberhart. "A discrete binary version of the particle swarm algorithm." 1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation. Vol. 5. IEEE, 1997.
20. Yang, Xin-She, and Suash Deb. "Cuckoo search via Lévy flights." 2009 World Congress on Nature Biologically Inspired Computing (NaBIC). IEEE, 2009.
21. Jin, Yaochu, Tatsuya Okabe, and Bernhard Sendhoff. "Neural network regularization and ensembling using multi-objective evolutionary algorithms." Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753). Vol. 1. IEEE, 2004.

# Theoretical foundation of detrending methods for fluctuation analysis such as detrended fluctuation analysis and detrending moving average

Marc Höll[1], Ken Kiyono[2] and Holger Kantz[3]
(1)Bar-Ilan University, (2)Osaka University (3)Max Planck Institute for the Physics of Complex Systems

## 1 Abstract

We present a bottom-up derivation of methods for fluctuation analysis with detrending and claim their basic principles. Such methods detect long-range correlations in time series even in the presence of additive trends or intrinsic nonstationarities. Long-range correlations which are often characterized by a power law decaying autocorrelation function are omnipresent in a tremendous amount of data sets. But especially in nonstationary time series, their existence has a non-negligible impact on the analysis, modeling and prediction of the time series. Therefore advanced methods are necessary to extract information about the correlation structure. The well-known detrended fluctuation analysis (DFA) and detrending moving average (DMA) are such methods but were introduced ad hoc. Both methods sample over mean-squared errors between the summed time series and its fitting polynomials in different time windows. We show that DFA and DMA fulfill our claimed basic principles of the general framework of detrending methods for fluctuation analysis. The bottom-up derivation of these principles starts with the observation that the standard sample estimator of the autocorrelation function has at least two fundamental statistical problems. First, the estimator fluctuates strongly around zero for large time lags which makes it difficult to observe a possible power law decay. Second, the estimator is only meaningful for stationary time series. Therefore we consider instead the mean-squared displacement of the summed time series. It contains the same information about long-range correlations as the autocorrelation function but overcomes the first problem as it is an increasing function depending on the time lag. However, the estimation of its scaling behavior on a single time series is affected not only by additive trends on the data but also by intrinsic nonstationarities. Exemplary, fractional Brownian motion without additive trends is an intrinsic nonstationary process. We show in detail, how the estimation of the mean-squared displacement of the summed time series is disturbed by these two types of nonstationarity. We relate this estimator with the autocorrelation function of the random signal of the time series so that we are able to identify analytically the reason of failure of correct estimation. If we write the mean-squared displacement of the summed time series with respect to the autocorrelation function we call this the increment representation as the time series is the increment process of the summed time series. A nonstationary process either having additive trends or being only fractional Brownian motion has a nonstationary autocorrelation function which therefore introduces a bias in the estimation of the

mean-squared displacement of the summed time series. This bias is exactly the reason why stationary detrending methods such as fluctuation analysis (FA) or DFA with zero order detrending always observe the same scaling behavior in the presence of nonstationarities no matter what random process is underlying. To erase the influence of the bias we introduce a weighting kernel for the mean-squared displacement of the summed time series in the increment representation which is the only possible presentation to formulate a general form of the weighting kernel. We call this function the fluctuation function of the time series and will later show that DFA and DMA are specific examples with each having their own weighting kernel. Now we can claim two basic principles of detrending methods for fluctuation analysis. First, we claim that the fluctuation function scales like the mean-squared displacement of the summed time series. We preserve the information of the latter. Second, we claim that the estimator of the fluctuation function is unbiased even in the presence of nonstationarities. Hence the weighting kernel removes the influence of the bias during the estimation procedure. This is called detrending in methods for fluctuation analysis which is usually introduced by removing fitting polynomials from the summed time series. Our novel and more general understanding of detrending provides a unified picture for different types of nonstationarity and especially explains detrending for the intrinsic nonstationary fractional Brownian motion. Both principles define the general framework for detrending methods and serve as basis to derive important characteristics of them such as the following three. First, we can derive a superposition principle assuming that the time series consists of independent additive processes so that then the fluctuation function is the sum of individual fluctuation functions. Second, we can relate our theory to the field of wavelet transforms as a factorizable weighting kernel of the fluctuation function is identical to the Haar wavelet transform. And third, our framework allows to formulate the practical implementation of detrending methods in a general way where only the specific form of weighting kernel is optional. Here DFA and DMA are suitable candidates because both are able to detect the correct scaling behavior in the presence of nonstationarities. However, their original description is different than above general framework meaning their original forms of the fluctuation function are not written in the increment representation. For both DFA and DMA we analytically derive their weighting kernels and show that above two principles are indeed fulfilled. Therefore DFA and DMA are examples of detrending methods. Although both methods are different their basic techniques of detrending works very similar. The main difference between them is the method of sampling. But no matter what sampling procedure one chooses the influence of the bias introduced by nonstationarities is oppressed by their weighting kernels. The knowledge of the weighting kernel allows us not only to show the fulfillment of the basic principles but furthermore also to calculate analytically the fluctuation function for different stochastic models. Notably, for autoregressive models the fluctuation function exhibits a crossover behavior in scaling. This crossover behavior demands long enough data sets which is often not the case in real applications. Summarized, we provide a theoretical framework of detrending methods for fluctuation analysis which serves as basis to investigate many application problems. This framework also allows to find new methods with more suitable weighting kernels for certain problems in fluctuation analysis of nonstationary time series.

# Seasonal Models for Forecasting Day-Ahead Electricity Prices

Catherine McHugh[1a], Sonya Coleman[1b], Dermot Kerr[1c], and Daniel McGlynn[2d]

[1] Intelligent Systems Research Centre (ISRC), School of Computing, Engineering and Intelligent Systems, Ulster University, Northern Ireland, UK
[2] Click Energy, Northern Ireland, UK
[a]mchugh-c24@ulster.ac.uk, [b]sa.coleman@ulster.ac.uk,
[c]d.kerr@ulster.ac.uk, [d]daniel.mcglynn@clickenergyni.com

**Abstract.** Prediction algorithms are increasingly popular techniques within the energy industry to assist in forecasting future prices and to help reduce costs. When predicting day-ahead energy prices several significant external factors that influence electricity prices need to be considered. Initially we use the transparent Nonlinear AutoRegressive Moving Average model with eXogenous inputs (NARMAX) to model the relationship of factors that contribute to energy costs and day-ahead electricity prices. Energy data from 2017 were analyzed separately for Spring, Summer, and Autumn periods to observe the effect of factors on seasonality. The seasonal NARMAX models identify important input factors and their correlated lagged values such as price, energy demand, generation type, and environmental temperature to be significant factors in accurately predicting day-ahead prices for all three seasons. The identified factors are used to refine a Seasonal AutoRegressive Integrated Moving Average model with eXogenous variables (SARIMAX). To conclude the seasonal NARMAX models proved useful for removing insignificant external factors and the seasonal refined SARIMAX models permitted accurate prediction of day-ahead prices.

**Keywords:** Energy Forecasting, Time-Series Modelling, Seasonality, Day-Ahead.

## 1 Introduction

Energy market price prediction is a difficult problem as many of the factors that contribute to market prices display nonlinear and nonstationary behaviour [1]. Volatility in contributing factors causes spikes in market trends due to imbalances between demand and supply, therefore a model that is robust and manages volatility is desirable [2]. Recently there has been an increase in electricity price forecasting with market traders trying to control volatility by applying prediction algorithms [3]. Short-term imbalances between supply and demand that cause this volatility means that forecasting should occur over a small prediction timeframe. Therefore with volatile energy data, short-term models are considered favorable for energy forecasting.

Price forecasting models learn from historical data to accurately make predictions and spot trends, hence it is important to correctly identify contributing factors to establish a successful price forecasting tool [4]. Energy forecasting is influenced by many economic and technical factors which result in price fluctuations [5] therefore it is necessary to include the key determinants when modelling. Including external factors in price forecasting models can help improve energy price prediction for trading [2]. Previous literature [6], [7] has mentioned important external factors to consider in energy forecasting models: wind distributes the largest volatile production ratio [6]. Pandey and Upadhyay [5] highlight that energy demand is a significant contributing factor as demand can vary in conjunction with other factors (for example, weather or temperature), but other factors like fuel prices and generator maintenance can also have an impact on electricity price. In some energy markets real-time factors can influence the cheapest bid (system marginal price) [8] therefore prediction models need to consider these to accurately predict prices. As well as this, prices during the same hour/day (trend) and spring/summer (seasonality) are observable since, over time, energy data display patterns and thus should be considered in model development.

Day-ahead energy forecasting has previously been applied to the Spanish and Californian energy markets [9] with results displaying small errors highlighting accurate predictions, especially for short-term forecasting. Gao et al., [3] explored Artificial Neural Networks (ANN) models [10] and AutoRegressive Integrated Moving Average (ARIMA) models and noted that RMSE was lowest for ARIMA approaches, emphasizing that the model precision was more accurate with a short-term forecast. Li et al., [11] explained that energy data peak at every 24-hour and demonstrates this to be true when applying autocorrelation analysis for energy load, suggesting strong correlation in factors of the same hour which are recommended as inputs. Historical input factors show if any causal relationship between explanatory variables and the price dependent variable exists [12]. It is best to select historical values with the most influence on electricity price to use as additional inputs to a forecasting model. This is important as not including such variables was considered a limitation in a Sweden energy study [13].

ARIMA models are a popular statistical time-series method that have demonstrated promising results in forecasting daily electricity prices [5]. These models can observe trends but require a large dataset to achieve accurate forecasts [14]. A Seasonal ARIMA model (SARIMA) is an extension of ARIMA and includes seasonal trends, however the model needs to display constant mean/variance (stationarity) before applying forecasting techniques [12]. A SARIMA model with exogenous variables (SARIMAX) can improve prediction accuracy, [15], where forecasted hourly load with interactions of exogenous variables improved the model accuracy compared with using SARIMA.

NARMAX models are popular methods that have been utilized within industry, for example studying the key determinants of house prices in China [16] and in predicting cash demands of Automatic Teller Machines (ATMs) [17] which considered seasonal factors as input and noticed a strong recurring cycle. NARMAX models, specifically polynomial models, are non-linear and transparent [18], providing full transparency of the statistically significant terms. NARMAX can be considered suitable for many time-series modelling applications as NARMAX models have the capability to represent both linear and non-linear structures [19].

We focus on developing short-term price forecasting models, which include the contribution of external factors that impact electricity prices. The Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) methodology is a parameter estimation methodology to identify the important model terms and their associated parameters of a non-linear polynomial model of an unknown non-linear dynamic system. The NARMAX methodology is suitable for modelling the input-output relationship and involves the following key steps: i) initial model structure detection, ii) parameter estimation, iii) model validation, iv) prediction, and v) analysis. Thus, the methodology is utilized to find the contributing factors that influence electricity price in each season and develop final seasonal models for forecasting day-ahead price. We consider day-ahead price to be one-step ahead prediction where the step size is 24 hours from any given point in time. These factors are further used to inform and refine a Seasonal AutoRegressive Integrated Moving Average model with eXogenous variables (SARIMAX). Performance is evaluated for both NARMAX and SARIMAX approaches.

The remainder of this paper is organized as follows: Section 2 outlines the proposed methodology for time-series modelling and the results obtained are presented and discussed in Section 3. The paper is concluded with a summary of the main findings in Section 4.

## 2 Time-Series Modelling

A polynomial NARMAX model is represented as:

$$y(t) = F^l\big[y(t-1),..,y\big(t-N_y\big),u(t),..,u(t-N_u),\varepsilon(t-1),..,\varepsilon(t-N_\varepsilon)\big] + \varepsilon(t) \quad (1)$$

where $u(t)$ is the input time-series, $y(t)$ is the output time-series, $N_u$ is the input lag order, $N_y$ is the output lag order, $N_\varepsilon$ is the prediction error lag order, $F^l$ is a nonlinear function, and $\varepsilon(t)$ is the prediction error [20].

There are various stages of the NARMAX methodology to estimate $F^l$ and to find the significant model terms. These include: structure selection, parameter estimation, model validation, and prediction analysis. The input stage is a particularly important part of the methodology as the obtained model fitness is dependent on the selected input terms. Polynomial NARMAX models are the most popular form of nonlinear NARMAX models due to their simple transparent structure [21].

Modelling numerous parameters results in a model which is opaque and difficult to analyze. Hence, the NARMAX methodology uses the orthogonal estimation algorithm [20] to prune parameters, estimating both parameter coefficients and prediction errors in each iteration until convergence. The Error Reduction Ratio (ERR) helps to maintain model accuracy by ranking regressors using the Mean Square Error (MSE) [22]. Once parameters are estimated, the final model is verified using unseen data with Billings and Voon's [23] prediction process.

A SARIMAX model is outlined in [24] as:

$$\varphi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D yt = \beta_k x'_{k,t} + \theta_q(B)\Theta_Q B^s \varepsilon_t \quad (2)$$

where $\varphi_p(B)$ is the AutoRegressive (AR) polynomial term for standard ARIMA denoted as $p$, $\theta_q(B)$ is the Moving Average (MA) polynomial term for standard ARIMA denoted as $q$, and $(1-B)^d$ is the differencing for standard ARIMA denoted as $d$. $\Phi_P(B^s)$ is the AutoRegressive (AR) polynomial term for seasonal ARIMA denoted as $P$, $\Theta_Q B^s$ is the Moving Average (MA) polynomial term for seasonal ARIMA denoted as $Q$, and $(1-B^s)^D$ is the differencing for seasonal ARIMA denoted as $D$. $y_t$ is the prediction, $B$ is the backward shift, $\beta_k x_{k,t}{}'$ is the exogenous variable and $\varepsilon_t$ is the error term [24]. When calculating terms the model follows the Box-Jenkins method of identification, estimation, and diagnostic checking [25].

The order terms $p$, $q$, $P$, and $Q$ need to be manually set and this is achieved using the Akaike Information Criterion (AIC) technique. The quality fit of the model is measured by the AIC [13]:

$$\mathrm{AIC} = -2\ln(L) + 2k \tag{3}$$

where $k$ is total number of parameters and $L$ is maximum likelihood. The model with the lowest AIC value is the best. To achieve stationarity the input data have to be seasonally differenced ($S$) with the seasonal pattern repeating itself, for example a 24-hour lag would be a daily seasonal difference [13]. Orders of integration ($d$ and $D$) should be defined as the number of differences required to make the input data stationary with constant mean and variance [15]. If the seasonal pattern of the time series is unstable over time then D=0.

## 3 Experiments and Results

Historical data were collected for electricity market prices and various energy and environmental related factors that potentially contribute to day-ahead prices. The Nordpool day-ahead exchange traded (N2EX) market report contained hourly electricity price data from 2017 [26]. The National Grid, Great Britain (GB) website includes hourly gas price data [27]. Additional energy related data that contributed to energy generation were obtained from the Gridwatch GB website [28]. The Speedwell weather website was used to obtain hourly temperature [29] averaged from five U.K. weather stations. The data were split into seasons: Spring (20th March to 19th June - 2208 records), Summer (21st June to 20th September - 2208 records), and Autumn (22nd September to 19th December - 2136 records). For each season first 50% of the records were used for model estimation and the remaining 50% were used for model validation.

Table 1 displays the external factors initially included in the proposed seasonal NARMAX models. Autocorrelation determined the corresponding lags for each factor, and as expected most peaked every 24 hours. These lags, also shown in Table 1, were subsequently included as initial model inputs.

**Table 1.** External factors used as inputs and identified lags (in hours)

| External Factor (Unit) | Model terms ($t$=1) |
|---|---|
| Price (Euro per Megawatt Hour) | $u1(t);\ u1(t-2);\ u1(t-24)$ |
| Demand (Megawatt) | $u2(t);\ u2(t-2);\ u2(t-24)$ |
| Gas (Pence per Kilowatt Hour) | $u3(t);\ u3(t-2);\ u3(t-24)$ |
| Wind (Megawatt) | $u4(t);\ u4(t-1)$ |
| Solar (Megawatt) | $u5(t);\ u5(t-2);\ u5(t-24)$ |
| Coal (Megawatt) | $u6(t);\ u6(t-2);\ u6(t-24)$ |
| Moyle Interconnector (Megawatt) | $u7(t);\ u7(t-2);\ u7(t-24)$ |
| Nuclear (Megawatt) | $u8(t);\ u8(t-1)$ |
| Pumped Storage (Megawatt) | $u9(t);\ u9(t-2);\ u9(t-24)$ |
| Hydroelectric Power (Megawatt) | $u10(t);\ u10(t-2);\ u10(t-24)$ |
| Biomass (Megawatt) | $u11(t);\ u11(t-1)$ |
| Combined Cycle Gas Turbine (Megawatt) | $u12(t);\ u12(t-2);\ u12(t-24)$ |
| Open Cycle Gas Turbine (Megawatt) | $u13(t);\ u13(t-2);\ u13(t-24)$ |
| Average Temperature (Celsius) | $u14(t);\ u14(t-2);\ u14(t-24)$ |

The NARMAX methodology was applied to each of the seasonal datasets and an individual model was generated for each season. Table 2 presents each seasonal NARMAX model. The Spring model is denoted as Spring_N, the Summer model is denoted as Summer_N, and the Autumn model is denoted as Autumn_N. For all three seasons, price was found to have the largest ERR and it was the variable with the highest contribution for Spring_N. However, gas was also identified as a significant factor as it was the variable with the most contribution in Summer_N and Autumn_N.

Each initial seasonal model consisted of 39 input factors, which include each external factor and a number of corresponding optimal lagged terms. 17 inputs were deemed significant for Spring_N, 18 inputs significant for Summer_N, and 20 inputs were found to be significant for Autumn_N as seen in Table 2. The most significant factors that influenced day-ahead price in all three seasons were price, demand, coal, pumped storage, hydroelectric power, and average temperature. A lag of 24 hours was significant for price, demand, coal, and pumped storage showing that optimal lags do influence electricity price. To summarize out of 14 external factors, 6 of these factors are important to consider in an electricity price forecasting model.

The RMSE values displayed in Table 2 are obtained during the model validation (testing) stage using unseen data. Observing the RMSE, which determines overall model accuracy, Autumn_N had the lowest RMSE value (7.32) and Spring_N had the largest RMSE value (11.64) thus Autumn_N is most accurate at forecasting energy prices.

**Table 2.** Summary results for seasonal linear polynomial NARMAX models

| Model | RMSE Model Validation | NARMAX Model | Most Weighted | Largest ERR |
|---|---|---|---|---|
| Spring_N | 11.64 | $y(t) = 8.04 + 0.49u1(t) + 0.026u1(t-2) + 0.24u1(t-24) + 0.000017u2(t) - 0.000013u2(t-24) - 0.000080u4(t) + 0.00014u4(t-1) - 0.000011u6(t) - 0.00015u6(t-24) + 0.00026u7(t-24) + 0.00021u9(t) + 0.00015u9(t-2) + 0.00026u9(t-24) + 0.00019u10(t-2) + 0.012u13(t) + 0.024u13(t-2) - 0.19u14(t-24)$ | Price (0.49) | Price (59.65) |
| Summer_N | 9.90 | $y(t) = 22.84 + 0.30u1(t) + 0.080u1(t-24) + 0.000042u2(t) + 0.0000027u2(t-2) - 0.000069u2(t-24) + 2.38u3(t) + 1.78u3(t-2) + 0.000031u6(t) - 0.00033u6(t-24) + 0.00042u9(t) + 0.00028u9(t-2) + 0.00011u9(t-24) - 0.000038u10(t) + 0.0011u10(t-2) + 0.0014u10(t-24) - 0.00029u11(t) - 0.012u13(t-24) + 0.44u14(t)$ | Gas (2.38) | Price (34.14) |
| Autumn_N | 7.32 | $y(t) = -6.91 + 0.40u1(t) + 0.21u1(t-24) + 0.00012u2(t) + 0.000048u2(t-24) + 1.17u3(t) - 0.00012u4(t) + 0.000059u4(t-1) - 0.000078u5(t) - 0.000021u5(t-24) - 0.000078u6(t) - 0.000077u6(t-24) - 0.00015u8(t) + 0.0000056u9(t) + 0.000091u9(t-24) - 0.0010u10(t) + 0.00090u10(t-24) + 0.000044u11(t) - 0.000081u12(t) - 0.000058u12(t-24) + 0.37u14(t-24)$ | Gas (1.17) | Price (50.39) |

The NARMAX model validation for Autumn_N is displayed in Figure 1 and illustrates that the predicted day-ahead price closely matches the majority of actual values. Nonetheless, Autumn_N misses the high peaks and therefore is under-fitting.

We also implemented statistical seasonal day-ahead forecasting models using SARIMAX. First, the parameter order terms ($p$, $q$, $P$, and $Q$) were selected by applying the AIC technique outlined in Equation (3). Each of the order terms were entered as a range: $p$=(0, 4), $q$=(0, 4), $P$=(0, 2), and $Q$=(0, 2), and verified with training data for each season. The lowest AIC values for each season are displayed in Table 3:

**Table 3.** Lowest AIC values for each season

| Season | Order Parameters (p,q,P,Q) | AIC |
|---|---|---|
| Spring | (3,3,0,1) | 7267.98 |
| Summer | (2,3,0,1) | 7012.96 |
| Autumn | (1,2,0,1) | 7045.64 |

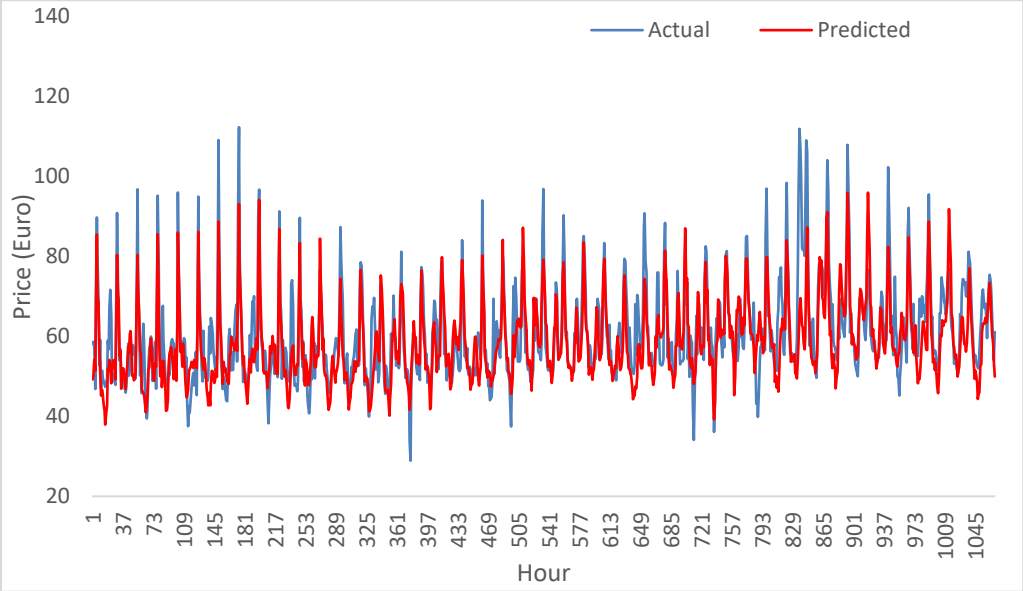



**Fig. 1.** Autumn NARMAX model

Since these data displayed volatility they are not stationary, so differencing was necessary. First difference was applied to the data by calculating the difference between the current price value and the previous price value. After first differencing the data displayed a pattern with constant mean and variance, therefore term *d* was set to 1. Since energy data are quite volatile, the seasonal pattern is unstable, so D was set to 0. Seasonality (*S*) was set to 24 as the input data were hourly records.

Therefore the SARIMAX model (p, d, q)(P, D, Q, S) implemented for Spring, denoted as Spring_S, was (3, 1, 3)(0, 0, 1, 24), displayed as:

$$Y_t = \beta_k x'_{k,t} + \varphi_1 \nabla Y_{t-1} + \varphi_2 \nabla Y_{t-2} + \varphi_3 \nabla Y_{t-3} + \theta_1 \nabla \varepsilon_{t-1} + \theta_2 \nabla \varepsilon_{t-2} + \theta_3 \nabla \varepsilon_{t-3} + + \Theta_1 \nabla^{24} \varepsilon_{t-1} + \varepsilon_t \quad (4)$$

The SARIMAX model (p, d, q)(P, D, Q, S) implemented for Summer, denoted as Summer_S, was (2, 1, 3)(0, 0, 1, 24), displayed as:

$$Y_t = \beta_k x'_{k,t} + \varphi_1 \nabla Y_{t-1} + \varphi_2 \nabla Y_{t-2} + \theta_1 \nabla \varepsilon_{t-1} + \theta_2 \nabla \varepsilon_{t-2} + \theta_3 \nabla \varepsilon_{t-3} + \Theta_1 \nabla^{24} \varepsilon_{t-1} + \varepsilon_t \quad (5)$$

Finally, the SARIMAX model (p, d, q)(P, D, Q, S) implemented for Autumn, denoted as Autumn_S, was (1, 1, 2)(0, 0, 1, 24), displayed as:

$$Y_t = \beta_k x'_{k,t} + \varphi_1 \nabla Y_{t-1} + \theta_1 \nabla \varepsilon_{t-1} + \theta_2 \nabla \varepsilon_{t-2} + \Theta_1 \nabla^{24} \varepsilon_{t-1} + \varepsilon_t \quad (6)$$

Since NARMAX is a transparent model keeping only dominant energy related factors, the seasonal SARIMAX models were refined to improve performance by only including the significant NARMAX parameters as model inputs. The new refined models are labelled as Spring_SR, Summer_SR, and Autumn_SR.

The Spring_SR model function is:

$$Y_t = -0.68\nabla Y_{t-1} + 0.22\nabla Y_{t-2} + 0.27\nabla Y_{t-3} + 0.27\nabla \varepsilon_{t-1} - 0.66\nabla \varepsilon_{t-2} - 0.43\nabla \varepsilon_{t-3} - 0.09\nabla^{24}\varepsilon_{t-1} + 42.56 \quad (7)$$

The Summer_SR model function is:

$$Y_t = 0.56\nabla Y_{t-1} + 0.08\nabla Y_{t-2} - 0.59\nabla \varepsilon_{t-1} - 0.17\nabla \varepsilon_{t-2} - 0.13\nabla \varepsilon_{t-3} - 0.05\nabla^{24}\varepsilon_{t-1} + 43.84 \quad (8)$$

The Autumn_SR model function is:

$$Y_t = 0.61\nabla Y_{t-1} - 0.77\nabla \varepsilon_{t-1} - 0.19\nabla \varepsilon_{t-2} - 0.08\nabla^{24}\varepsilon_{t-1} + 43.09 \quad (9)$$

Autumn_SR, displayed in Figure 2, shows that a refined SARIMAX is slightly better than NARMAX at predicting high peaks and thus can accurately forecast day-ahead prices, as the predicted day-ahead prices are very similar to the actual day-ahead price values.

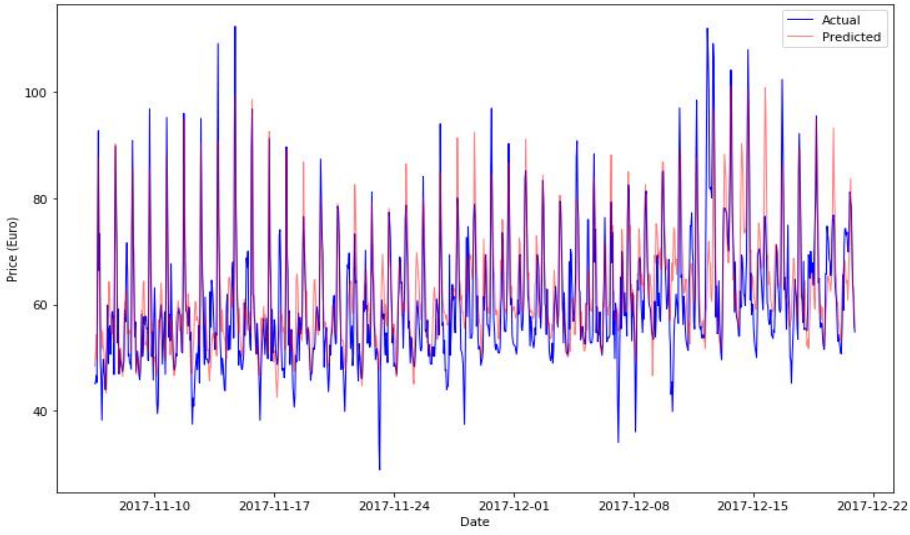



**Fig. 2.** Autumn_SR model performance

For comparison, we calculate the RMSE for each of the seasonal models for each approach. The results are presented in Table 4 where SARIMAX_R denotes the refined SARIMAX models. It can be seen that although the NARMAX models perform better than the SARIMAX models, when the significant factors determined by NARMAX are used to refine the SARIMAX model, the performance increases. Therefore, transparent

NARMAX modelling can be utilized for determining key terms in well-known statistical models to enhance overall prediction performance.

**Table 4.** RMSE values

| Model | Spring | Summer | Autumn |
|---|---|---|---|
| NARMAX | 11.64 | 9.90 | 7.32 |
| SARIMAX | 12.33 | 10.87 | 10.47 |
| SARIMAX_R | 9.48 | 8.93 | 7.29 |

## 4 Conclusion

This paper explored the use of a NARMAX polynomial to predict day-ahead electricity prices and to identify contributing external factors needed for energy forecasting. The data was split into Spring, Summer, and Autumn to determine seasonality differences. Peak lags for each input factor were identified and included to examine if lagged factors influenced prediction accuracy.

Overall, six significant factors remained dominant for each season: price, demand, coal, pumped storage, hydroelectric power, and average temperature. For the majority of these significant factors, a lag of 24 hours was dominant in influencing electricity prices. The RMSE value was lowest for Autumn highlighting that out of the three seasonal models Autumn_N was best at accurately forecasting day-ahead price.

SARIMAX models for each season were also implemented. After applying the necessary stationarity and seasonality checks the best models were determined for electricity price forecasting. Price was highly significant in all seasons, coal was significant for Spring_S, pumped storage was significant for Spring_S and Autumn_S, and hydroelectric power was significant for Spring_S and Summer_S. However, the RMSE values were higher in SARIMAX compared with NARMAX. Therefore, SARIMAX was refined to include only significant factors from NARMAX. For the refined models, price and pumped storage were significant for all seasons, coal and hydroelectric power were significant for both Spring_SR and Summer_SR, and demand was significant for Spring_SR. For all seasons, the RMSE was much lower than using the NARMAX or standard SARIMAX models.

To conclude, a NARMAX model is beneficial to distinguish the significant external factors and these factors would be best to include as inputs in a SARIMAX model to predict day-ahead electricity prices accurately. Further work will examine seasonality in more detail by looking at all seasons as well as focussing on weekends and holidays only.

**Acknowledgment.** This work was funded by DfE CAST scholarship in collaboration with Click Energy.

## References


1. Mosbah, H., El-Hawary, M.: Hourly Electricity Price Forecasting for the Next Month Using Multilayer Neural Network. Can. J. Electr. Comput. Eng. 39, 283–291 (2016). doi:10.1109/CJECE.2016.2586939
2. Amjady, N., Hemmati, M.: Energy price forecasting: Problems and proposals for such predictions. IEEE Power Energy Mag. 4, 20–29 (2006). doi:10.1109/MPAE.2006.1597990
3. Gao, G., Lo, K., Fan, F.: Comparison of ARIMA and ANN Models Used in Electricity Price Forecasting for Power Market. Energy Power Eng. 09, 120–126 (2017). doi:10.4236/epe.2017.94B015
4. Haupt, S., Kosovic, B.: Variable generation power forecasting as a Big Data problem. IEEE Trans. Sustain. Energy. 8, 1–1 (2016). doi:10.1109/TSTE.2016.2604679
5. Pandey, N., Upadhyay, K.G.: Different price forecasting techniques and their application in deregulated electricity market : A comprehensive study. Int. Conf. Emerg. Trends Electr. , Electron. Sustain. Energy Syst. 1–4 (2016). doi:10.1109/ICETEESES.2016.7581342
6. Cerjan, M., Matijaš, M., Delimar, M.: Dynamic hybrid model for short-term electricity price forecasting. Energies. 7, 3304–3318 (2014). doi:10.3390/en7053304
7. McHugh, C., Coleman, S., Kerr, D., McGlynn, D.: A Linear Polynomial NARMAX Model with Multiple Factors to Forecast Day-Ahead Electricity Prices. Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018. 2125–2130 (2019). doi:10.1109/SSCI.2018.8628694
8. Mirakyan, A., Meyer-Renschhausen, M., Koch, A.: Composite forecasting approach, application for next-day electricity price forecasting. Energy Econ. 66, 228–237 (2017). doi:10.1016/j.eneco.2017.06.020
9. Nogales, F.J., Contreras, J., Conejo, A.J., Espínola, R.: Forecasting next-day electricity prices by time series models. IEEE Trans. Power Syst. 17, 342–348 (2002). doi:10.1109/TPWRS.2002.1007902
10. Vijayalakshmi, S., Girish, G.P.: Artificial neural networks for spot electricity price forecasting: A review. Int. J. Energy Econ. Policy. 5, 1092–1097 (2015)
11. Li, P., Arci, F., Reilly, J., Curran, K., Belatreche, A., Shynkevich, Y.: Predicting short-term wholesale prices on the Irish single electricity market with artificial neural networks. 2017 28th Irish Signals Syst. Conf. ISSC 2017. (2017). doi:10.1109/ISSC.2017.7983623
12. Ghalehkhondabi, I., Ardjmand, E., Weckman, G.R., Young, W.A.: An overview of energy demand forecasting methods published in 2005 – 2015. Energy Syst. 8, 411–447 (2017). doi:10.1007/s12667-016-0203-y
13. Xie, M., Sandels, C., Zhu, K., Nordström, L.: A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden. 2013 10th Int. Conf. Eur. Energy Mark. (2013). doi:10.1109/EEM.2013.6607293
14. Torbat, S., Khashei, M., Bijari, M.: A hybrid probabilistic fuzzy ARIMA model for consumption forecasting in commodity markets. Econ. Anal. Policy. 58,

22–31 (2018). doi:10.1016/j.eap.2017.12.003
15. Elamin, N., Fukushige, M.: Modeling and forecasting hourly electricity demand by SARIMAX with interactions. Energy. 165, 257–268 (2018). doi:10.1016/j.energy.2018.09.157
16. Zhang, Y., Hua, X., Zhao, L.: Exploring determinants of housing prices: A case study of Chinese experience in 1999-2010. Econ. Model. 29, 2349–2361 (2012). doi:10.1016/j.econmod.2012.06.025
17. Acuna, G., Ramirez, C., Curilem, M.: Comparing NARX and NARMAX models using ANN and SVM for cash demand forecasting for ATM. Proc. Int. Jt. Conf. Neural Networks. 10–15 (2012). doi:10.1109/IJCNN.2012.6252476
18. Zito, G., Landau, I.D.: A methodology for identification of narmax models applied to diesel engines. IFAC World Congr. 16, 374–379 (2005). doi:10.3182/20050703-6-CZ-1902.00063
19. Taib, M.N.: Time Series Modelling and Prediction Using Neural Networks. Univ. Sheff. Thesis, (1993)
20. Korenberg, M., Billings, S.A. and Liu, Y.P.: An Orthogonal Parameter Estimation Algorithm for Nonlinear Stochastic Systems. Acse Rep. 307. (1987)
21. Billings, S., Coca, D.: Control Systems, Robotics, and Automation – Vol. VI - Identification of NARMAX and Related Models. VI, (2001)
22. Amisigo, B.A., van de Giesen, N., Rogers, C., Andah, W.E.I., Friesen, J.: Monthly streamflow prediction in the Volta Basin of West Africa: A SISO NARMAX polynomial modelling. Phys. Chem. Earth. 33, 141–150 (2008). doi:10.1016/j.pce.2007.04.019
23. Billing, S.A. and Voon, W.S.F.: Correlation Based Model Validity Tests for Nonlinear Models. Acse Rep. 285. (1985)
24. Vagropoulos, S.I., Chouliaras, G.I., Kardakos, E.G., Simoglou, C.K., Bakirtzis, A.G.: Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. 2016 IEEE Int. Energy Conf. ENERGYCON 2016. (2016). doi:10.1109/ENERGYCON.2016.7514029
25. Chakhchoukh, Y., Panciatici, P., Bondon, P.: Robust estimation of SARIMA models: Application to short-term load forecasting. IEEE Work. Stat. Signal Process. Proc. 77–80 (2009). doi:10.1109/SSP.2009.5278636
26. Nordpool: N2EX market prices, https://www.nordpoolgroup.com/historical-market-data
27. NationalGrid: Gas Transmission data, http://mip-prod-web.azurewebsites.net/DataItemExplorer/Index
28. Gridwatch: G.B. National Grid status, http://www.gridwatch.templar.co.uk/
29. Speedwell: Temperature data, https://www.speedwellweather.com/WeatherData

A Lotka-Volterra model for diffusion of electric vehicles in the US: competition and forecasting

The transport sector requires a reduction in CO2 emissions, so that industry, policy makers and researchers are forced to think about the diffusion of plug-in electric vehicles. Market forecasting is a well-developed field of study, but in the electric vehicle domain this is a complex task due to the relative newness of the market.
In this paper, we propose and apply a new Lotka-Volterra model to the diffusion of electric vehicles in the US market. Specifically, we study the competitive interaction between the two main market players, namely Tesla and Nissan Leaf, by considering the time series of monthly sales. After a suitable model reduction, the results of the selected model statistically confirm a significant interaction between the two and indicate that Tesla has been able to cannibalize the market of Nissan Leaf. At the same time, the model suggests that Nissan Leaf exerted a collaborative effect towards Tesla, rather a competitive one. The analysis is completed with short term forecasts and a SARMAX refinement which accounts for seasonal and autodependent components.

# Extreme Value Analysis of Power System Data

Per Westerlund* and Wadih Naim

KTH Royal Institute of Technology
School of Electrical Engineering
and Computer Science
Teknikringen 31
SE-100 44 Stockholm
Sweden
{perw,wadih}@kth.se
http://www.kth.se

**Abstract.** In the electric system, the consumption varies throughout the year, during the week and during the day. The consumption should be balanced by the production, which is not easy with solar power and wind power, as they lack storage like the dams for hydropower. Then these sources should be modelled as time series.
The time series analyzed is the solar and wind power production in Sweden during 5 months. The method is called *Peaks over Threshold* or *POT* and it calculates the frequency of peaks above a certain threshold. It also determines the distribution of the size of the peaks, which is the generalized Pareto distribution in this case.
Two different clustering methods are tried; one by Lindgren and the other one a modification of Leadbetter's method. The latter one gives the best fit. However, some ten years of data should be analyzed in order to include the seasonal effects.

**Keywords:** extreme value theory, energy production

## 1 Introduction

In a power system, production must be balanced with the power demand. Power consumption follows various fluctuation patterns on annual, weekly, and daily scales. Annually, fluctuations in consumption are driven by seasonal changes, where, for instance, power demand is higher during winter than that in summer due to the need for heating in cold climates. While on weekly and daily scales, the fluctuation pattern follows cycles of work days and weekends, as well as, daytime and nighttime.

From the perspective of a system operator, production is planned while keeping in mind the costs. Thus, renewable sources can be more favorable. Unlike

* The research is financed by Energimyndigheten (Swedish Energy Agency), Svenska kraftnät and E.on Eldistribution AB through SweGRIDS, the Swedish Centre for Smart Grids and Energy Storage. http://www.swegrids.se

traditional power generation, renewables do not have a stable availability of production. Solar power and wind power, in particular, lack the storage capacity that is available in dams of hydropower plants. Then, it is of interest for the operator to know how these sources are distributed over time. Moreover, the maximum values of produced power are needed in order to guarantee that the transmission capacity is sufficient.

## 2 Data

The time series are the hourly production of electric power in Sweden in total as well as divided according to sources. Also the consumption and the import and export are included. The source is Svenska Kraftnät, which is responsible for the transmission of electric power across Sweden as well as keeping the frequency by matching consumption with production. The period of the data covers 151 days (January 2019 – May 2019). [1]

The data chosen are the solar power and wind power production. The respective time series are displayed in figures 1 and 2 and their distributions in figures 3 and 4.

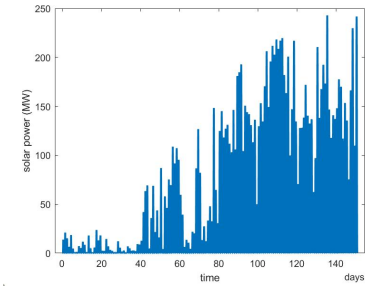



**Fig. 1.** Time series of the solar power production in Sweden.

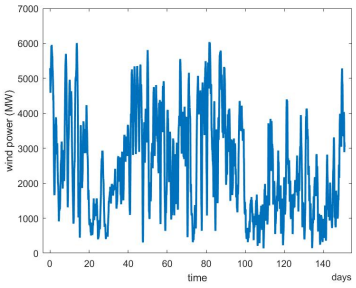



**Fig. 2.** Time series of the wind power production in Sweden.

## 3 Method

The method is called *Peaks over Threshold* or *POT* and it calculates how often peaks above a certain threshold appear and what is the distribution of the size of the peaks. It is described in [3, p 64–65] as:

- Get a large number of observations $x_1, x_2, \ldots, x_N$.
- Choose only those that exceed the threshold $A$ and denominate the excesses as $y_1, y_2, \ldots, y_n$ and the times as $t_1, t_2, \ldots, t_n$.

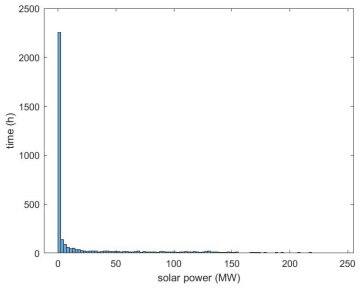



**Fig. 3.** Histogram of the solar power production in Sweden.

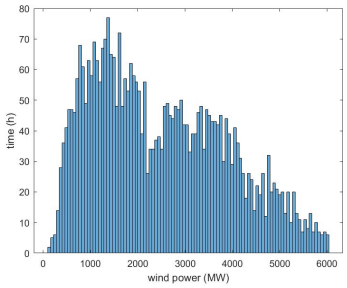



**Fig. 4.** Histogram of the wind power production in Sweden.

- Estimate the intensity $\mu_A$ of the Poisson process defined as exceeding the threshold.
- Estimate the shape parameter $k$ and the scale parameter $B$ for a generalized Pareto distribution.
- Determine the generalized extreme value distribution based on the previous estimations.
- Calculate the probability that the maximum exceeds a certain level during the period, which serves as the unit of the intensity $\mu_A$.

The explanation is that the extreme value distribution can be regarded as a Poisson process giving time points where the value will exceed a certain threshold together with a generalized Pareto distribution for the amount exceeding that threshold. [2]

The events in the Poisson process should be independent of each other. Since it is probable that the threshold will be exceeded for some samples away from the peak, Lindgren proposes that only the peak should be recorded, not the samples above the threshold around it. [3, p 65] Another way to get the independence between the events is to take the maximum during blocks of several time points, as explained in more detail in [2, p 360].

Here Lindgren's method as well as Leadbetter's method will be used. However the actual position of the peak in the block will be used instead of the first time point as proposed by Leadbetter. As the studied time series are not long enough for a Poisson modelling of the position of the peak, this part will not be carried out.

### 3.1 Estimation of the generalized Pareto distribution

The distribution of the generalized Pareto distribution with shape parameter $k$ and scale parameter $B$ is given by:

$$F(x; B, k) = 1 - \left(1 - k\frac{x}{B}\right)^{\frac{1}{k}}$$

An interesting property is that the mean excess above a certain level $A$ is a linear function of that level:

$$E[X - A|X \geq A] = \frac{B - kA}{1 + k}$$

It will be used to check that the data follow a generalized Pareto distribution by plotting the mean of the excess for different thresholds. The parameters can be estimated from that graph.

## 4 Results

The results are divided according to the clustering carried out. Sections 4.1 and 4.2 show the results of clustering using methods by Lindgren and Leadbetter respectively.

### 4.1 Peaks together with surrounding samples above the threshold

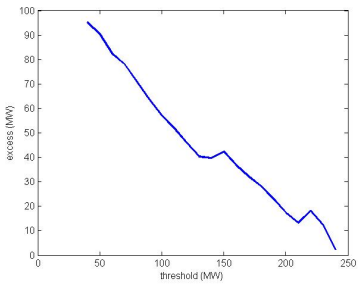



**Fig. 5.** Mean of the excess for the solar power production above different thresholds with Lindgren's clustering.

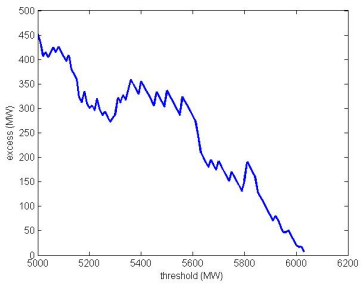



**Fig. 6.** Mean of the excess for the wind power production above different thresholds with Lindgren's clustering.

Figs. 5–6 show the test for the generalized Pareto distribution when the surrounding samples above the peak are grouped together with the peak according to Lindgren.

### 4.2 Maximum of each block

Figs. 7–8 show the test for the generalized Pareto distribution when the peaks are based on the maximum of each day, a modification of Leadbetter's clustering.

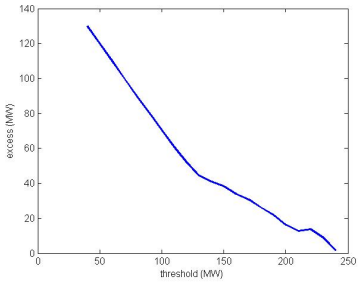



**Fig. 7.** Mean of the excess for the solar power production above different thresholds with a modified variant of Leadbetter's clustering.

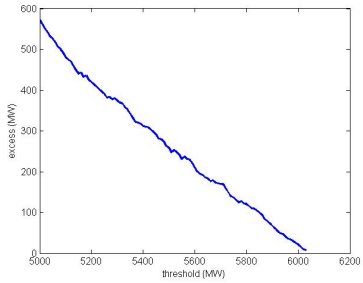



**Fig. 8.** Mean of the excess for the wind power production above different thresholds with a modified variant of Leadbetter's clustering.

## 5 Discussion and future work

Taking the maximum of a period is better as it makes the curves smoother, which means that the generalized Pareto distribution describes the data better. For the wind power production the effect is large, from a variation of 150 MW to just around 10 MW. The solar power production gives fairly smooth curves, although with some peaks. With the modified Leadbetter clustering, it gets smoother, but there is still a knee.

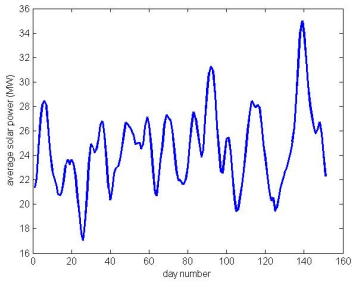



**Fig. 9.** Daily averages of the solar power production.

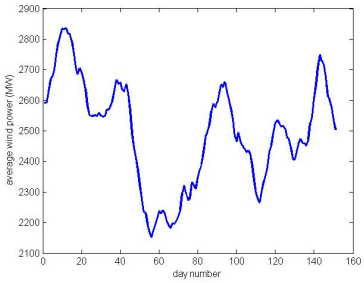



**Fig. 10.** Daily averages of the wind power production.

One explanation of that knee is that there is natural seasonal variation in the solar production, which should be eliminated. This requires longer periods of data, at least ten years, which also would enable the estimation of the intensity of the appearance of peaks, which is the parameter of the Poisson process. The daily averages of solar and wind power are shown in Figs 9–10. The solar production

shows an increasing trend, which is natural since the time period goes from winter to summer.

## 6 Conclusion

The distribution of the size of the peaks in the solar and wind power production in Sweden can be estimated by a generalized Pareto distribution. A modified variant of Leadbetter's clustering appears to be better than the Lindgren's clustering as it provides smoothers curves. However longer time series should be used together with elimination of seasonal variation especially in the case of solar power. Analyzing the production data over a long time period (around 10 years) would provide essential information to the system operator to prevent overloading of trasmission lines, and to make more informed decisions when it comes to planning and balancing power production.

## References

1. Svenska kraftnät. *Förbrukning och tillförsel per timme (normaltid).* `https://www.svk.se/aktorsportalen/elmarknad/statistik/`
2. M.R. Leadbetter. *On a basis for Peaks over Threshold modeling.* Statistics & Probabilities Letters, 12(4):357362, October 1991.
3. Georg Lindgren. *Hundraårsvågen – något nytt?* In Göran Grimvall and Olof Lindgren, editors. *Risker och riskbedömning.* Studentlitteratur and Ingenjörsvetenskapsakademien IVA, 1995

# Reconstruction of the transition probability density function from persistent time series

Zbigniew Czechowski

Institute of Geophysics, Polish Academy of Sciences, Ks. Janusza 64, 01-452 Warsaw, Poland
zczech@igf.edu.pl

**Keywords:** probabilistic forecast; persistent processes; transition probability density; Langevin equation; reconstruction procedure.

In some models (e.g., diffusion Markov models) the probabilistic forecast is given by the transition probability density function. For the case of the standard Langevin equation (which describes a time evolution of Markov processes) Siegert et al. (1998) introduced the numerical procedure of reconstructing from time series the drift and diffusion functions that are creating the transition probability density for small times. Then the method was developed in other papers (see Friedrich et al. 2011).

However, the standard Langevin equation describes Markov processes only. In this work we introduce the generalized discrete Langevin equation for some class of non-Markov processes, namely for persistent time series of order $p$. Therefore, non-local effects must be considered. We assume that the next state of the process is dependent not only on the present state but also on signs of $p$ previous jumps. To this aim, the standard discrete Langevin equation is modified by introducing a new random function which determines the sign of the diffusion term. The function depends on the vector random variable (i.e., the chain of $p$ previous signs), the random scalar variable with the uniform distribution in [0, 1] and on the vector persistence parameter (with $2^p$ components). The term is keeping the tendency of increase/decrease of the process in the next step according to given persistence parameters. When all the parameters are equal to 0.5 the modified equation reduces to the standard Langevin equation. The proposed model is a significant extension of our previous approach (Czechowski 2016) in which persistent processes of order $p = 1$ were taken into account. The generalization opens a wide possibilities of nonlinear modeling of data in which persistence and antipersistence of different orders can be mixed in a time series under investigation.

In order to construct the transition probability density function the forms of drift and diffusion functions are needed. The standard procedure (Siegert et al. 1998, Friedrich et al. 2011) of reconstruction of the Langevin equation from time series leads to the proper estimation of the diffusion function but to the wrong reconstruction of the drift function in the case of the modified equation for persistent processes. To estimate the deviation in the drift we propose a new reconstruction procedure in which the standard procedure is used three times. The algorithm can be summarized in five steps. At the beginning the vector persistence parameter is estimated by using a histogram method. In the second step the standard procedure is applied to the input persis-

adfa, p. 1, 2011.

tent time series. As the result first reconstructions of the diffusion function and the drift function are obtained, however, the reconstructed drift is deviated from the input drift. In the next step a new time series generated by the modified Langevin equation with the estimated persistence parameter and first reconstructions of the drift and diffusion functions is treated as the input to the second use of standard procedure. In this way the second reconstruction of the drift function is obtained. The fourth step of the procedure is similar to the third step, however, here the second reconstruction of the drift (instead of the first reconstruction) is applied. At the result the third reconstruction of the drift function is found. The last step of the procedure bases on the assumption (which has been verified analytically for the case $p = 1$) that ratios of deviations are dependent on the persistence parameter only. Therefore, the ratio calculated from deviations found in the third and fourth step is used for estimation of the deviation between the unknown input drift function and the first reconstruction. This leads to the final reconstruction of the drift function from the input $p$-persistent time series.

In order to test an efficiency of the procedure many time series were generated by using the modified Langevin equation with different drift and diffusion functions and different persistences. This enables to compare the input functions and parameters to the reconstructed ones. A good efficiency of the modified reconstruction procedure has been shown.

Having the proper forms of reconstructed drift and diffusion functions enables derivation of the short-time transition probability density function. For the case of Markov processes the function has a Gaussian form, however non-Markovian features of persistent processes make the problem more complex. Therefore, a correction term appears in the formula for short-time transition probability density function. For the persistence of order 1 (i.e., for $p = 1$) the correction term can be derived analytically but for higher orders numerical estimations are necessary. It should be underlined that the presented method applied to forecasting time series generates a probability distribution for the next point, rather than a single point estimate as in autoregression. The parameters in the modified Langevin model are dynamically estimated from past data.

An important advantage of the proposed approach is that it offers simultaneously the reconstruction of the stochastic model of the phenomena under investigation and the method of probabilistic forecast.

Acknowledgements

This work has been financed by the project of the National Science Centre (contract No. 2016/21/B/ST10/02998).

1. Siegert S., R. Friedrich, and J. Peinke, 1998, Analysis of data sets of stochastic systems, Phys. Lett. A **243**, 275-280.
2. Friedrich R., Peinke J., Sahimi M., Reza Rahimi Tabar M., 2011. Approaching complexity by stochastic methods: From biological systems to turbulence, Physics Reports 506, 87-162.
3. Czechowski Z., 2016, Reconstruction of the modified discrete Langevin equation from persistent time series, CHAOS **26**, 053109.

# A covariance function for time dependent Laplacian fields in 3D

Gyorgy Terdik
University of Debrecen (Hungary)

## Abstract

A homogeneous and isotropic Laplacian field is considered. We compare the decays of the covariance functions of an AR model in 1D, 2D and 3D. It is well known that the decay is exponential for a stationary AR time series and it has slower decay in 2D by a celebrated result of Whittle. We show that the decay is again exponential in 3D. A covariance function for spatio-temporal 3D-Lapalcian fields will also be considered in time-frequency domain, see Subba Rao -Terdik: A New Covariance Function and Spatio-Temporal Prediction, JTSA 12017.

# Do Google Trends Forecast Bitcoins? Stylized Facts and Statistical Evidence

Argimiro Arratia[1] and Albert X. López Barrantes[2]

[1] Universitat Politècnica de Catalunya,
Dept. of Computer Science, Barcelona, SPAIN
argimiro@cs.upc.edu
[2] Universitat Autónoma de Barcelona, SPAIN
albertxavier.lopez@gmail.com

**Abstract.** In early 2018 Bitcoin prices peaked at US$ 20,000 and, almost two years later, we still continue debating if cryptocurrencies can actually become a currency for the everyday life or not. From the economic point of view, and playing in the field of behavioral finance, this paper analyses the relation between Bitcoin prices and the search interest on Bitcoin since 2014. We questioned the forecasting ability of Google Bitcoin Trends for the behavior of Bitcoin price by performing linear and nonlinear dependency tests, and exploring performance of ARIMA and Neural Network models enhanced with this social sentiment indicator. Our analyses and models are founded upon a set of statistical properties common to financial returns that we establish for Bitcoin, Ethereum, Ripple and Litecoin.

**Keywords:** Google Trends, Bitcoin, causality, ARIMA, Neural Networks

## 1 Introduction

Bitcoin is the most popular and prominent cryptocurrency in the world. The first design was published in 2008 under the pseudonym of Satoshi Nakamoto [13]. A currency where everyone from anywhere can execute transactions with no need of a traditional financial institution involved in the process generated a huge expectation around the world rising Bitcoin prices to a peak of US$ 20,000 in January 2018. However, the low number of transactions per second that are able to support and the non-recognition as a currency by most companies and governments, caused a big drop in the price to the current US$ 3,000 price. A new economic bubble in the exact sense of the term: a huge increase on expectations on a short period of time, often coming from the irrational decisions we do as human beings. We can indeed observe that a classical financial bubble have formed on the Bitcoin price time series (Figure 1 left). And with the intention to build upon research on the use of big data in social media to construct predictors for various economic variables or social events as in, for example, [1, 3, 8] and most relevant for this project the paper [6], we looked at the trend of the topic

"bitcoin" in Google Trends for the last four years, and obtained the time series depicted in Figure 1 (right).

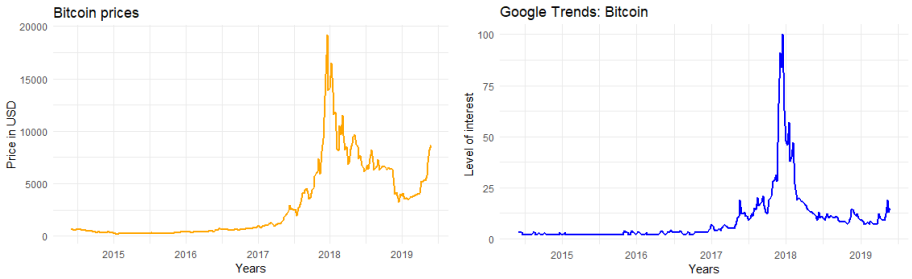



**Fig. 1.** Bitcoin prices and Google Trends on topic "bitcoin" series

The strong resemblance among both time series may lead one to think of the Google Bitcoin Trends (GBTrends) as predictor of Bitcoin price behavior. But being cautious about drawing conclusion from pictures and aware of the failure of the Google Flu Trends [11], we turn to traditional statistical methods to prove hypothesis.

## 2 Stylized empirical facts and statistical issues

Financial time series have a common set of characteristics that should be studied and considered before any further analysis, modelling or conclusions. These are commonly known as stylized empirical facts [4], and knowing which of these hold will guide on proper modeling of the financial asset. However, a basic underlying hypothesis is that of stationarity, and more often than not, these data sets are non stationary which means that we have to transform them, usually by considering returns, in order to recognize some statistical properties of the data which remain invariant over time, and so that modeling is possible.

### 2.1 Data gathering

We obtained Bitcoin prices from *CoinMarketCap.com*. The data set contains Bitcoin daily prices since 2014 with access to Open and Closing prices as well as volume traded and the total market capitalization over time. On the other hand, Google Trends on the word "Bitcoin" can be obtained through the R package *gtrendsR* where one can query everything with the same parameters over time and geography as in the web page of Google Trends. One has only to take into account the minimum time scale one can get from Google Trends, which for this paper we got a weekly time scale since 2014.

### 2.2 Stationarity

Kendall [10] was the first one to realize financial time series are seldom stationary. To get close to be stationary, or at least to be *second-order stationary*, a common technique is to apply successive differences to the series. Hence, it is recommendable to work instead with the series logarithmic returns: $r_t = \log\left(\frac{P_t}{P_{t-1}}\right)$.

There are several tests for stationarity and a few for second order stationarity; a survey of the former kind of stationarity and a proposal for the latter can be seen in [15]. In this work we applied two tests for second order stationarity: the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test with null hypothesis that an observable time series is stationary around a deterministic trend; and the Priestley-Subba-Rao (PSR) test that investigates how "non-constant" the time-varying Fourier spectrum of the series is. The KPSS is implemented in the R-package `tseries` test, and PSR is implemented in `fractal` package.

**Table 1.** Priestley-Subba-Rao (PSR) test

| **series** | **p-value for T** |
|---|---|
| Bitcoin returns | 3.74812e-05 |
| Ethereum returns | 0.001743593 |
| Ripple returns | 2.279732e-12 |
| Litecoin returns | 4.931523e-08 |
| GBTrends returns | 0 |

We ran both tests for all four cryptocurrencies and GBTrends returns, and for all the null hypothesis of stationarity was rejected. Hence modeling is in principle in this context not statistically well-founded.

### 2.3 Aggregational Gaussianity

One of the most used conventions when working with financial data is the assumption that returns are log-normally distributed, which is equivalent to the assumption that log-returns are normally distributed. Early in 1953 Kendall, Mandelbrot in 1960 or Fama in 1965, among other researchers, signaled the non normal distribution of asset returns and the heavy tails [10, 5]. However, it is remarkably important that distributions become closer to normal when the timescale increases. This convergence in distribution towards normality as timescale increases is called Aggregational Gaussianity and is widely documented across assets around the world. In our case, we checked for this phenomenon in all cryptocurrencies returns series.

For Bitcoin daily returns we found the typical leptokurtic distribution, sharp peaked and heavy tailed, which is far from being normally distributed. However, once we increase the timescale to weekly samples, the distribution gets closer

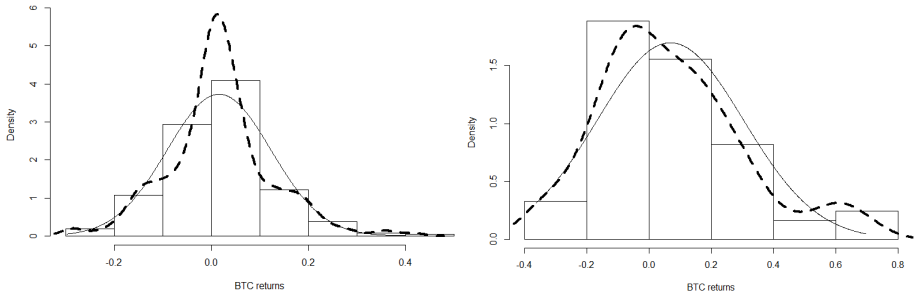



**Fig. 2.** Weekly (left) and monthly (right) bitcoin returns estimated density (dashed line) and normal density fit (solid line)

to a normal distribution. It is no until we get the monthly returns distribution when we closer to a normal distribution. On the timescale of weekly returns, if we try to fit the GBTrends series we get a similar shape as the weekly Bitcoin prices, far from a normal shape with considerable fat tails and a considerable peak, which can be observed also in the skewness comparative at the end.

However, a visual analysis is not enough to say if the aggregational gaussianity is happening on cryptocurrency's returns. To properly check this, we will use Shapiro-Wilk and Jarque-Bera normality tests on different time scales of Bitcoin, Ethereum, Ripple and Litecoin returns. The null hypothesis of Shapiro-Wilk normality test is data is normally distributed, and the null hypothesis of Jarque-Bera test is a joint hypothesis of $skewness = 0$ and excess $kurtosis = 0$.

Overall, Bitcoin returns is in line with literature on Aggregational Gaussianity, the more we increase the timescale, the closer we get to have log normal returns, specifically when we reach a monthly times scale. For the other three cryptocurrencies results are more extreme, since we only get normality once we reach yearly returns in Jarque-Bera test. Since our data starts at 2014 for all cryptocurrencies and GBTrends, yearly returns are calculated on 5 observations, which invalidates any results of these tests of this timescale.

The work by Chan et al [2] goes a step further by fitting non-normal distributions to each of the cryptocurrencies. They find that for Bitcoin and Litecoin, the generalized hyperbolic distribution gives the best fit, while the other cryptocurrencies return distributions are better fitted by the normal inverse Gaussian, generalized $t$ and Laplace distributions.

### 2.4 Autocorrelations

We now check the possibility of self linear association of the cryptocurrencies return time series, by computing the ACF and PACF. The auto-correlation function (ACF) gives us the values of autocorrelation between the time series and its lagged values. The partial auto-correlation function (PACF) gives the correlation values of the residuals with the next lag value.

**Table 2.** Normality tests

| **Data** | **Shapiro-Wilk** | **Jarque-Bera** |
|---|---|---|
| | $p$-value | $p$-value |
| Weekly returns GBTrends | 0.00001 | 0.00001 |
| Daily returns Bitcoin | 0.00001 | 0.00001 |
| Weekly returns Bitcoin | 0.00006 | 0.00001 |
| Monthly returns Bitcoin | 0.03808 | 0.07850 |
| Quarterly returns Bitcoin | 0.00633 | 0.01168 |
| Yearly returns Bitcoin | 0.00126 | 0.20780 |
| Daily returns Ethereum | 0.00001 | 0.00001 |
| Weekly returns Ethereum | 0.00001 | 0.00001 |
| Monthly returns Ethereum | 0.00002 | 0.00001 |
| Quarterly returns Ethereum | 0.00004 | 0.00001 |
| Yearly returns Ethereum | 0.00103 | 0.40190 |
| Daily returns Ripple | 0.00001 | 0.00001 |
| Weekly returns Ripple | 0.00001 | 0.00001 |
| Monthly returns Ripple | 0.00001 | 0.00001 |
| Quarterly returns Ripple | 0.00001 | 0.00001 |
| Yearly returns Ripple | 0.00002 | 0.16900 |
| Daily returns Litecoin | 0.00001 | 0.00001 |
| Weekly returns Litecoin | 0.00001 | 0.00001 |
| Monthly returns Litecoin | 0.00001 | 0.00001 |
| Quarterly returns Litecoin | 0.00001 | 0.00001 |
| Yearly returns Litecoin | 0.00009 | 0.17430 |

Visualizing the ACF and PACF for Bitcoin and Google Bitcoin Trends returns, we can not observe any significant auto-correlations in both series (Figure 3). These results are in line with financial literature, where it is a common characteristic the lack of auto-correlation, and widely supported by "The Efficient Market Hypothesis". In a competitive market where participants use all available information, market prices should be very close to the intrinsic value of the company, leaving very few opportunities to arbitrage [5].

Computing ACF and PACF for the other top three cryptocurrencies, we see that Ripple and Litecoin show no significant autocorrelation, while Ethereum have some positive autocorrelations around lag 3 (Figure 4).

### 2.5 Volatility clustering

Volatility clustering refers to the observation, first noted by Mandelbrot [12], that "*large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes.*" This is tested computing the ACF and PACFs of the squared of returns. It is generally expect that log returns to be serially uncorrelated, but the squared log returns to show significant autocorrelations. This is the case for the four cryptocurrencies but not for Google

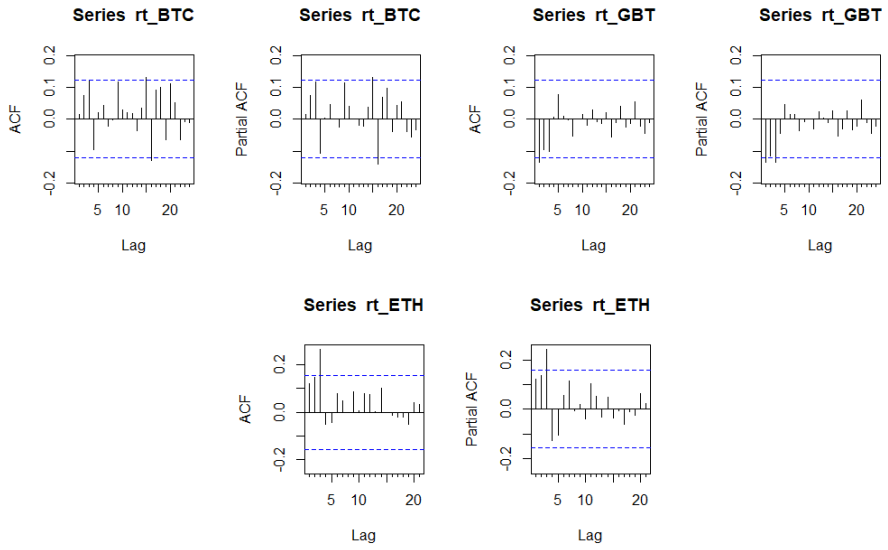



**Fig. 3.** ACF and PACF of Bitcoin, GBTrends and Ethereum returns

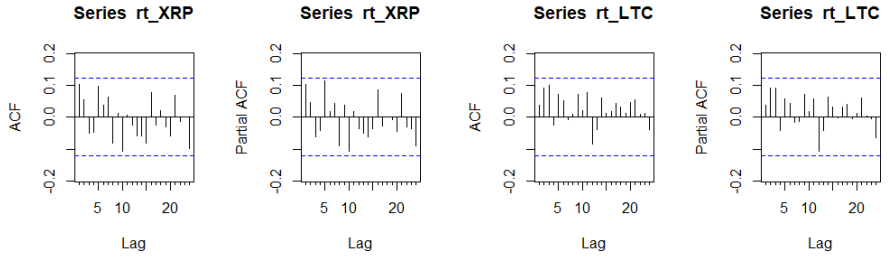



**Fig. 4.** ACF and PACF of Ripple and Litecoin returns

Bitcoin Trends: while Bitcoin, Ethereum, Ripple and Litecoin squared log returns show significant autocorrelations, GBTrends squared returns does not.

### 2.6 Causality

When analyzing relationships between time series, correlation only captures the linear dependency between two variables, but one would like to know in which direction the information flows from one series to the other. In other words, if Bitcoin prices and GBTrends are correlated we want to know which one causes the move of the other. In this context potential outcomes could be:

a) The interest in Bitcoin around the world and captured by Google Trends is actually a good proxy for measuring the social interest on the cryptocurrency triggering a Bitcoin price movement.
b) Big swings on Bitcoin prices create media and social attention triggering an increase in the interest of people around Bitcoin.
c) A combination of both events, were both series move at the same time and direction, up or down.
d) There is no relation at all between them and none of them influences the other.

We measure causality using Granger causality test [7]. The basic idea of Granger causality is that *$X$ causes $Y$, if $Y$ can be better predicted using the histories of both $X$ and $Y$ than it can by using the history of $Y$ alone*. Formally one consider a bivariate linear autoregressive model on $X$ and $Y$, making $Y$ dependent on the history of $X$ and $Y$, together with a linear autoregressive model on $Y$, and then test for the null hypothesis of "$X$ does not causes $Y$", which amounts to test that all coefficients accompanying the lagged observations of $X$ in the bivariate linear autoregressive model are zero. Then, we can evaluate the null hypothesis through an F-test. To perform this test on our time series we have used the R package `MSBVAR` which provides methods for estimating frequentist and Bayesian Vector Autoregression (VAR) models and other tools such as the Bivariate Granger Causality Test.

We perform this test at four different epochs marked by full calendar years 2015, 2016, 2017, 2018, and sampling weekly returns. Lag lengths considered to compute this test were the first 4 lags, to cover any autocorrelation through the month. We run the causality test for GBTrends, Bitcoin, Ripple and Litecoin returns (Ethereum has not been included in this analysis because of the lack of historical data, since it started to trade in late 2015, and low volume of trade).

Table 3 shows the results for GBTrends and Bitcoin for 2017. We see significant causality from GBTrends to Bitcoin in all four lags. This is also the case for the year 2015 (although only for lag 1), but not for 2016 and 2018 where rather contemporaneous correlation (causality in both directions) has been shown at all lags. A quick explanation of these results is that as the hype for Bitcoin was building up, both GBTrends and Bitcoin were moving each other, but as the Bitcoin prices started to rise exponentially news started to anticipate the

next jump until the peak in 2017-2018. At that point the curiosity on the topic was globally widespread, leading to an increase in searches in Google on this cryptocurrency due to its incredibly high price.

**Table 3.** Granger Tests results for 2017

| **Lag length used** | **1 lag** | **2 lags** | **3 lags** | **4 lags** |
|---|---|---|---|---|
| Causal relations | $p$-val | $p$-val | $p$-val | $p$-val |
| GBTrends to BTC | 0.2740 | 0.1156 | 0.1052 | 0.1985 |
| BTC to GBTrends | 0.0030 | 0.0130 | 0.0168 | 0.0342 |

Table 4 shows causality results for GBTrends, Ripple and Litecoin for year 2017. Here we observe causality in both directions (GBTrends to and from the cryptocurrencies). This situation is more or less the same on the other epochs considered. A plausible explanation could be that these two cryptocurrencies do not induce massive searches, and are somewhat surrogates for Bitcoin.

**Table 4.** Granger Tests results for 2017

| **Lag length used** | **1 lag** | **2 lags** | **3 lags** | **4 lags** |
|---|---|---|---|---|
| Causal relations | $p$-val | $p$-val | $p$-val | $p$-val |
| GBTrends to XRP prices | 0.1687 | 0.3184 | 0.2668 | 0.3486 |
| GBTrends to LTC prices | 0.9368 | 0.7929 | 0.8024 | 0.9205 |
| BTC prices to XRP prices | 0.3262 | 0.6521 | 0.6598 | 0.3956 |
| BTC prices to LTC prices | 0.3862 | 0.6928 | 0.8505 | 0.8956 |
| XRP prices to GBTrends | 0.8432 | 0.7538 | 0.8005 | 0.8803 |
| LTC prices to GBTrends | 0.3827 | 0.7749 | 0.2705 | 0.4101 |

### 2.7 Neglected nonlinearity

A multivariate test of nonlinearity to ascertain if two time series are nonlinearly related can be achieved with the neural network test for neglected nonlinearity developed by White [14]. The basic idea is to perform a test of the hypothesis that a given neural network defines a perfect mapping between its input and output and that all the errors are due to randomness. For our experiments we use the Teräsvirta linearity test, presented in [16] and based on Whites neural network test for neglected nonlinearity. An implementation of this algorithm is available in the `tseries` R library.

Again we only have space to show results for 2017 in Table 5. The null is the hypotheses of linearity in mean. Then, results from Teraesvirta test indicate

existence of a non-linear map from GBTrends to Bitcoin. The presence of a non-linear relation justifies the use of non-linear models such as neural networks that profit from all those existing relations between time series.

**Table 5.** Terasvirta Neural Network Tests results

| **2017** | **X-squared** | ***p*-value** |
|---|---|---|
| Bitcoin to GBTrends | 5.8703 | 0.0531 |
| GBTrends to Bitcoin | 2.8758 | 0.2374 |
| Bitcoin to Ethereum | 0.002 | 0.9989 |
| Bitcoin to Ripple | 2.111 | 0.3479 |
| Bitcoin to Litecoin | 0.9126 | 0.6336 |
| GBTrends to Ethereum | 0.7732 | 0.6794 |
| GBTrends to Ripple | 0.1757 | 0.9159 |
| GBTrends to Litecoin | 3.0942 | 0.2129 |

## 3 Modelling

To sum up, we have seen the stylized facts for Bitcoin prices (and other three cryptocurrencies) and Google Bitcoin Trends. As expected, both of them share most characteristics seen in stock market literature:

- Both time series are non stationary, so further statistical analysis are performed on their returns time series. As well as Bitcoin, Ethereum, Ripple and Litecoin are not stationary.
- Bitcoin returns series converge to normality when increasing the sampling period. The other three major cryptocurrencies Ethereum, Ripple and Litecoin do not converge to normality.
- Bitcoin prices and GBTrends returns time series do not experience autocorrelation with their first few lags.
- GBTrends has a significant causal effect on Bitcoin price changes at certain epochs (2015 and more strongly at 2017), so there could be some value in GBTrends as predictor. Other epochs, and for most of the cryptocurrencies we observe contemporaneous correlations, among themselves and with GBTrends.

With this information, our approach in this section is to test predictability on top of some models. To do so we will be using a control model, only using past information from Bitcoin returns and comparing errors against models using GBTrends series as external variables to predict. We used Bitcoin price changes on a weekly timescale. This is forced by the fact that we could only get Google Trends data on a weekly time scale.

### 3.1 ARIMA back-testing

First, we start using an ARIMA model to predict Bitcoin prices only using past observations, and an additive model including GBTrends data as external variable. ARIMA models are the most general class of models for forecasting time series, so it is going to be good as a base model for later iterations. To fit the ARIMA models we have used the auto.arima() function from the R library *forecast* developed by Hyndman [9]. This function returns the best ARIMA model according to either AIC, AICc or BIC value. The function conducts a search over possible model within the order constraints provided. To proceed with the model comparison, we fit a base model only using Bitcoin returns and later a second one adding GBTrends as an external variable. We fitted the data year by year from 2015 to 2018 on different models, and averaged the error measures.

**Table 6.** arima model comparison

| **models** | **AIC** | **RMSE** | **MAE** |
|---|---|---|---|
| ARIMA | -90.23 | 592.31 | 343.41 |
| ARIMA + GBTrends | 688.22 | 525.61 | 342.58 |

Using ARIMA models, we see a decrease in errors when adding GBTrends as external regresor, although the model fit is not as good according to AIC.

### 3.2 Neural Networks

We modelled with feed-forward neural networks to profit from non linear relations we observed in the neglected nonlinearity section. As we are dealing with time series data, lagged values of them can be used as inputs to a neural network. This particular version of models are called neural network autoregression or NNAR models. We used the function nnetar() from the R package *forecast* which fits feed-forward neural networks with a single hidden layer and lagged inputs as well as external variables. In our case, we choose lags 1 to $p = 5$ in order to consider the first 5 lags. Notation on this models is NNAR$(p, k)$ which declares a model that uses up to the last $p$ observations and $k$ neurons in the hidden layer. We fixed the size of the hidden layer to $k = 4$. We note that although the function NNAR allows for tuning the $p$ and $k$ parameters, we do not use that feature as we are interested in fixing a base neural network model and test if that model improves forecasting with the addition of GBTrends. Hence, as in previous experiment, we repeat the procedure fitting a base model only using Bitcoin returns and later a second one adding GBTrends as an external variable. We fit the data year by year on different models, as we did in the stylized facts, and finally we averaged the error measures:

One can observe that using neural networks we get considerably lower errors than modelling with ARIMA. One can also see that when introducing GBTrends

**Table 7.** neural networks model comparison

| **NNAR(p,k)** | **RMSE** | **MAE** |
|---|---|---|
| NNAR(5,4) | 392.7452 | 210.73 |
| NNAR(5,4) + GBTrends | 285.38 | 151.78 |

to predict Bitcoin prices, errors decrease even more. This is in line with the results obtained in Teraesvirta test of existence of a non-linear map from GBTrends to Bitcoin. Hence the NNAR model enhanced with GBTrends shows better forecasting power for Bitcoin than NNAR alone and ARIMA based models.

## 4 Conclusions

In the past years, Bitcoin and the world of cryptocurrencies have gained some level of establishment. Some people see them as the new world currency far from the intervention of governments and central banks. For others, cryptocurrencies are considered as a new type of commodity like gold which can offer some protection against inflationary cycles. And for the rest, just another creation from the digital age we are living right now, as many other trends we have seen through the past in many other fields. The main motivation for this work was to explore the idea of using Google Trends data to forecast Bitcoin prices, under the hypothesis that Google Trends could be a good proxy of the interest of people around a specific topic in internet. In order to properly face this question we built a proper study around the main statistical issues to be covered in order to build a solid basis prior to any modeling and give more consistency to our results.

From our analysis of stylized empirical facts, the most relevant fact we found is the causal relation from GBTrends to Bitcoin in some years, which signaled the potential for using this source of data to improve accuracy when forecasting Bitcoin prices. After a careful consideration of relations found in the study of linear and non linear dependencies, we fitted a linear model (ARIMA) and a non linear model (Neural Network) trying to profit from the initial idea of using GBTrends as an alternative source of data from a social sentiment point of view. Doing so, we found a significant reduction in the out-of-sample error in the neural network model when introducing this new variable. There is no much improvement in the case of forecasting with autoregressive linear models. All in all we found evidence to believe that Google Bitcoin Trends could serve as predictor of Bitcoin prices, although not consistently through time.

From an economic perspective, these results could be explained following the supply and demand model of price determination in a competitive market. If supply is relatively fixed in the long-term then demand should be the largest single contributor to Bitcoin prices, given there are no actual quarterly earnings or interest rates associated with Bitcoin. In this case demand, or at least the

interest from the demand side, can be observed from the search interest in the biggest search engine of internet, Google.

**Acknowledgments.** A. Arratia acknowledge support by grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad) and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya)

## References

1. Arias, M., Arratia, A. & Xuriguera, R. (2013). Forecasting with Twitter Data. ACM Transactions on Intelligent Systems and Technology, (TIST) 5 (1).
2. Chan, S., Jeffrey, C., Nadarajah, S., & Osterrieder,J. (2017) *A Statistical Analysis of Cryptocurrencies*. Journal of Risk and Financial Management, 10(2), 12.
3. Ciulla, F., Mocanu, D., Baronchelli, A., Gonalves, B., Perra, N., & Vespignani, A. (2012). Beating the news using social media: the case study of American Idol. EPJ Data Science, 1(1), 8.
4. Cont, R. (2000) *Empirical properties of asset returns: stylized facts and statistical issues*. Quantitative Finance.
5. Fama, E. (1965). The behavior of stock market prices. The Journal of Business.
6. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012.
7. Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. Econometrica 37:424-438.
8. Grcar, M., Cherepnalkoski, D., Mozetic, I., & Novak, P. K. (2017). Stance and influence of Twitter users regarding the Brexit referendum. Computational social networks, 4(1), 6.
9. Hyndman, R.J. & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 26(3).
10. Kendall, M. (1953). The Analysis of Economic Time Series, Part I: Prices. Journal of the Royal Statistical Society, 96.
11. Lazer,D., Kennedy, R., King, G. & Vespignani, A. (2014) *The Parable of Google Flu: Traps in Big Data Analysis*. Science (vol 343).
12. Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. The Journal of Business.
13. Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*.
14. Lee, T.H., White, H., Granger, C.W.J. (1993) Testing for neglected nonlinearity in time series models, Journal of Econometrics 56, 269–290.
15. Nason, G. (2013). A test for secondorder stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. J. R. Stat. Soc. B, 75: 879-904.
16. Teraesvirta, T., Lin, C.F., Granger, C.W.J. (1993) Power of the Neural Network Linearity Test, Journal of Time Series Analysis 14, 209–220.

# Random Forest-controlled Sparsity of High-Dimensional Vector Autoregressive Models

Dmitry Pavlyuk[1][0000-0003-3710-9678]

[1] Transport and Telecommunication Institute, Riga, Latvia
Dmitry.Pavlyuk@tsi.lv

**Abstract.** This paper introduces a new random forest-based approach to control the sparsity of vector autoregressive (VAR) models. VAR models are the popular forecasting tool in many applied areas, but high dimensionality of data is widely acknowledged as a significant hurdle for their efficient application. Sparse specifications of VAR models are designed to overcome this problem and keep the model specification parsimonious and interpretable. We consider the control of VAR models' sparsity as a special case of feature filtering problem and propose application of the random forest learning for its solution. The proposed approach includes preliminary ranking of features (model autoregressive and cross-sectional lags) by the random forest algorithm and further application of obtained feature importance values for a sparse specification of the VAR model. We tested the proposed approach against other specifications of VAR models' sparsity (refined VAR, penalized VAR) using the real-world urban traffic data set and demonstrated its statistical advantages: higher and more stable forecasting accuracy, manageable level of sparsity, and good computational performance for extremely high-dimensional time series.

**Keywords:** urban traffic forecasting, spatiotemporal model, feature selection, high dimension, big data

## 1 Introduction

In the era of big data, high-dimensional multivariate time series becomes a usual data structure in many applied areas such as intelligent transportation, distributed energy production, financial forecasting, and telecommunications. Data from spatially distributed providers (sensors, power stations, access points) form highly interrelated implicit structures, which should be identified and modelled for efficient forecasting. Historically these data sets were analyzed as independent univariate time series, but recent advances of time series forecasting methodology and growth of the computational power allow learning temporally lagged cross relationships between time series from data (in the context of spatially distributed data providers, structures of these relationships are called spatiotemporal). In addition to a high dimensionality of cross relationships in real-world applications, they are frequently observed as highly dynamic – appear only in specific conditions and exist for a short period of time. Effi-

cient leaning of spatiotemporal relationships in such environments is an emerging problem of multivariate time series forecasting.

In this study we contribute to the methodology of spatiotemporal structure learning by application of the random forest technique for controlling active relationships in multivariate time series in real time. The methodology represents a special case of feature filtering and applied for popular sparse vector autoregressive (VAR) models.

The proposed methodology is applied for the problem of spatiotemporal urban traffic forecasting, which is emerging in the transportation research area[1]. Data sets of a traffic management system include information from thousands of sensors (inductive loop, cameras, etc.), deployed within a city-wide road network, and perfectly demonstrate the problem of spatiotemporal structure learning for high-dimensional multivariate time series. Spatiotemporal relationships in traffic flows caused both by physical reasons (movement of cars from one spatial location to another) and by latent reasons (simultaneous traffic flows to a city center during morning rush hours). A structure of these relationships is highly dynamic and depends on traffic conditions, e.g. some relationships appear in a congested regime for a short period of time. Using the large real-world data set, we demonstrate the utility of the proposed methodology for spatiotemporal traffic forecasting.

## 2 Sparse Vector Autoregressive Models: State of the Art

Since the seminal paper of Sims[2], VAR models have become a popular tool of multivariate time series forecasting. Nowadays VAR models are intensively applied not only in macroeconomics, but also in many areas like health research, video stream control, traffic engineering, among many others. In many applications multivariate time series are high-dimensional and includes data for hundreds or thousands of indicators. However, performance of parameter-rich VAR models is degraded for high-dimensional data (the famous “curse of dimensionality” problem), which makes direct application of complete VAR models impractical. Several methodologies are suggested in literature to deal with the curse of dimensionality. Following the terminology of feature engineering, we divide all methodologies into two classes: feature extraction and feature selection methods.

Feature extraction corresponds to reduction of the dimensionality by transforming of the high-dimensional data set into a derivative feature set of a smaller dimension. Dynamic factor models[3], which combine time series into linear factors, are the popular representatives of this class of methods.

Feature selection techniques reduce number of VAR model’s parameters by setting restrictions on many model coefficients. Such limited specifications of VAR models are called sparse VAR. Sparse VAR models are in focus of this research.

Existing feature selection techniques are conventionally subdivided[4] to filter methods, wrapper methods, and embedded methods. Filter methods use preliminary feature ranking for selecting most valuable features. In the context of VAR models, Davis et al. [5] applied a partial spectral coherence, based on conditional correlation, for feature selection in their two-step sparse model specification procedure. Other

correlation-based VAR feature filtering approaches are proposed by Yang et al. [6, 7], Tanizawa et al.[8] and Yuen et al.[9].

Popular Bayesian VAR models also shrink complete VAR models towards a parsimonious specification by applying informative prior distributions of model parameters. Among several recent studies on Bayesian VAR[10, 11], Billio et al.[12] suggested Bayesian nonparametric prior distributions for VAR that combines clustering and shrinking restrictions.

The second class of feature selection techniques, wrapper methods, utilize information about VAR model performance in their iterative procedure of parsimonious specification search. Popular search strategies include stepwise-elimination of regressors and application of heuristic routines (genetic algorithms, particle swarm optimization). Classical wrapper strategies to model reduction are presented by Brüggemann[13]. Despite a good theoretical background and several promising evidences of wrapper technique application (e.g. PcGets algorithm and software[14]), this approach to sparse VAR model specification is related to significant computational complexity and rarely used for high-dimensional time series.

The third class of feature selection techniques, embedded methods, incorporate feature selection into model estimation process. Most popular embedded methods utilize different types of regularization penalties in VAR model estimators: $L_1$ (least absolute shrinkage and selection operator, LASSO) or elastic net (combination of $L_1$ and $L_2$ (Tikhonov) penalties). Penalties could be applied for all VAR parameters independently or groups by lag (to force sparsity in the temporal dimension) or by time series (for force sparsity in the indicator interrelationship structure). Regularization of high-dimensional VAR models is an emerging topic in literature: recently it was addressed by Basu and Michailidis[15], Barigozzi and Brownlees[16], and Nicholson et al.[17].

In addition to different classes of feature selection methods, discussed above, it should be mentioned that two general strategies are available: system strategy and single equation (equation-wise) strategy[13]. The system strategy implements feature selection jointly for all VAR equations, while the single equation strategy deals with each equation independently. As VAR models are the special case of seemingly unrelated regressions and deleting features from one equation affects the estimates of others, the system strategy is more natural. At the same time, in case of absence of instantaneous causality, single equation strategies also lead to optimal results[13] and could demonstrate better computational performance.

In this study we propose to apply a random forest as a feature selection tool for controlling the sparsity of VAR models. The random forest[18] is a popular statistical learning approach, widely used for feature selection and forecasting[19]. Advantages of random forests include: ability to learn under extremely large number of candidate features; low computational complexity and easy parallelization of the learning algorithm; embedded estimation of feature importance; resistance to overfitting and data preprocessing problems (scaling, outliers, missing data). To the best of our knowledge, random forests are not previously applied to learning the sparsity structure of VAR models. Recently random forests were applied by Furqan and Siyal[20], Papagiannopoulou et al.[21], and Chikahara and Fujino[22] for efficient and stable

learning of Granger causalities in multivariate time series, but without further application of discovered relationships. Tyralis and Papacharalampous[23] applied the random forest for feature selection in univariate autoregressive moving average models and demonstrate its preferable forecasting performance.

We apply the proposed random forest-controlled sparse VAR models to a spatio-temporal urban traffic forecasting problem and demonstrate its good computational complexity and forecasting performance. Thus, the study contributes both to the methodology of high-dimensional time series modelling and to the applied area of traffic forecasting.

## 3 Methods and Data

This section introduces notation and briefly summarizes vector autoregressive models and several feature selection techniques.

### 3.1 Methods

A multivariate time series in discrete time is defined as a sequence of $T$ observations of $k$-dimensional vector $Y_t = \left(y_{1,t}, y_{2,t}, \dots, y_{k,t}\right)', \ t = 1, \dots, T$. The complete (unrestricted) vector autoregressive model of order $p$, VAR($p$), is conventionally written as:

$$Y_t = \mu + \sum_{l=1}^{p} \Phi^{(l)} Y_{t-l} + \varepsilon_t, \tag{1}$$

where $\Phi^{(l)} = \left\{\phi_{i,j}^{(l)}\right\}$ are $k \times k$ coefficient matrixes ($l = 1, \dots, p; \ i, j = 1, \dots, k$), $\mu = \{\mu_i\}$ is a optional $k \times 1$ vector of constant terms, $\varepsilon_t = \left\{\varepsilon_{i,t}\right\}$ is a $k \times 1$ vector of unobservable zero mean disturbances with non-singular covariance matrix $\sum_{\varepsilon}$.

Sparsity of VAR(p) models corresponds to setting elements of coefficient matrixes $\Phi^{(l)}$ to zero to reduce number of model parameters. In this study we consider the filter approach to controlling the model sparsity, which is based on selection of non-zero coefficient before model estimation. Thus, we formulate the sparse VAR(p) model introducing a set of binary matrixes $\mathrm{S}^{(l)} = \left\{s_{i,j}^{(l)}\right\}$ that represent relationships in VAR(p):

$$Y_t = \mu + \sum_{l=1}^{p} \mathrm{S}^{(l)} \Phi^{(l)} Y_{t-l} + \varepsilon_t. \tag{2}$$

We will refer $S = \left[\mathrm{S}^{(1)}, \mathrm{S}^{(2)}, \dots, \mathrm{S}^{(p)}\right]$ and $\Phi = \left[\Phi^{(1)}, \Phi^{(2)}, \dots, \Phi^{(p)}\right]$.

VAR(p) model can be fit by the ordinary least squares (OLS) estimator, but number of estimated parameters equals to $(pk^2 + k)$ and becomes extremely large for high-dimensional time series. Regularization is a usual way to overcome the curse of dimensionality, which introduces a penalty function $\mathcal{P}(\Phi)$ into the estimator objective function:

$$\underset{\Phi,\mu}{\operatorname{argmin}}\left(\left\|Y_t-\mu-\sum_{l=1}^{p}\mathrm{S}^{(l)}\Phi^{(l)}Y_{t-l}\right\|_F+\lambda\mathcal{P}(\Phi)\right) \quad (3)$$

where $\|A\|_F$ denotes the Frobenius norm of a matrix $A$, $\lambda$ is a regularization hyperparameter ($\lambda = 0$ for non-regularized OLS). In this study we utilize popular LASSO penalty function $\mathcal{P}(\Phi) = \|\Phi\|_1$, where that $\|\Phi\|_1$ is the $L_1$-norm of the coefficient matrix. Regularization substantially decreases number of VAR model parameters and could improve model forecasting performance.

As an alternative approach, we propose to use random forests for feature filtering. We limit the research scope to the single equation strategy, so random forests will be independently applied for each equation of the VAR model and obtained results will be used for a final specification of $\mathrm{S}^{(l)}$.Within the single equation strategy, the problem of sparsity definition is reduced to feature selection in *k* equations:

$$y_{i,t}=\mu_i+\sum_{l=1}^{p}\sum_{j=1}^{k}s_{i,j}^{l}\phi_{i,j}^{l}y_{j,t-l}+\varepsilon_{i,t}, \quad (4)$$

where $s_{i,j}^{l}=1$ corresponds to used features.

Random forest is a popular machine learning technique, proposed by Breiman in 2001[18]. This technique is widely used for feature selection and includes following steps[19]:

1. Sample with replacement of *n* training sets $\{Y_{ts}\}, ts \subseteq \{1,\dots,T\}$
2. Training of a regression tree for equation (4) for every training set, randomly selecting features for every tree node
3. Estimating of importance of each feature in every regression tree
4. Combining obtained feature importance values (e.g. by averaging over training sets)

We apply increase of mean squared error (MSE) as a metric of feature importance at step 4. The metric is calculated as a difference between out-of-sample MSE estimate for the original training set and the same indicator for a training set with randomly permutated values of a specific feature. This difference (decrease of importance) is averaged over all training sets.

The next step is to utilize feature importance values to choose most important features. We consider two approaches for this step: combine feature importance values, obtained by random forests for all equations, or implement feature selection for every equation independently.

The resulting feature set is used for sparse VAR model specification and estimation. Summarizing the methodologies, stated above, we formulate 4 alternative model specifications:

- Unrestricted VAR model.
- Refined VAR (a model with excluded insignificant coefficients by backward elimination procedure, the system strategy) – a representative of wrapper feature selection.
- Penalized VAR (LASSO penalties; the system strategy) – a representative of embedded feature selection.

- Random forest-controlled VAR (single-equation strategy) – the proposed representative of feature filtering.

The primary research question lays in a comparative forecasting performance of candidate models. We applied the rolling analysis [24] with a constant window size (look-back interval) for tuning of hyperparameters and estimation of models' out-of-sample forecasting accuracy. Parameters of every model specification were tuned independently:

- Complete VAR model: look-back interval; maximal lag $p$.
- Refined VAR model: look-back interval; maximal lag $p$.
- Penalized VAR: look-back interval; regularization parameter λ; maximal lag $p$.
- Random forest-controlled (RF-controlled) VAR with equation-wise feature filtering: look-back interval; maximal lag $p$; selection: sparsity (percent of non-zero coefficients).

The out-of-sample mean absolute error (MAE), averaged by time series, is used as the primary forecasting accuracy metric:

$$MAE_t = \frac{1}{k}\sum_{i=1}^{k}\left|y_{i,t} - \hat{y}_{i,t}\right|, \quad (5)$$

where $\hat{y}_{i,t}$ is a predicted value for a spatial location *i* and time point *t*.

### 3.2 Data: Urban Traffic Forecasting

We applied the proposed methodology to a multivariate time series of urban traffic volume values, obtained from 103 stations on arterial roads at Minneapolis, USA. All stations are located within 6 minutes of travel time in uncongested traffic conditions from a city center. We collect data for 40 weeks (01 Jan 2017 – 07 Oct 2017) and temporally aggregated them in 1-minute time frames. First 30 weeks of data were used for learning of historical patterns and the last 10 weeks – for model validation. Historical patterns are learned independently for every univariate time series as median values, specific to a day of the week and time of the day. We also tested exponential smoothing model for learning and excluding trends and periodic fluctuations, but the resulting patterns did not significantly differ from simple median values (due to absence of significant trends and changes in periodic components in data). Data for 10 weeks, designed for model validation was detrended by subtracting historical patterns (see Fig. 1 for a sample plot of original and detrended time series for one day).

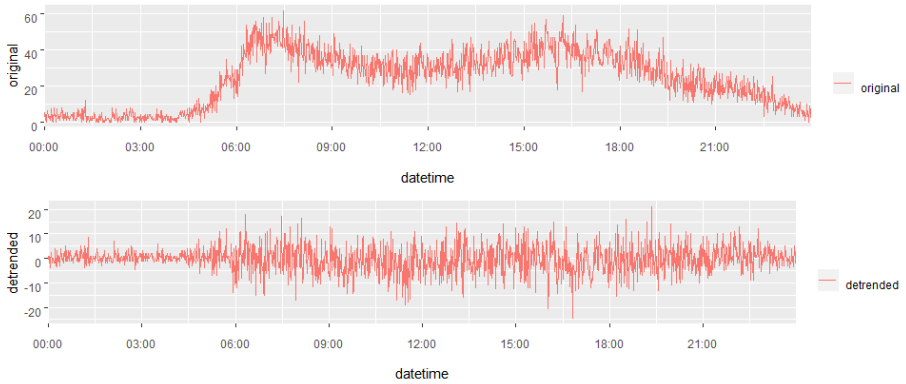



**Fig. 1.** Typical original and detrended traffic volume values

In addition, we implemented standard data preprocessing procedures: removal of outliers (based on physical capacity of roads); imputing missed data (by linear interpolation), and winsorization (by lower and upper bounds, identified by the inter quartile range technique of outlier detection).

## 4 Results

Dimensionality of modelled time series is a key input for sparse model specification. We tested all candidate models for two datasets:

- Random sample of 10 stations (k=10) that referred as the low-dimensional data set.
- Complete set of 103 sensors (k=103) that referred as the high-dimensional data set.

We assume that dimensionality of the first multivariate time series is small enough to keep stable estimates of the complete VAR model, while for the second data set sparse VAR specification will be beneficial.

Hyperparameter tuning was executed by a grid search, where every combination of hyperparameter values is tested by rolling window cross-validation. The rolling window was shifted over 69 days (10 weeks minus the first day for look-back interval) every 4 hours, which resulted to 414 model estimates per hyperparameter combination. The resulting hyperparameters values were selected as:

- Look-back interval is 16 hours for all models ($T$=960 minutes). The value is selected as a minimal length of time series that ensure stabilized results on model forecasting performance. The 16-hour interval was required for the complete VAR model, while sparse model specifications required approximately twice less data (8 hours) for stable estimation.
- Optimal order of VAR models is 3 ($p$ = 3), which was expected due to a limited spatial area of analysis (maximum travel time between sensors is 6 minutes in normal traffic conditions).

- Regularization parameter λ for the penalized VAR model was selected in a flexible manner for every cross-validation subsample. The time-specific optimal value is obtained by splitting the data set into two equal part and using of the second part for local cross-validation of the regularization parameter λ [17]. In addition to the basic penalty function, we tested its own-other variant, that deals with own autoregressive and cross time series lags differently. Despite our expectations, the own-other penalty function doesn't demonstrate significant improvements of forecasting performance, so we excluded the related results from this paper.
- Sparsity (percent of non-zero elements) for the RF-controlled VAR model is selected as 30% for the low- and 20% for the high-dimensional data set (that corresponds to 3 non-zero coefficients for the former and 20 non-zero coefficients for the latter data set per every lag in every VAR equation). We applied an increase in MSE as a metric of feature importance and dealt with every VAR equation independently. More precisely, we excluded features with negative values of increase in MSE and after that selected most important features according to the specified sparsity of the resulting VAR model.

The penalized VAR and RF-controlled sparse VAR model specifications allows tuning of parameters for a specific forecasting horizon. In this study we arbitrary trained both models to optimize the one-step ahead forecasting performance.

Resulting forecasting performance of the candidate models with optimally selected hyperparameters is presented in Table 1.

**Table 1.** One-step ahead forecasting performance of the candidate models

| | Low- dimensional data set, k=10 | | High-dimensional data set, k=103 | | |
|---|---|---|---|---|---|
| Model | Average MAE | 95$^{th}$ percentile of MAE | Average MAE | 95$^{th}$ percentile of MAE | Complexity*, seconds |
| Complete VAR | 4.64 | 6.87 | 4.53 | 8.47 | 0 |
| Penalized VAR | 4.69 | 6.89 | 4.14 | 6.67 | 682 |
| Refined VAR | 4.62 | 6.78 | 3.99 | 7.31 | 1169 |
| RF-controlled sparse VAR | 4.61 | 6.67 | 4.06 | 6.57 | 871 |

* Computation complexity, average seconds per model for feature selection

The obtained one-step ahead forecast average MAE values are almost identical for all candidate model for the low-dimensional data set, while differ significantly for the high-dimensional one. In addition to average MAE values, we provide their 95$^{th}$ percentiles to explore spatial and temporal stability of obtained forecasts. Discussion on the presented results is provided in the next paper section.

Another point of our interest is the stability of model forecasting performance for longer forecasting horizons. We constructed *h*-step ahead forecasts for all models (*h* is a forecasting horizon, *h*=1, ..., 12) using the iterative one-step ahead strategy (so forecasts for the next step were calculated using values, forecasted for the previous steps)

and combined them into aggregated forecasts for longer intervals (from 1 to *h* minutes). Further average MAE values were calculated for aggregated forecasts. A comparison of obtained results is presented on Fig. 2 (for the low-dimensional data set) and Fig. 3 (for the high-dimensional data set).

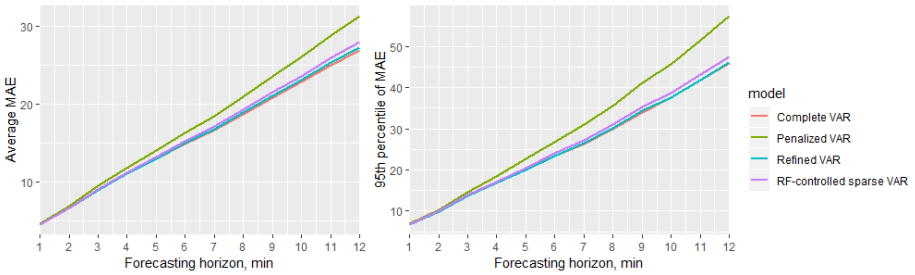



**Fig. 2.** Accuracy of the candidate models (mean and 95th percentile of MAE values) by forecasting horizon: low-dimensional data set

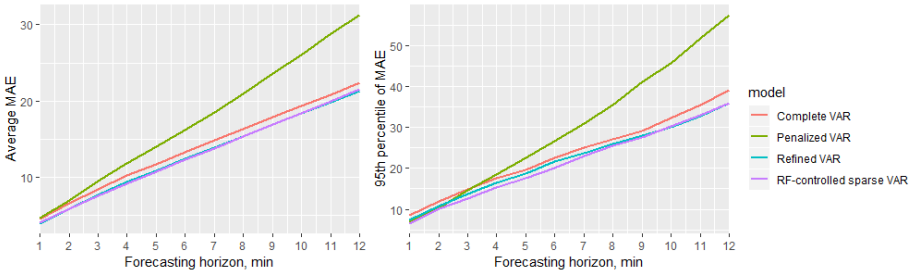



**Fig. 3.** Accuracy of the candidate models (mean and 95th percentile of MAE values) by forecasting horizon: high-dimensional data set

Note that *h*-step ahead forecasting accuracy is almost identical for all models, except the penalized VAR, for low-dimensional data, but differ significantly for high-dimensional ones.

Finally, to support reproducibility of the obtained results, we provide open source codes (R language) for all described routines in the public repository: http://bit.ly/ITISE2019.

## 5 Discussion

The primary research interest is comparison of forecasting performance of the proposed RF-controlled VAR model specification against other alternatives. For the low-dimensional data set, RF-controlled VAR model's forecasting performance is almost identical to the complete VAR model (average MAE is 4.61 against 4.64 for one-step ahead forecasts, Table 1). While the forecasting performance of models is similar, the parsimonious specification of the RF-controlled VAR model could be considered as

an advantage, because it provides easier understanding of existing relationships and leads to more interpretable and manageable results. These results become more important under the observed stability of RF-controlled VAR model's forecasting performance for longer time intervals (Fig. 2). Performance of RF-controlled VAR model is degrading almost with the same speed as of the complete VAR model, while average MAE of the competitor penalized VAR is growing much faster (note that both RF-controlled VAR and penalized VAR are trained to optimize one-step ahead forecasts, so no model has a prespecified advantage). Thus, we conclude the good performance of the RF-controlled VAR model for low-dimensional data sets, where the unrestricted VAR is widely considered as a primary model specification.

For larger dimensionality the proposed RF-controlled VAR model demonstrates a clear advantage in forecasting performance over complete and penalized VAR models. Its average one-step ahead forecast MAE value is 4.06 against 4.53 of the complete VAR model (Table 1), which is a statistically significant difference for the utilized number of cross-validation subsamples. This advantage keeps stable over longer forecasting intervals (Fig. 3), while forecasting accuracies of the complete and penalized VAR are degraded faster. Forecasting performance of the RF-controlled VAR model is similar to the refined VAR, but its computational complexity is much lower (Table 1 contains averages computation times for feature selection on the identical environment). Another comparative advantage of the RF-controlled VAR model against all competitor specifications is demonstrated by the $95^{th}$ percentile values of MAE – 6.57 for the RF-controlled VAR model against 8.47 for the complete VAR (Table 1). This fact is considered as an evidence of improvement of the forecasting performance stability over space and time and supports our hypothesis about the general advantage of the proposed approach.

In addition to the primary research interest, we should mention several observations from the obtained results: 1) forecasting performance of the refined VAR model overcomes the complete VAR specification both for low-dimensional and high-dimensional data sets, but requires intensive computations; 2) the penalized VAR model's performance strictly depends on the prespecified target forecasting horizon (one-step ahead in our experiments), so this model should be separately tuned for every forecasting horizon; 3) computational complexity of the proposed RF-controlled VAR model is growing fast with increasing dimensionality of the time series, but keeps manageable (at least for several hundreds of dimensions) and the related algorithm is easily parallelized.

## 6 Conclusions

In this paper we propose a new random forest-based approach to sparse specification of vector autoregressive models. Within the proposed approach, we utilize the random forest for equation-wise feature selection and further apply most important features for sparse VAR model specification.

The proposed approach was applied to the real-world urban traffic data set and tested against alternative model specifications: unrestricted VAR; refined VAR with

excluded insignificant coefficients; and LASSO-penalized VAR. Obtained experimental results demonstrated the advantage of the proposed RF-controlled sparse VAR model in several aspects. First, forecasting performance of the RF-controlled sparse VAR model overcomes the performance of analyzed competitive models for high-dimensional data. Second, a parsimonious specification of the RF-controlled sparse VAR is also appropriate for low-dimensional data, which is an advantage in terms of model interpretability. Third, the proposed approach inherits advantages of random forests such as ability to learn under extremely large number of candidate features; low computational complexity, easy parallelization; resistance to overfitting, which makes it appropriate for high-dimensional modelling of big data.

Finally, we should mention a wide area for the future research in this direction. We applied the equation-wise strategy of feature filtering for VAR model specification, while the system-wide strategy, where all equations are analyzed simultaneously, is expected to be more efficient. To implement this strategy, methodological advances of multi-output random forests and their application for multivariate time series are highly required.

## Acknowledgement

The author was financially supported by the specific support objective activity 1.1.1.2. “Post-doctoral Research Aid” (Project id. N. 1.1.1.2/16/I/001) of the Republic of Latvia, funded by the European Regional Development Fund. Dmitry Pavlyuk’s research project No. 1.1.1.2/VIAA/1/16/112 “Spatiotemporal urban traffic modelling using big data”.

## References

1. Ermagun, A., Levinson, D.: Spatiotemporal traffic forecasting: review and proposed directions. Transport Reviews. 1–29 (2018). https://doi.org/10.1080/01441647.2018.1442887.
2. Sims, C.A.: Macroeconomics and Reality. Econometrica. 48, 1 (1980). https://doi.org/10.2307/1912017.
3. Forni, M., Lippi, M.: The Generalized Dynamic Factor Model: Representation Theory. Econometric Theory. 17, 1113–1141 (2001).
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering. 40, 16–28 (2014). https://doi.org/10.1016/j.compeleceng.2013.11.024.
5. Davis, R.A., Zang, P., Zheng, T.: Sparse Vector Autoregressive Modeling. Journal of Computational and Graphical Statistics. 25, 1077–1096 (2016). https://doi.org/10.1080/10618600.2015.1092978.
6. Yang, K., Yoon, H., Shahabi, C.: CLe Ver: A Feature Subset Selection Technique for Multivariate Time Series. In: Ho, T.B., Cheung, D., and Liu, H. (eds.) Advances in Knowledge Discovery and Data Mining. pp. 516–522. Springer Berlin Heidelberg, Berlin, Heidelberg (2005). https://doi.org/10.1007/11430919_60.
7. Yang, K., Yoon, H., Shahabi, C.: A Supervised Feature Subset Selection Technique for Multivariate Time Series. 10 (2005).

8. Tanizawa, T., Nakamura, T., Taya, F., Small, M.: Constructing directed networks from multivariate time series using linear modelling technique. Physica A: Statistical Mechanics and its Applications. 512, 437–455 (2018). https://doi.org/10.1016/j.physa.2018.08.137.
9. Yuen, T.P., Wong, H., Yiu, K.F.C.: On constrained estimation of graphical time series models. Computational Statistics & Data Analysis. 124, 27–52 (2018). https://doi.org/10.1016/j.csda.2018.01.019.
10. Carriero, A., Clark, T.E., Marcellino, M.: Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. Journal of Econometrics. S030440761930079X (2019). https://doi.org/10.1016/j.jeconom.2019.04.024.
11. Koop, G., Korobilis, D., Pettenuzzo, D.: Bayesian compressed vector autoregressions. Journal of Econometrics. 210, 135–154 (2019). https://doi.org/10.1016/j.jeconom.2018.11.009.
12. Billio, M., Casarin, R., Rossini, L.: Bayesian nonparametric sparse VAR models. Journal of Econometrics. S0304407619300776 (2019). https://doi.org/10.1016/j.jeconom.2019.04.022.
13. Brüggemann, R.: Model Reduction Methods for Vector Autoregressive Processes. Springer Berlin Heidelberg, Berlin, Heidelberg (2004).
14. Hendry, D., Krolzig, H.-M.: Automatic Econometric Model Selection Using PcGets. Timberlake Consultants Press, London (2001).
15. Basu, S., Michailidis, G.: Regularized estimation in sparse high-dimensional time series models. Ann. Statist. 43, 1535–1567 (2015). https://doi.org/10.1214/15-AOS1315.
16. Barigozzi, M., Brownlees, C.: NETS: Network estimation for time series. J Appl Econ. 34, 347–364 (2019). https://doi.org/10.1002/jae.2676.
17. Nicholson, W.B., Matteson, D.S., Bien, J.: VARX-L: Structured regularization for large vector autoregressions with exogenous variables. International Journal of Forecasting. 33, 627–651 (2017). https://doi.org/10.1016/j.ijforecast.2017.01.003.
18. Breiman, L.: Random Forests. Machine Learning. 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York, New York, NY (2009). https://doi.org/10.1007/978-0-387-84858-7.
20. Furqan, M.S., Siyal, M.Y.: Random forest Granger causality for detection of effective brain connectivity using high-dimensional data. J. Integr. Neurosci. 15, 55–66 (2016). https://doi.org/10.1142/S0219635216500035.
21. Papagiannopoulou, C., Miralles, D.G., Decubber, S., Demuzere, M., Verhoest, N.E.C., Dorigo, W.A., Waegeman, W.: A non-linear Granger-causality framework to investigate climate–vegetation dynamics. Geosci. Model Dev. 10, 1945–1960 (2017). https://doi.org/10.5194/gmd-10-1945-2017.
22. Chikahara, Y., Fujino, A.: Causal Inference in Time Series via Supervised Learning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. pp. 2042–2048. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden (2018). https://doi.org/10.24963/ijcai.2018/282.
23. Tyralis, H., Papacharalampous, G.: Variable Selection in Time Series Forecasting Using Random Forests. Algorithms. 10, 114 (2017). https://doi.org/10.3390/a10040114.
24. Zivot, E., Wang, J.: Rolling Analysis of Time Series. In: Modeling Financial Time Series with S-PLUS. pp. 313–360. Springer New York, New York, NY (2006). https://doi.org/10.1007/978-0-387-32348-0_9.

# Unsupervised Anomaly Detection in Time Series with Convolutional-VAE

Emanuele La Malfa[1,a)] and Gabriele La Malfa[2,b)]

[1]*Computer Science Engineer*
[2]*Student EMLYON Business School*

[a)]Corresponding author: emanuele.la.malfa@outlook.it
[b)]Corresponding author: gabriele.lamalfa@edu.emlyon.com

**Abstract.** We propose an unsupervised machine learning algorithm for anomaly detection that exploits self-learnt features of mono-dimensional time series. A Variational Autoencoder, where convolution takes place of dot product, is trained to compress each input to a low-dimensional point from a normal distribution, detecting an anomaly as low probability and high density sequence. We validate our work on different public datasets, obtaining results that shed new light on Variational Autoencoders applied to anomaly detection.

## I. INTRODUCTION

Forecasting and anomaly detection represent two critical research topics in the analysis of continuous phenomena over time. These studies are applied to different disciplines: physics, healthcare, robotics, artificial intelligence, finance, product analysis, etc. Each discipline employs its own corpus of knowledge [1, 2], build on the top of many factors: from the nature of input variables and their relationships, to the presence of trends, seasonality in the time series etc. [3, 4]. Often physical processes exhibit patterns that can be modeled with simple functions that repeat themselves around their fundamental period, while for many others, like financial and economical series, random walks models [5] are employed and their predictability is still an open issue [6]. A fundamental factor, in terms of predictability of the process, is how information propagates through time. If from one side in Markovian processes the information useful to forecast the next time period depends only on the previous state of the system, on the other in chaotic system a small perturbation of the initial state may influence the behavior or the phenomenon in a remote future.

Sequential models applied to time series have been widely used in recent years in many disciplines [8, 9, 13]. Particular attention has been devoted to explore learning methods [14, 15], enabled by the capacity of those models to process unidimensional and multidimensional datasets, extract features autonomously [16] and model complex systems' dynamics. [17]. Malhotra et al. [20] use Long Short-Term Memory Networks (LSTM) [21] on physical time series: once forecasting is reliable, anomaly detection is based on modeling the prediction errors. On the other hands, Tsang et al. [22] provide a learning method applied to financial time series: after preprocessing data with a Symlet Wavelet Thresholding, a Stacked Autoencoder (SAE) is used for a pre-train session and finally an LSTM is used to forecast and identify anomalies. When it comes to financial time series forecasting, is often necessary a multivariate dataset which provides the missing information about the stochastic process that is not present in the mono variate market index. Laptev et al. [18] feed the features extracted from an autoencoder to a LSTM model, hence the model is used for anomaly detection. Being able to extract only the relevant features for the process may also benefit extreme event forecasting [19].

In this work, we use a Variational Autoencoder (VAE) [24], where dot product is replaced by convolution: this operation has been used extensively [8, 9, 10] for signal processing, hence it can enhance VAE so the model learns relevant features by compressing each input sequence to a point drawn from a low-dimensional gaussian distribution, hence labeling as anomalous dense timesteps (i.e. they are close each other) whose probability is low. To the best of our knowledge Convolutional VAE has been applied only recently to clustering problems [11, 12] while anomaly

detection applied to physical time series is an original contribute of this work.
The paper is organized as follows: in Section II the Convolutional Variational Autoencoder (CVAE) model is described in details. In Section III the we present the results on several physical datasets, while the last Section is dedicated to the conclusion and future directions of this work.

## II. METHOD

We approach the problem of anomaly detection in mono dimensional time series with a Variational Autoencoder [24] where the dot product, that involves the affine transformation between each stage input and the neural network's parameters, is replaced by convolution. The main idea behind this choice is that convolution is the state of the art method in many challenges where signals are involved [7], mainly due to ability of convolutional networks to build on top of the self-extracted features increasingly complex representations of the input that are used for tasks like classification, outliers detection etc. Differently from simple dot product, convolution is characterized by weights that are shared along the input, making it possible to spot patterns that are invariant to translations and rotations, i.e. robust to noise and perturbations.
As the Variational Autoencoder learns to map each training input to a point belonging to a low-dimensional gaussian distribution, the model emphasizes the local characteristics of each sequence. Figure 1. shows separately the key points of both Variational Autoencoders and the convolution operation. In the next section the math behind the model and the CVAE architecture are described in details, while the full source code for the models employed in this work is provided[1].

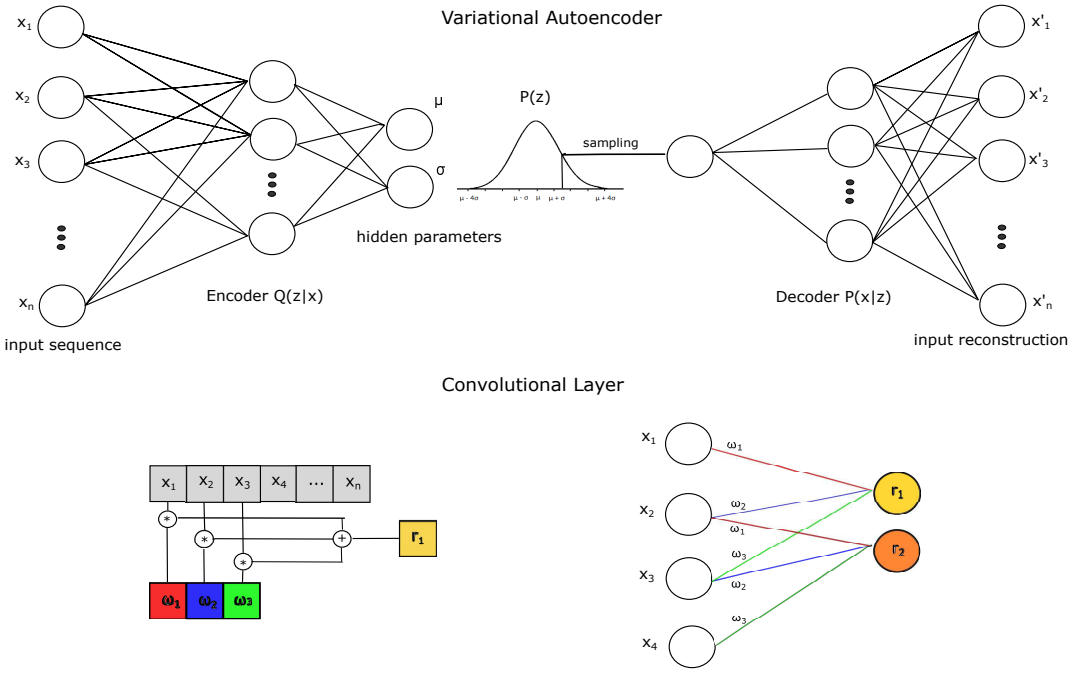



FIGURE 1: The top image shows a VAE architecture where the neural network's parameters are not shared (i.e. each layer is dense). The encoder learns a compressed representation of each input sequence by learning the hidden distribution's parameters (in our case, being it a standard distribution, a vector for the mean and a matrix for the standard deviation). From that representation, the decoder reconstructs the input through another step of affine transformations. The bottom image shows how convolution is employed in our work to replace the neural network's dense layers: in the convolutional layers the parameters are locally connected, hence the model is expected to learn a representation of the series which is invariant to translations and rotations.

[1] https://github.com/EmanueleLM/CVAE

## Convolutional Variational Autoencoder

A Variational Autoencoder is used to make inference and learning with a probabilistic based method characterized by latent variables with a posterior intractable distribution.
By defining latent variables $z$ that describe our data it is possible to obtain a generative model: formally, one can model the probability of the input data as $P(X) = \int P(X|z)P(z)dz$, where $P(z)$ is the probability distribution function of the latent variable $z$, also called *prior*, and $P(X|z)$ the conditional distribution of data.
In order to obtain $P(z)$, one can use the conditional probability $P(z|X)$: unfortunately approximating this distribution is often hard, hence variational inference approximates $P(z|X)$ with another tractable distribution $Q(z|X)$. This approximation problem can be optimized by a convolutional neural network where the first half layers, i.e. the *encoder*, map $X$ to the low-dimensional gaussian distribution $P(z)$ employing $Q(z|X)$, while the second part of the network rebuilds (hence the name *decoder*) the input by approximating $P(X|z)$ from its low-dimensional representation $z$.
In order to rebuild the original input sequence, the *deconvolution* operation is employed [26], while the hidden representation of each input sequence is obtained with the so-called *reparametrization trick* [27].
The network's parameters are learnt through backpropagation, by minimizing the Kullback-Leibler (KL) divergence between encoder and the intractable prior distributions, namely $D_{KL}[Q(z)||P(z|X)]$. The objective function, known as Evidence Lower Bound (ELBO), takes the following form: $logP(x) - D_{KL}[Q(z|X)||P(z|X)] = E[logP(X|z)] - D_{KL}[Q(z|X)||P(z)]$. In the last equation $E[logP(X|z)]$ measures the reconstruction error from the input sequence, while $D_{KL}[Q(z|X)||P(z)]$ accounts for the divergence between the encoder and the prior function.
In our work we train the CVAE on data $D_{train}$ that does not contain anomalies and test it on unseen data $D_{test}$ that instead may contain anomalies: in this way, when the compressed representation of an input sequence $z$ from $D_{test}$ is very different from the pool of patterns seen so far in the time series, it will be assigned a low probability and marked as anomalous.
Since minimizing the CVAE loss is computationally expensive even in the condition of defining a trainable encoder function, we explored several normalizations techniques: from l2 regularization, that is known to benefit CVAE [11], to l1 (that induces sparsity in the solution), to a combination of both l1 and l2 regularizations. Moreover, we have experienced that assigning importance weights to the different loss' terms (reconstruction error and KL divergence between decoder and prior) benefits the anomaly detection.

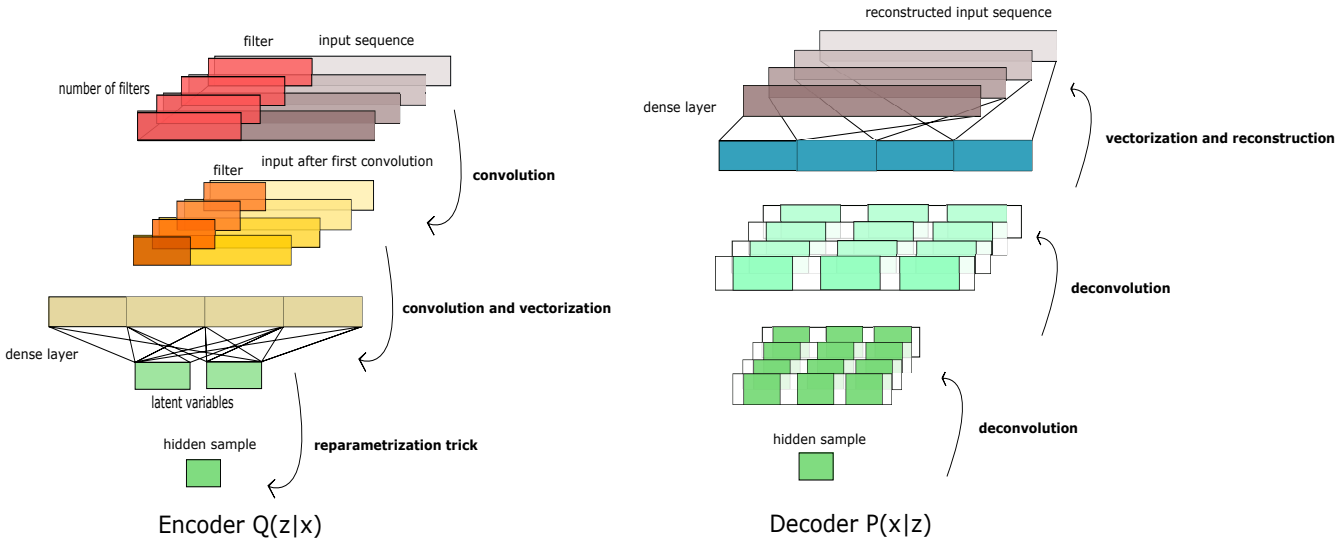



FIGURE 2: The CVAE model proposed in this work: at each stage of the encoder (left part of the image), convolution replaces dot product (fully connected layers), except for the layer before the *bottleneck*, i.e. where the *reparametrization trick* is used to sample from the latent distribution. At each stage of the decoder (right part of the image), deconvolution is used to reconstruct the input sequence, except for the last layer that is dense. Fully connected layers are necessary to make the input match the dimension of respectively its reconstruction and the latent variables distribution.

## III. Anomaly Detection

We introduce a method to detect anomalies based on both probability and density of the candidate sequences: given a series of input sequences $s = (x_1, .., x_n)$, $s$ is anomalous if the following two conditions on $s$ hold. First, the probability that CVAE assigns to each sequence is below a threshold $\tau$, namely $\forall\ x_i \in s\ P_M(x_i) \leq \tau$, for example the $5^{th}$ and $95^{th}$ percentiles, being $z$ drawn from a normal distribution. The density based condition requires that for each sequence in $s$, the temporal distance between each couple $(x_i, x_{i+1})$ is a fraction of the entire sequence length, namely $\forall\ x_i, x_{i+1} \in s\ d(x_i, x_{i+1}) \leq k|x|,\ s.t.\ k \in (0, 1)$, where $d(\cdot, \cdot)$ is a measure of the distance on the $x$-axis (the *time* axis) between two candidate sequences, while $|x|$ is the length of each input sequence.
We test our model on three publicly available physical[2] datasets, plus a synthetic one. All the datasets are arranged so the anomalies are not present in train/validation, while in the test part one or more anomalies need to be detected. As regards the synthetic dataset, it is a *sin* function (the model for each data point is $y = sin(t + \rho)$) where some flat zones are introduced in the test set, as it was thought to explicitly show how CVAE algorithm discovers and highlights anomalies (see Figure 3 for respectively the train set (a), the test set (b) and each sequence's likelihood (c)). Even if all the 4 datasets come in the form of univariate quasi predictable time-series, they set different challenges: data from Space Shuttle Marotta Valve (Figure 4) contain a localized anomaly which can be easily spotted when a sequence is long enough to capture the unseen pattern, while in the Power Consumption dataset (Figure 5) the anomaly can be spotted if the algorithm *captures* the 7 days periodicity, finally spotting in the test set that two out of five consumption's peaks are not present at the end of the sequence. Data are preprocessed with normalization methods and subsampled (up to a factor of 5) when possible, to speedup computation.
The results obtained with CVAE algorithm are reported in following images [3].

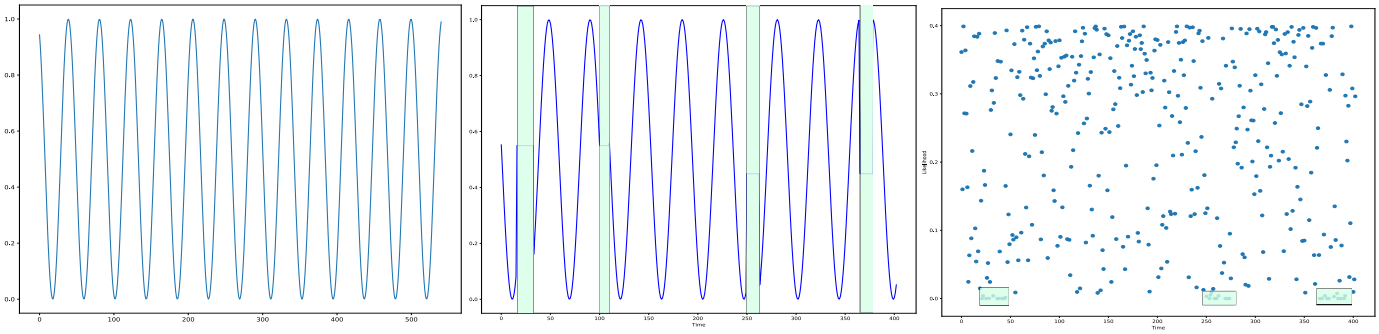



FIGURE 3: Anomaly detection on the synthetic dataset: it is a simple sin function where in the test set (figure (b)) some flat lines has been substituted to the origin function (the green vertical bars highlight each anomalous zone). The CVAE algorithm is able to spot the anomalies (green rectangles on figure (c), where on $y$-axis it is reported the probability of each timestep) assigning low probability to each sequence that has been artificially modified.

## IV. CONCLUSION

We have presented an unsupervised method for anomaly detection that exploits two different concepts: Variational Autoencoders and Convolutional Neural Networks. We have shown that anomalies are highlighted as high-density/low-probability points. We reserve to extend the analysis to other datasets: in future works the aim is to apply those methods also to multidimensional financial time series, to capture the highly complex relations between features and hidden variables.

[2] `http://www.cs.ucr.edu/~eamonn/discords`.

[3] Since the algorithm is fully unsupervised, one may obtain the parameters $\tau$ and $k$ as the parameters that enhance anomaly detection on one or more synthetic datasets. On the other hands, one may wish to find $\tau$, $k$ so they maximize the $F_\beta$ score between anomalous and non-anomalous sequences on validation, but this would make the problem partially supervised.

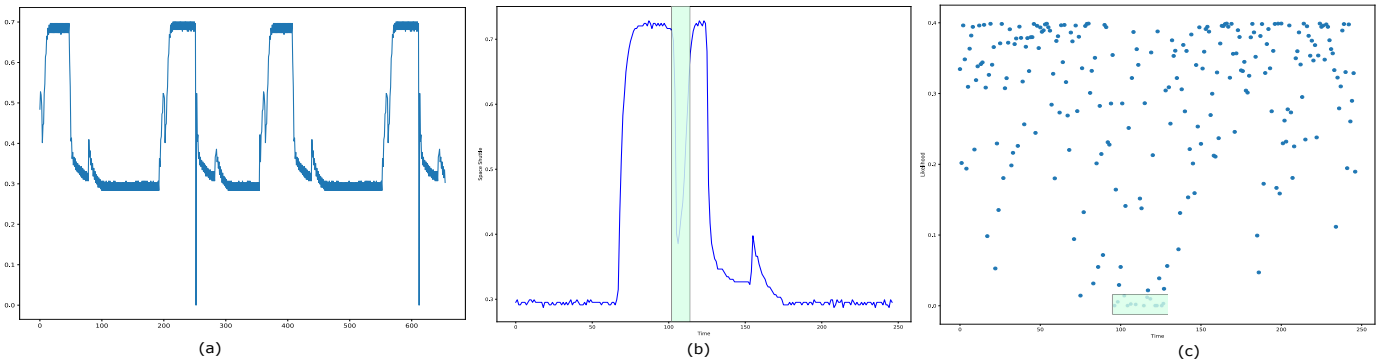



FIGURE 4: Anomaly detection on the space shuttle dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b)) there's a phase of decompression and compression that constitutes an anomaly that is detected by the CVAE algorithm with certainty (green rectangles on figure (c), where on *y*-axis it is indicated the probability of each timestep).

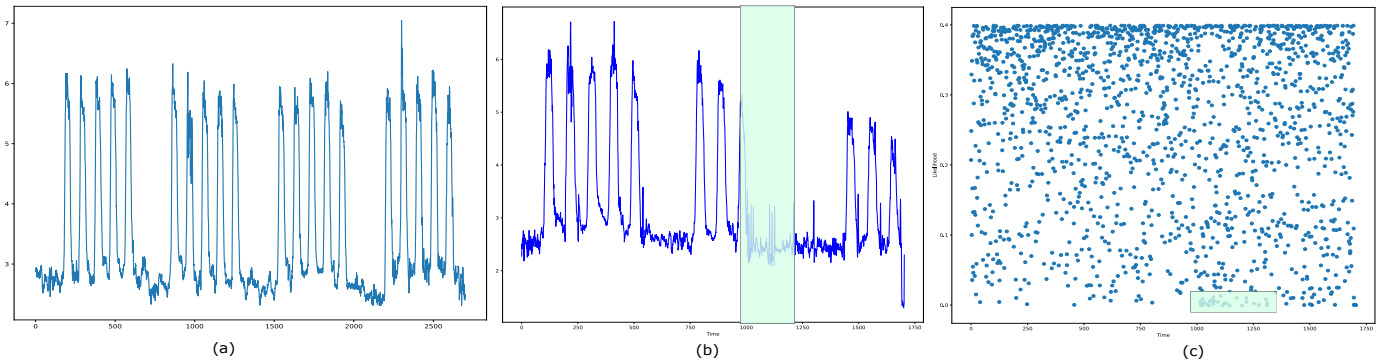



FIGURE 5: Anomaly detection on the power consumption dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b), where the green vertical bar highlight the anomalous zone) the anomalous sequence is identified by two out of five missing peaks of consumption. The CVAE algorithm spots it with a high density zone of low probability sequences (green rectangles on figure (c), where on *y*-axis it is indicated the probability of each timestep).

## REFERENCES


[1] V. Chandola, A. Banerjee, V. Kumar, *Anomaly detection: A survey*, ACM Computing Surveys (CSUR), Vol. 41, Issue 3, 2009.

[2] V. J. Hodge, J. Austin, *A Survey of Outlier Detection Methodologies*, Artificial Intelligence Review, Vol. 22, Issue 2, pp. 85126, 2004.

[3] R. H. Shumway, D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, Springer International Publishing, 2017.

[4] R. S. Tsay, *Analysis of Financial Time Series*, John Wiley & Sons, 3rd Edition, Hoboken, 2010.

[5] E. F. Fama, *Random Walks in Stock Market Prices*, Financial Analysts Journal, Vol. 21, no. 5, pp. 55-59, 1965.

[6] A. W. Lo, A. C. MacKinlay, *A Non-Random Walk Down Wall Street*, Princeton University Press, 2002.

[7] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, *A survey of deep neural network architectures and their applications*, Neurocomputing, 234, 11-26, 2017.

[8] L. Deng, *Three classes of deep learning architectures and their applications: a tutorial survey*, APSIPA Transactions on Signal and Information Processing, 2012.

[9] Y. LeCun, Y. Bengio, G. Hinton, *Deep learning*, Nature, Vol. 521, no. 7553, 2015.

[10] Y. LeCun, Y. Bengio, *Convolutional networks for images, speech, and time series*, The handbook of brain theory and neural networks 3361 (10), 1995.

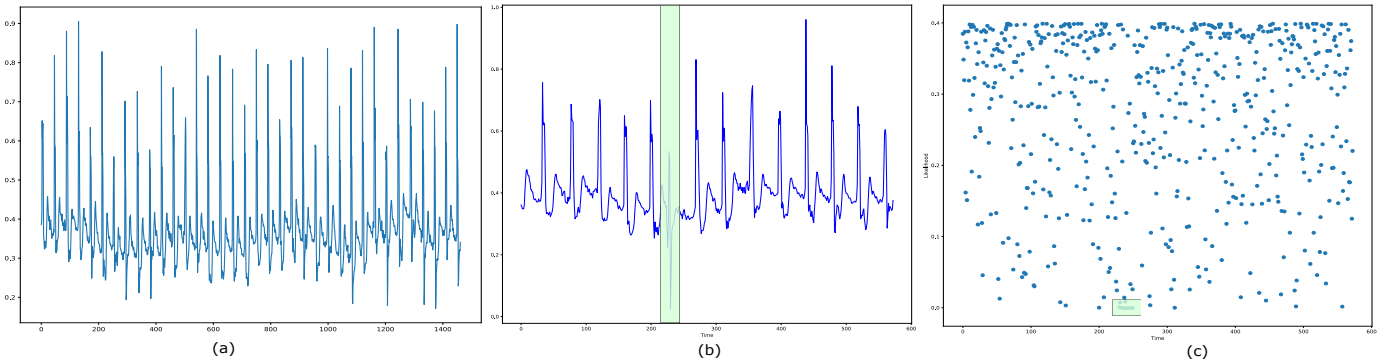



FIGURE 6: Anomaly detection on the ecg dataset: as it is easy to spot the pattern in the train set (figure (a)), in the test set (figure (b), where the green vertical bars highlight the anomalous zone) there's an anomalous sequence between two non-anomalous peaks. The CVAE algorithm spots it with a high density zone of low probability sequences (green rectangles on figure (c), where on $y$-axis it is indicated the probability of each timestep).


[11] X. Guo, X. Liu, E. Zhu, J. Yin, *Deep clustering with convolutional autoencoders*, International Conference on Neural Information Processing Springer, Cham, 2017.

[12] Aytekin, C., Ni, X., Cricri, F., Aksu, E. *Clustering and Unsupervised Anomaly Detection with l 2 Normalized Deep Auto-Encoder Representations*, International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE, 2018

[13] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer-Verlag Berlin Heidelberg, Vol. 385, 2012.

[14] K. Yamanishi, J. Takeuchi, *A unifying framework for detecting outliers and change points from time series*, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, Issue 4, pp. 482-492, 2006.

[15] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu , *Recurrent Neural Networks for Multivariate Time Series with Missing Values*, Scientific Reports 8, Article Number 6085, 2018.

[16] M. Assaad, R. Bon, H. Cardot, *A New Boosting Algorithm for Improved Time-Series Forecasting with Recurrent Neural Networks*, Information Fusion 9, pp. 41-55, 2008.

[17] O. Ogunmolu, X. Gu, S. Jiang, N. Gans, *Nonlinear systems identification using deep dynamic neural networks*, arXiv: 1610.01439, 2016.

[18] N. Laptev, J. Yosinski, L. E. Li, S. Smyl, *Time-series Extreme Event Forecasting with Neural Networks at Uber*, Uber AI, ICML, 2017.

[19] L. de Haan, A. Ferreira, *Extreme Value Theory: An Introduction*, Springer Series in Operations Research and Financial Engineering, 2006.

[20] P. Malhotra, L. Vig, G. Shroff, A. Puneet, *Long short term memory networks for anomaly detection in time series*, ESANN 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[21] S. Hochreiter, J. Schmidhuber, *Long short-term memory*, Neural Computation, Vol. 9, Issue 8, pp. 1735-1780, 1997.

[22] G. Tsang, J. Deng, X. Xie, *Recurrent Neural Networks for Financial Time-Series Modelling*, 24th International Conference on Pattern Recognition, ICPR, 2018.

[23] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, G. Shroff, A. Puneet, *LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection*, ArXiv: 1607.00148, 2016.

[24] D. P. Kingma, M. Welling, *Auto-Encoding Variational Bayes*, arXiv: 1312.6114, 2014.

[25] J. Armstrong, F. Collopy, *Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons*, International Journal of Forecasting, Vol. 8, Issue 1, pp. 69-80, 1992.

[26] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, *Deconvolutional networks*, Cvpr (Vol. 10, p. 7), 2010.

[27] D. J. Rezende, S. Mohamed, *Variational inference with normalizing flows*, arXiv preprint arXiv:1505.05770, 2015.

[28] S. Makridakis, *Accuracy Measures: Theoretical and Practical Concerns*, International Journal of Forecasting, Vol. 9, Issue 4, pp. 527-529, 1993.

[29] R. J. Hyndman, A. B. Koehler, *Another Look at Measures of Forecast Accuracy*, International Journal of Forecasting, Vol. 22, Isuue 4, pp. 679-688, 2006.

# Feature Selection based Multivariate Time Series Forecasting: An Application to Antibiotic Resistance Prediction

José Palma[1], Fernando Jiménez[1], Gracia Sánchez[1], David Marín-García[2], Francisco Palacios, MD[1], and Lucía López, MD[3]

[1] Artificial Intelligence and Knowledge Engineering Group. Faculty of Computer Science. University of Murcia, Spain
[2] International Doctorate School. University of Murcia, Spain
[3] Universitary Hospital of Getafe, Madrid, Spain

**Abstract.** In this paper we propose a methodology to build a model for predicting future outbreaks of *Methicilin-resistant Staphylococcus aereus* (MRSA). Infection incidence forecasting is approached as a *feature selection based time series forecasting* problem using multivariate time series composed of incidence of *Staphylococcus Aereus* and MRSA infections, influenza incidence and total days of therapy of both of *Levoflaxin* and *Oseltamivir* antimicrobials. Data were collected from the University Hospital of Getafe (Spain) from January 2009 to January 2018, using months as time granularity. The proposed methodology includes the application of wrapper multivarite feature selection methods on transformed datasets with a different number of lagged variables. The search strategy of the feature selection method is based on *multi-objective evolutionary algorithms* and the most powerful state-of-the-art regression algorithms are used as evaluators. The performance of the feature selection methods has been measured using both *root mean square error* ($\mathcal{RMSE}$) and *mean absolute error* ($\mathcal{MAE}$) metrics. In order to select the most satisfactory regression model, including multivariate *ARIMA* and *VAR* autoregressive models, a novel multi-criteria decision-making process is proposed. Results show that the best model according to the proposed multi-criteria decision making process provides a $\mathcal{RSME} = (0.1349, 0.1304, 0.1325)$ and a $\mathcal{MAE} = (0.1003, 0.096, 0.0987)$ for 1, 2, and 3 steps-ahead predictions.

**Keywords:** Feature Selection, Multi-objective Evolutionary Algorithms, Multivariate Time Series, Antibiotic Resistance Forecasting, Multiple Criteria Decision Making.

## 1 Introduction

The massive use of antimicrobials and, more important, their misuse is threatening with an increasing spread of multi-resistant bacteria which can cause infections with fatal consequences. According to recent studies, antimicrobial resistance (AMR) is estimated to be responsible for 25,000 death per year in the EU [1] and 700,000 deaths per year globally. As a result, AMR has become one of the most important health problems and

global action plans have been proposed [2,3]. Prevention plays a key role in these action plans and, in this context, we proposed the use of *Artificial Intelligence*, specifically *Time Series Forecasting* techniques, for predicting outbreak of multi-resistant bacteria from hospital level data.

To this end, we have focused on infections caused by *Staphylococcus Aureus* (SA) and *Methicilin-resistant Staphylococcus Aereus* (MRSA). MRSA is a methicilin-resistant SA strain which is able to persist not only at hospital level (where the use of antimicrobial agents is high) but also at the community level. Fighting SA and MRSA infections requires a huge effort both in health and economic terms. Therefore, early and reliable detection of infections outbreaks will allow to efficiently reallocate the available scarce resources to avoid infection propagation.

In this work, infection incidence forecasting is approached using Machine Learning techniques. To this end, time series data must be transformed by removing the temporal ordering of individual input examples and adding a set of delays to the input which are called *lagged attributes*, to approach the problem with regression techniques. This approach to time series forecasting is more powerful and more flexible than classical statistics techniques such as *ARMA* and *ARIMA* [4] which make more emphasis on modelling aspects. Feature selection methods are applied for the selection of lagged variables. We have also considered multivariate autoregressive models, such us vector autorregresive models (*VAR*) and multivariate *ARIMA* (*MARIMA*) in the set of experiments. A novel multi-criteria decision making process is applied to choose the most satisfactory model for the *1,2,3-step-ahead* predictions, where both *root mean square error* ($\mathcal{RMSE}$) and *mean absolute error* ($\mathcal{MAE}$) performance metrics are considered. The experiments have been carried out using the *Waikato Environment for Knowledge Analysis* (*Weka*) [5] and the `marima` and `vars` R packages [6,7].

The paper has been organized as follows: section 2 shows the related works. The data set used in this work are described in section 3. Section 4 proposes a methodology for multivariate time series forecasting of antibiotic resistance based on feature selection. Section 5 analyses and discusses the results, and finally section 6 concludes the paper and outlines futures works.

## 2 Related Work

An analysis of the optimal number of lag variables that should be used for time series forecasting with *Random Forest* can be found in [8]. In [9], FS wrapper methods have been used for time series prediction using *Neural Networks*. Granger causality discovery has been used to identify key features with effective sliding window sizes, considering the influence of lagged observations of features on the target time series [10]. Other studies have searched for the optimal time-windows and time lags for each variable based on feature pre-processing and sparse learning to configure the input dataset [11].

The first works in the application of time series analysis to antibiotic resistance are shown in [12]. In [13], the same authors demonstrated a temporal relationship between antimicrobial use and resistance, they also presented a technique to quantify the effect of the use of antimicrobials on resistance and how to estimate the delay between variations of use and subsequent variations in resistance. The association between an-

timicrobial use and resistance rates in *Pseudomonas aeruginosa* (PA) using time series analysis has been presented in [14]. In [15], two statistical methods (Pearson's correlation and distributed lags time series analysis) are compared to determine their ability to analyse the relationship between antipseudomonal antimicrobial consumption and resistance rates of PA. In [16], time series analysis on multiple longitudinal datasets has demonstrated their potential in microbiome research. In [17], antimicrobial drug consumption has been predicted using selected lag variables of time series of web search frequency associated with antimicrobial consumption and making use of *Linear Elastic Net* and *Gaussian Processes* models.

## 3 Antibiotic resistance dataset

For this experiment, we have used a multivariate time series dataset with five time series. Each time series is composed of 109 events collected from a hospital between January 2009 to January 2018, using months as time granularity. Figure 1 shows the five time series used in this work and the unit in which each series is measured is shown in table 1. We have selected these time series since they are involved in the influenza protocol. Influenza first symptoms are treated with Oseltamivir antiviral drug so as to improve disease symptoms. In order to prevent bacterial infections as a complication of influenza, Levofalxin antibiotic is also administrated. The most common, and more risky, bacterial infections are those provoked by SA and MRSA which, as said before, can lead to fatal consequences.

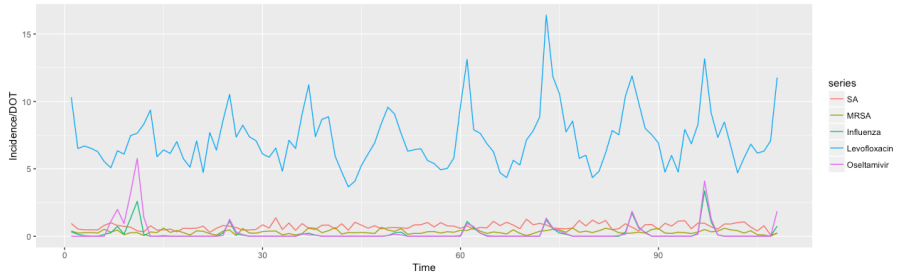



Fig. 1: Time series used in this work.

SA and MRSA time series are measured in monthly incidences and Levoflaxin and Oseltamivir in total days of therapy (DOT) by months. Incidence is calculated as a proportion between the number of inpatients that, before the moment that it is calculated, are affected by the event (that is, new events) and the total number of inpatients. DOT represents the number of days in which a patient is treated with the corresponding antimicrobial. Therefore, Levofalxin and Oseltamivir time series represent the sum of DOT over a month in the hospital.

| Series | Unit |
|---|---|
| Staphylococcus aureus (SA) | Incidence |
| Staphylococcus aureus meticilin resistant (MRSA) | Incidence |
| Influenza | Incidence |
| Levoflaxin | Days of Therapy (DOT) |
| Oseltamivir | Days of Therapy (DOT) |

Table 1: Time series considered in this work and their units.

## 4 A methodology for multivariate time series forecasting of antibiotic resistance based on feature selection

The following five steps have been systematically applied: dataset transformation, feature selection, regression, decision making and forecasting. Figure 2 summarizes the methodology proposed.

### 4.1 Database transformation

The first step of our methodology is to transform datasets by creating lagged versions of variables for use in the time series problem. We have experienced setting the *maximum lag length* to 2, 4 and 6. Therefore, three transformed datasets have been created (one for each lag length 2, 4 and 6) containing respectively 16, 26 and 36 attributes in total ($(laglength + 1) \cdot 5 + 1$ where 5 is the number of lagged attributes). These transformed datasets will be used later in the forecasting phase.

### 4.2 Feature selection

Once the task of transforming the dataset is done, the next step is to apply FS on the transformed datasets. We applied 12 different multivariate FS methods. We are interested in the wrapper methods due to its greater precision. In all the applied wrapper methods, a Multiobjective Evolutionary Search has been used as a search strategy. Specifically, *ENORA* multi-objective evolutionary algorithm has been used [18]. *ENORA* is an elitist Pareto-based multi-objective evolutionary algorithm that uses a $(\mu + \lambda)$ survival with uniform random initialisation, binary tournament selection, ranking based on local non-domination level with crowding distance, self-adaptive uniform crossover and self-adaptive one-bit flip mutation. Previous works have demonstrated that *ENORA* provides great efficiency in feature selection problems for regression [18] and performs better than the well-known *NSGA-II* in terms of *hypervolume* [19,20] for regression tasks.

The use of multi-objective evolutionary techniques allows the optimisation of two objectives when performing FS with wrapper methods. The first one is to maximise the performance metric chosen in the evaluator. The second one is to minimise the subset cardinality. The non-dominated solutions in the last population with the best fitness for the first objective are shown as output.

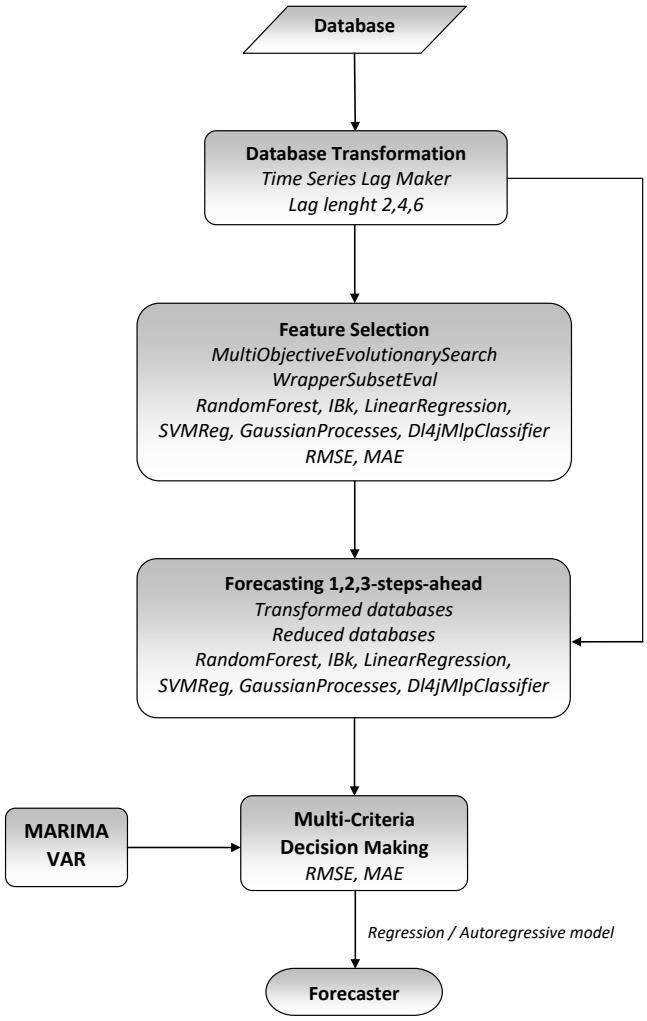



Fig. 2: Methodology for feature selection for antibiotic resistance multivariate time series forecasting.

Apart from the search strategy, a multivariate FS wrapper method requires an evaluator to evaluate every attribute subset generated by the search strategy. As evaluators, we have considered six regression algorithms: *Random Forest* [21], *K-nearest neighbours classifier* [22], *Linear Regression* [23], *Support Vector Machine* [24], *Gaussian Processes* [25] and *Multilayer Perceptron*.

### 4.3 Forecasting

We have considered 3 lag lengths (2, 4 and 6) and 12 feature selection methods. This gives a total of $3 + (3 \cdot 12) = 39$ datasets (3 transformed datasets plus 36 reduced datasets). The following predictions have been made for these datasets:

- The six regression algorithms have been used for forecasting in the transformed datasets with 2, 4 and 6 lag length ($3 \cdot 6 = 18$ regression models) and for 1, 2 and 3 step-ahead ($18 \cdot 3 = 54$ predictions).
- Predictions for the reduced datasets (12 FS selection method applied to the 3 transformed datasets) have been obtained applying the regression algorithm used for the wrapper feature selection method. A total of $3 \cdot 12 = 36$ regression models have been obtained with a total of $36 \cdot 3 = 108$ 1, 2 and 3 step-ahead predictions.

Therefore, a maximum number of $54$ regression models and $162$ predictions are obtained with this methodology. However, those cases in which the feature selection method does not select any lag variable for the output variable are discarded. In our case, 11 reduced datasets have been discarded, resulting in a total of $43$ regression models and $129$ predictions.

Finally, our methodology also makes predictions with *autoregressive models* using the `marima` and `vars` packages. The `vars` package [7] fits a *VAR* model, and the `marima` package fits *MARIMA* model using the Spliid's algorithm [6]. Since *VAR* model includes only autoregresive terms only [26], a *VARMA* model, which includes both autoregressive and moving average terms [27], has also been built. However, the contribution of the moving average component of the *VARMA* model was neglible, so only *VAR* model was taking into account, together with the *MARIMA* model. All regression models have been trained on the first $70\%$ instances and tested on the last $30\%$ instances

### 4.4 Multiple criteria decision making

The next step in our methodology is to compare all the predictions made in order to choose, either a regression method obtained through dataset transformation + feature selection, or *MARIMA* or *VAR* autoregressive models. For this purpose, we propose the following *multiple criteria decision making* process:

Let $X = \{x_1, \ldots, x_n\}$ a set of $n$ prediction models. In our case, $n = 45$ (43 regression models plus 2 autoregressive models). We consider the following *multi-objective optimization problem*:

$$\begin{array}{l} Min\ \mathcal{RMSE}(x,i),\ i = 1, \ldots, p \\ Min\ \mathcal{MAE}(x,i), \quad i = 1, \ldots, p \end{array} \tag{1}$$

In (1), $x \in \{X\}$ is a decision variable that represents a regression model. We use, as performance metrics the $\mathcal{RMSE}$ and the $\mathcal{MAE}$ of $n$ step-ahead predictions (a total of $2 \cdot n$ objective functions for minimisation) on test data ($30\%$).

Solution of (1) is a set $S = \{s_1, \ldots, s_m\} \subset X$, $m \leq n$, of *non-dominated* (or *Pareto*) solutions [28]. In order to choose a solution $s^* \in S$, we take into account the

sum of the values of $\mathcal{RMSE}$ and $\mathcal{MAE}$ in the 1, 2 and 3 step-ahead, together with the sum of the slopes (in absolute value) of the prediction lines evaluated with $\mathcal{RMSE}$ and $\mathcal{MAE}$ in 1, 2 and 3 step-ahead. Algorithm 1 describes the full multiple criteria decision making process.

**Algorithm 1** Multiple criteria decision making

**Require:** $X = \{x_1, \ldots, x_n\}$ {Set of $n$ prediction models}
1: $S = \{s_1, \ldots, s_m\} \leftarrow$ Solution of the multi-objective optimization problem (1)
2: $\mathcal{RMSE}'(s_j, i) \leftarrow$ Normalized $\mathcal{RMSE}(s_j, i), j = 1, \ldots, m, i = 1, \ldots, p$
3: $\mathcal{MAE}'(s_j, i) \leftarrow$ Normalized $\mathcal{MAE}(s_j, i), j = 1, \ldots, m, i = 1, \ldots, p$
4: $sRMSE_j \leftarrow \sum_{i=1}^{p} \mathcal{RMSE}'(s_j, i), j = 1, \ldots, m$
5: $sMAE_j \leftarrow \sum_{i=1}^{p} \mathcal{MAE}'(s_j, i), j = 1, \ldots, m$
6: $mRMSE_j \leftarrow \sum_{i=1}^{p-1} \mid \mathcal{RMSE}'(s_j, i+1) - \mathcal{RMSE}'(s_j, i) \mid, j = 1, \ldots, m$
7: $mMAE_j \leftarrow \sum_{i=1}^{p-1} \mid \mathcal{MAE}'(s_j, i+1) - \mathcal{MAE}'(s_j, i) \mid, j = 1, \ldots, m$
8: $v_j \leftarrow sRMSE_j \cdot mRMSE_j + sMAE_j \cdot mMAE_j, j = 1, \ldots, m$
9: $s^* \leftarrow s_{min} \mid v_{min} = \min_{j=1}^{m}\{v_j\}$
10: **return** $s^*$

## 5 Analysis of results and discussion

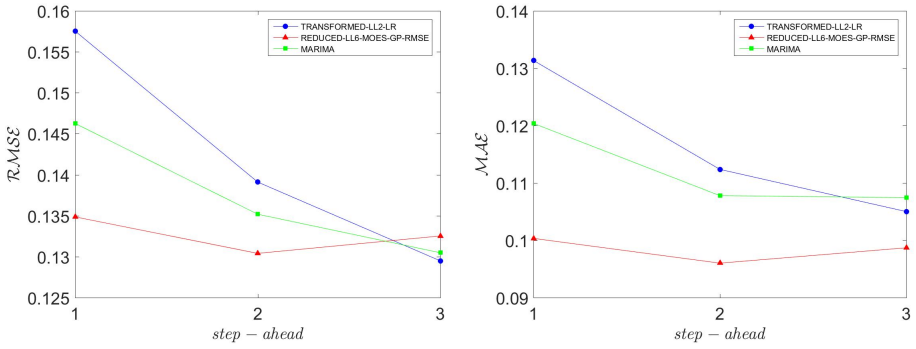
we have obtained a set $S = \{s_1, s_2, s_3\}$ composed of 3 non-dominated solutions as solutions to (1). Solution $s_1$ is the regression model obtained from the transformed dataset with lag length 2 (without feature selection) and the *Linear Regression* algorithm. Solution $s_2$ is the regression model obtained from the reduced dataset obtained with the wrapper FS method *MOES-GP-RMSE* applied to the transformed dataset with lag length 6 and the *Gaussian Processes* algorithm. Solution $s_3$ is the autoregressive model obtained with *MARIMA*. Table 2 shows the $\mathcal{RMSE}$ and $\mathcal{MAE}$ of each solution in 1, 2 and 3 step-ahead. The value $v_j$ of each solution, calculated by algorithm 1, is also shown. According to algorithm 1, the best solution is $s_2$. Not only does it has the lowest sum of $\mathcal{RMSE}$ and $\mathcal{MAE}$ in the 1, 2 and 3 step-ahead predictions, but the sum of the its prediction lines slope (both with $\mathcal{RMSE}$ and $\mathcal{MAE}$) in the 1, 2 and 3 step-ahead are minimum (see figures 3(a) and 3(b))

In this way, the multiple criteria decision making process takes into account the following aspects:

1. Optimality of the multi-objective problem (1), which identifies the best solutions in each step-ahead for both $\mathcal{RMSE}$ and $\mathcal{MAE}$ metrics.
2. To distinguish between non-dominated solutions, the following aspects are taken into account:
   (a) The joint optimality of the solution in all the steps ahead taking into account $\mathcal{RMSE}$ and $\mathcal{MAE}$ metrics separately.

| | | *1-step-ahead* | *2-step-ahead* | *3-step-ahead* |
|---|---|---|---|---|
| $s_1$ | $\mathcal{RMSE}(s_1,i)$ | 0.1575 | 0.1391 | 0.1295 |
| | $\mathcal{MAE}(s_1,i)$ | 0.1313 | 0.1124 | 0.1050 |
| | $v_1$ | 0.0636 | | |
| $s_2$ | $\mathcal{RMSE}(s_2,i)$ | 0.1349 | 0.1304 | 0.1325 |
| | $\mathcal{MAE}(s_2,i)$ | 0.1003 | 0.0960 | 0.0987 |
| | $v_2$ | **0.0030** | | |
| $s_3$ | $\mathcal{RMSE}(s_3,i)$ | 0.1463 | 0.1352 | 0.1305 |
| | $\mathcal{MAE}(s_3,i)$ | 0.1204 | 0.1078 | 0.1074 |
| | $v_3$ | 0.0232 | | |

Table 2: Non-dominated solutions and performances in the 1, 2 and 3 step-ahead.




(a) (b)

Fig. 3: $\mathcal{RMSE}$ (a) and $\mathcal{MAE}$ (b) in 1, 2 and 3 step-ahead of the non-dominated solutions.

(b) The robustness of the regression model along all the steps ahead using the sum of the slopes of the prediction lines (in absolute value) between each two steps ahead, again in both $\mathcal{RMSE}$ and $\mathcal{MAE}$ metrics separately.
(c) The joint optimality and robustness of the regression model are aggregated into a single function by the multiplication operator (step 8 of algorithm 1).

To analyse the effectiveness of the FS process, we compared the best regression model obtained with our proposal (solution $s_2$) with the equivalent regression model without applying feature selection (regression model obtained with the transformed dataset with lag length 6 and the *GaussianProcesses* algorithm). Figure 4 graphically shows the predictions in 1, 2 and 3 step-ahead in test data for both models (without feature selection on the top, and with feature selection on the bottom). The following statements can be made:

1. The solution with feature selection dominates the solution without feature selection.
2. Solution without feature selection obtained $\mathcal{RMSE} = (0.1733, 0.1752, 0.1740)$ and $\mathcal{MAE} = (0.1349, 0.1426, 0.1461)$. This means that the feature selection pro-

cess has reduced the $\mathcal{RMSE}$ by $23.17\%$, and the $\mathcal{MAE}$ has been reduced by $30.36\%$.

3. In addition to being more accurate, the regression model obtained with feature selection is more robust, since the predictions in 1, 2 and 3 step-ahead are the same. Note that, in the graph on the bottom of figure 4, the prediction lines in 2 and 3 step-ahead are covered by the prediction line in the 1-step-ahead. However, table 2 shows different $\mathcal{RMSE}$ and $\mathcal{MAE}$ values for the predictions in 1, 2 and 3 step-ahead, since, in the 2-step-ahead, one instance is evaluated less than in 1-step-ahead, and in the 3-step-ahead, two less instances are evaluated.

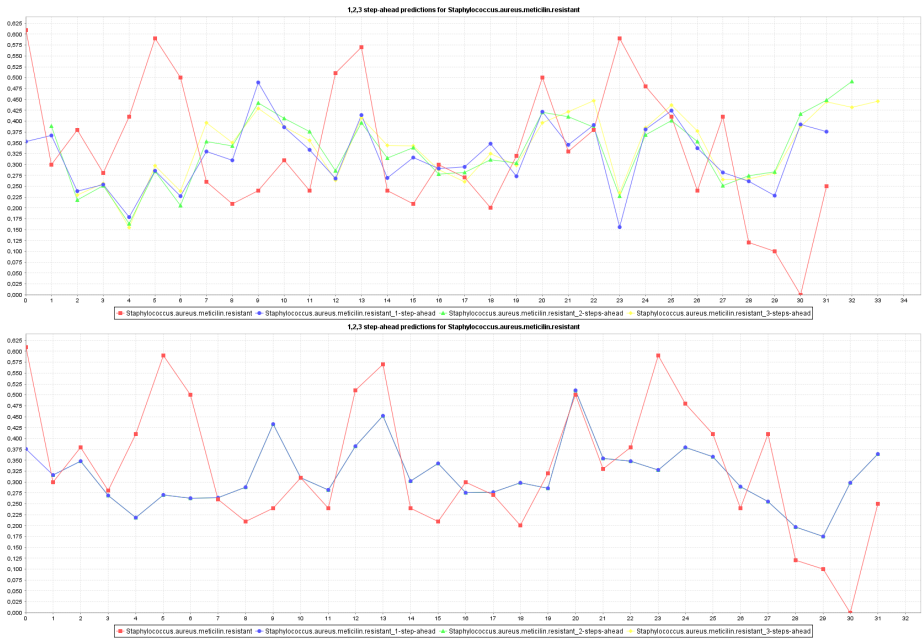



Fig. 4: 1,2,3-step-ahead predictions for MRSA evaluated on test data with *Transformed dataset, lag length 6, gaussian processes* (top) and with *Reduced dataset with MOES-GP-RMSE* from transformed dataset with lag length 6, gaussian processes (bottom).

In this paper, we propose a methodology that allows the selection of the most accurate time series forecasting model from a set of models that differ in the number of lagged variables, selected variables and regression algorithms used to build the model. Although the methodology can potentially be applied in any area, to the best of our knowledge this is the first time that a methodology for multivariate time series based on feature selection is proposed for antibiotic resistance forecasting. The main contributions of the work are, on the one hand, the application of wrapper multivariate feature selection methods with search strategy based on *multi-objective evolutionary algorithms* and the most powerful state-of-the-art regression algorithms as evaluators

(*Random Forest*, *K-nearest neighbours classifier*, *Linear Regression*, *Support Vector Machine*, *Gaussian Processes* and *Multilayer Perceptron*); and, on the other hand, a novel multi-criteria decision making process in order to select the most satisfactory model, using $\mathcal{RMSE}$ and $\mathcal{MAE}$ as performance metrics, as well as the prediction lines slopes at 1, 2 and 3 step-ahead for the sake of robustness.

## 6 Conclusions

In this paper, we propose a methodology that allows the selection of the most accurate time series forecasting model from a set of models that differ in the number of lagged variables, selected variables and regression algorithms used to build the model. Although the methodology can potentially be applied in any area, to the best of our knowledge this is the first time that a methodology for multivariate time series based on feature selection is proposed for the forecasting of antibiotic resistance infections outbreaks. The main contributions of the work are, on the one hand, the application of wrapper multivariate feature selection methods with search strategy based on *multi-objective evolutionary algorithms* and the most powerful state-of-the-art regression algorithms as evaluators (*Random Forest*, *K-nearest neighbours classifier*, *Linear Regression*, *Support Vector Machine*, *Gaussian Processes* and *Multilayer Perceptron*); and, on the other hand, a novel multi-criteria decision making process, which is approached as a multi-objective optimization problem in which the $\mathcal{RMSE}$ and $\mathcal{MAE}$ metrics, in different step-ahead predictions, are used in defining the problem objectives. The robustness of the regression models along the step-ahead predictions is also taken into account. Results show that the regression model obtained by feature selection improves by $23.17\%$ and by $30.36\%$ the $\mathcal{RMSE}$ and $\mathcal{MAE}$ respectively of the regression model without applying feature selection, as well as its robustness in the 1, 2 and 3 step-ahead predictions.

From the clinical point of view, the proposed mathematical models can provide more objectivity and quantification capabilities to the visual analysis of the temporal series carried out by epidemiologist experts. Furthermore, the models and the selected variables, make possible to extract knowledge from the temporal series. Predictions of future infections outbreaks allow the reallocation of resources (scarce and insufficient) to control de infection and avoid its propagation. Finally, in a context with a high probability of an outbreak according to predictions, epidemiological active surveillance techniques could adjust its sensitivity and specificity improving the outbreak early diagnosis.

Among future works, the use of information regarding doses is to be approached. Other open lines are related to the automation of the methodology proposed, including an automatic selection of the time series relevant for forecasting. Finally, to make possible the integration of the process in clinical practice, providing results in terms of probability and confidence intervals are going to be tackled.

## Acknowledgments

This work was partially funded by the Spanish Ministry of Science, Innovation and Universities under the SITUS project (Ref: RTI2018-094832-B-I00), and by the European Fund for Regional Development (EFRD, FEDER).

## References

1. ECDC/EMEA Joint Technical Report, The bacterial challenge: time to react, Technical Report EMEA/576176/2009, European Centre for Disease Prevention and Control (2009).
2. European Commission, A european one health action plan against antimicrobial resistance (amr), Technical report, European Commission (2017).
3. World Health Organization, Global action plan on antimicrobial resistance, Technical Report WHA68/2015/REC/1, Annex 3, World Health Organization (2015).
4. R. Adhikari, R. K. Agrawal, An introductory study on time series modeling and forecasting, CoRR abs/1302.6613 (2013). arXiv:1302.6613.
5. I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston, 2011 (2011).
6. H. Spliid, Multivariate ARIMA and ARIMA-X Analysis, CRAN, license GPL-2, Version 2.2, RoxygenNote 5.0.1 (2017).
7. B. Pfaff, Var, svar and svec models: Implementation within r package vars, Journal of Statistical Software 27 (4) (2008).
   URL http://www.jstatsoft.org/v27/i04/
8. H. Tyralis, G. Papacharalampous, Variable selection in time series forecasting using random forests, Algorithms 10 (4) (2017).
9. S. F. Crone, N. Kourentzes, Feature selection for time series prediction–a combined filter and wrapper approach for neural networks, Neurocomputing 73 (10-12) (2010) 1923–1936 (2010).
10. Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, R. Wang, Using causal discovery for feature selection in multivariate numerical time series, Machine Learning 101 (1-3) (2015) 377–395 (2015).
11. S. Hido, T. Morimura, Temporal feature selection for time-series prediction, in: 2012 21st International Conference on Pattern Recognition (ICPR 2012), IEEE, 2012, pp. 3557–3560 (2012).
12. J.-M. López-Lozano, D. L. Monnet, P. C. Alonso, A. C. Quintero, N. G. Jiménez, A. Y. Muñoz, C. Thomas, A. Beyaert, M. Stevenson, T. V. Riley, Applications of time-series analysis to antibiotic resistance and consumption data, in: Antibiotic Policies, Springer, 2005, pp. 447–463 (2005).
13. J.-M. López-Lozano, D. L. Monnet, A. Yagüe, A. Burgos, N. Gonzalo, P. Campillos, M. Saez, Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis, International Journal of Antimicrobial Agents 14 (1) (2000) 21 – 31 (2000).
14. M. Willmann, M. Marschal, F. Hölzl, K. Schröppel, I. B. Autenrieth, S. Peter, Time series analysis as a tool to predict the impact of antimicrobial restriction in antibiotic stewardship programs using the example of multidrug-resistant pseudomonas aeruginosa, Antimicrobial agents and chemotherapy 57 (4) (2013) 1797–1803 (2013).
15. V. Erdeljić, I. Francetić, Z. bošnjak, A. Budimir, S. Kalenic, L. Bielen, K. makar ausperger, R. Likic, Distributed lags time series analysis versus linear correlation analysis (pearson's r) in identifying the relationship between antipseudomonal antibiotic consumption and the

susceptibility of pseudomonas aeruginosa isolates in a single intensive care unit of a tertiary hospital, International journal of antimicrobial agents 37 (2011) 467–71 (05 2011).
16. K. Faust, L. Lahti, D. Gonze, W. M. de Vos, J. Raes, Metagenomics meets time series analysis: unraveling microbial community dynamics, Current Opinion in Microbiology 25 (2015) 56 – 66, environmental microbiology * Extremophiles (2015).
17. N. Dalum Hansen, K. Mølbak, I. J. Cox, C. Lioma, Predicting antimicrobial drug consumption using web search data, in: Proceedings of the 2018 International Conference on Digital Health, DH '18, ACM, New York, NY, USA, 2018, pp. 133–142 (2018).
18. F. Jiménez, G. Sánchez, J. García, G. Sciavicco, L. Miralles, Multi-objective evolutionary feature selection for online sales forecasting, Neurocomputing 234 (2017) 75–92 (2017).
19. E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: empirical results, Evolutionary Computation 8 (2) (2000) 173 – 195 (2000).
20. E. Zitzler, L. Thiele, M. Laumanns, C. Fonseca, V. Grunert da Fonseca, Performance assessment of multiobjective optimizers: An analysis and review, IEEE Transactions on Evolutionary Computation 7 (2002) 117–132 (2002).
21. L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32 (2001).
22. D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1) (1991) 37–66 (jan 1991).
23. X. Yan, Linear Regression Analysis: Theory and Computing, World Scientific Publishing Company Pte Limited, 2009 (2009).
24. S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, K. R. Murthy, Improvements to the smo algorithm for svm regression, Trans. Neur. Netw. 11 (5) (2000) 1188–1193 (Sep. 2000).
25. C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005 (2005).
26. H. Lutkepohl, New Introduction to Multiple Time Series Analysis, Springer Publishing Company, Incorporated, 2007 (2007).
27. H. Spliid, A fast estimation method for the vector autoregressive moving average model with exogenous variables, Journal of The American Statistical Association 78 (1983) 843–849 (12 1983).
28. Y. Collette, P. Siarry, Multiobjective Optimization: Principles and Case Studies, Springer Berlin Heidelberg, 2004 (2004).

# Multi-Objective Evolutionary Optimization for Time Series Lag Regression

Fernando Jiménez[1][0000−0001−9906−4132],
Joanna Kamińska[2][0000−0002−0157−516X],
Estrella Lucena-Sánchez[3,4][0000−0001−9312−1175],
José Palma[1][0000−0003−2502−4378], and
Guido Sciavicco[3][0000−0002−9221−879X]

[1] Dept. of Knowledge Engeneering and Communications,
University of Murcia (Spain)
jtpalma|fernan@um.es
[2] Dept. of Mathematics,
Wrocław University of Environmental and Life Sciences (Poland)
joanna.kaminska@upwr.edu.pl
[3] Dept. of Mathematics and Computer Science,
University of Ferrara (Italy)
estrella.lucenasanchez|guido.sciavicco@unife.it
[4] Dept. of Physics, Informatics, and Mathematics
University of Modena e Reggio Emilia (Italy)

**Abstract.** It is well-known that in some regression problems the effect of an independent variables on the dependent one(s) may be delayed; this phenomenon is known as lag. Lag regression is one of the standard techniques for time series explanation and prediction. However, using lagged variables to transform a multivariate time series so that a propositional algorithm such as a linear regression learner can be used requires to decide, at preprocessing time, which independent variables must be lagged and by how much. In this paper, we propose a novel optimization schema to solve this problem. We test our solution, implemented with a multi-objective evolutionary algorithm, on real data taken from a larger project that aims to construct an explanation model for the study of atmospheric pollution in the city of Wrocław (Poland).

**Keywords:** Regression; Lag; Multi-objective evolutionary computation; Time series explanation

## 1 Introduction

A *time series* is a series of data points labelled with a temporal stamp. If each data point contains a single time-dependent value, then the time series is *univariate*; otherwise, it is called *multivariate*. Time series arise in multiple contexts, for example, medical patients, who can be considered as time series in which every interesting medical value varies over time (e.g., fever, pain level, blood pressure), or environmental monitoring stations, which can also be considered time series,

in which atmospheric values change over time (e.g., pressure, concentration of chemicals).

There are two main problems associated with single time series: *time series explanation* and *time series forecasting*; explanation is a necessary step for forecasting, but the latter does not necessarily follow the former in every application and context. Explaining a time series aims to construct a (possibly interpretable) model that explains the present values; forecasting a time series implies testing and using the model to predict future values. In the univariate case, a model of a time series is based uniquely on the values of the series itself. For example, a forecasting model for the stock price of a certain company would allow one to predict the future price (e.g., in the next two days) based on the prices of the same company (e.g., the price in each day of the past week). The simplest univariate forecasting approach is commonly known as *Simple Moving Average* (SMA) model: in essence, a simple moving average is calculated over the time series by considering its last $n$ values, used to perform a smoothing process of the series, and then used to forecast for the next value. Although such an approach has some clear limitations, it is still useful to establish a baseline, against which to compare more complex solutions [3]. Based on the observation that the most recent values may be more indicative of a future trend than older ones, *Simple Exponential Smoothing* (SES) models consider a weighted average over the last $n$ observations, assigning exponentially decreasing weights as they get older [3]. Other than this first, simple type of smoothing, it is also worth mentioning *Holt's Exponential Smoothing* (HES) models [9], which can consider an increasing or decreasing trend in the time series, and *Holt-Winters' Exponential Smoothing* (HWES) [14] models, that can take into account seasonality effects. Technically, exponential smoothing belongs to the broader *AutoRegressive Integrated Moving Average* (ARIMA) family [11], which includes models that can be fitted to time series data either to better understand the data itself or to predict future points in the series, when it shows evidence of non-stationarity. Specifically, the methods that are capable of dealing with periodical variations in the time series fall under the umbrella of *Seasonal ARIMA* [3]. Relevant to this study is also the algorithm presented in [1], in which a multi-objective evolutionary method is employed for the optimization of the parameters of an ARIMA-like model. The common aspect among all univariate models is that they make a prediction based on a weighted linear sum of recent past observations; in the multivariate case, instead, one identifies one dependent variable (time series), and aims to construct a model to explain and/or predict its future values based on the past and present values of other, independent variables (which themselves are time series): this is usually done with *lagged* models. While ARIMA-type models emerge from computational statistics, lagged models belong to the machine learning domain, and, in general, they consist of creating *lagged* version of (a subset of) the independent variable to construct a larger data set that is then used to create a model of the dependent time series using classical, propositional algorithms (such as, for example, linear regression). Among the available packages to this purpose we mention WEKA's *timeseriesForecasting* [7]. Other approaches to multivari-

ate time series modelling include *Recurrent Neural Networks* (RNNs)[8], which have been used for time series forecasting with promising results, but at the expenses of the interpretability of the resulting model; in some recent works, neural networks for time series forecasting have been trained and optimized with multi-objective evolutionary algorithms. Autoregressive techniques can be combined with lagged methodologies; in the simplest case, it is sufficient to create, in a lagged extended data set, one or more lagged version(s) the dependent variable as well, whose values are combined with those of the independent ones.

The main limitation of multivariate lagged models is precisely the choice of lag variables and lag amounts. In some cases, it is difficult to foresee the necessary lag amount. Moreover, uncontrolled lag variable creation may lead to very large data bases which, when treated with propositional algorithms, may lead to poorer results, as unnecessary lag variables become noise. Finally, even if lagged variables increase the quality of the result, the obtained function may not be easy to interpret. In this work we present a very simple optimization schema that avoids the above problems for time series explanation using regression. The distinctive characteristics of our method are: *(i)* it is a *wrapper* algorithm based on well-known and easy-to-implement components, *(ii)* it may use any *black box* regression algorithm, and *(iii)* it includes an intrinsic feature selection mechanism. Our algorithm is an instantiation of the more general *dynamic preprocessing* mechanism, which generalizes the concept of wrapper by allowing the (possibly simultaneous) optimization of several aspects of data.

We test our model on a real data set taken from a larger project that aims to construct an explanation and prediction model for the study of atmospheric pollution in the city of Wrocław (Poland).

## 2 Lag Regression

### 2.1 Mathematical Formulation

Regression is a common statistical data analysis technique, used to determine the extent to which there is a mathematical relationship between a dependent variable and one or more independent variables, and its applications range from biology, to agriculture, to food and water resources optimization (see, e.g. [2, 12, 13]. Regression can be *univariate*, when there is only one independent variable, or *multivariate*, otherwise. Moreover, regression is usually *linear*, that is, it is usually the case that we search for a linear relationship; it becomes *non-linear*, when we search for any function (whose form is unknown) that links the independent variable(s) and the dependent one. Linear regression is not only the most common type, but it is also the one that presents the clearest mathematical formalization. In the following, and in our experiments as well, we assume that the relations that we search for are, in fact, linear; the entire optimization model, however, works for any type of regression.

Given a data set $A$ with $n$ independent variables $A_1, \ldots, A_n$ and one observed variable $B$, solving a linear regression problem consists of finding a vector $\bar{c} = (c_0, c_1, \ldots, c_n)$ of $n+1$ *parameters* (or *coefficients*) so that the equation:

$$B = c_0 + \sum_{i=1}^{n} c_i \cdot A_i + \epsilon, \tag{1}$$

where $\epsilon$ is a random value, is satisfied. Starting from a data set of observations:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{bmatrix} \tag{2}$$

the regression problem is usually solved by suitably estimating $\bar{c}$ so that, for each $1 \leq j \leq m$:

$$b_j \approx c_0 + \sum_{i=1}^{n} c_i \cdot a_{ij} + \epsilon. \tag{3}$$

The performance of such an estimation can be measured in several (standard) ways, such as *correlation, covariance, mean squared error*, among others. When $A$ is a multivariate time series, composed by $n$ independent and one dependent time series, then data are temporally ordered and associated to a timestamp:

$$A = \begin{bmatrix} t_1 & a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ t_2 & a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ t_m & a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{bmatrix} \tag{4}$$

Using linear regression to explain $B$, then, entails that, instead of (1), we are finding optimal coefficients for:

$$B(t) = c_0 + \sum_{i=1}^{n} c_i \cdot A_i(t) + \epsilon, \tag{5}$$

because we aim to explain $B$ at a certain point in time $t$ using the values $A_1(t), \ldots, A_n(t)$.

Lag (linear) regression consists of solving a more general equation, whose formulation is:

$$B(t) = c_0 + \sum_{i=1}^{n} \sum_{k=0}^{p_i} c_{i,k} \cdot A_i(t-k) + \epsilon. \tag{6}$$

In other words, we use the value of each independent variable $A_i$ not only at time $t$, but also at time $t-1, t-2, \ldots, t-p_i$, to explain $B$ at time $t$; each $A_i(t-k)$ is associated to a coefficient $c_{i,k}$, which must be estimated, along with each $m_i$. We work under the additional assumption that, for each $i$, there is precisely one

lag $k$, denoted $k_i$, such that $A_i(t-k_i)$ influences the output more than any other lag. Our purpose is to devise an optimization schema that allows one to estimate both the value $k_i$ and the coefficient $c_i$ that corresponds to it, to obtain the best solution to the following, simpler, equation:

$$B(t) = c_0 + \sum_{i=1}^{n} c_i \cdot A_i(t - k_i) + \epsilon. \tag{7}$$

### 2.2 Applications Scenarios

Multivariate time series emerge in many real contexts. Consider, for example, the medical context. Each patient can be described, during the observation period, by collecting all relevant numerical values of his/her indicators: blood pressure, temperature, body weight, amount of all drugs that are administered to him/her, and so on. In this way, a patient becomes a multivariate time series. Now, if we identify one particular variable of interest (e.g., the temperature), we can approach the problem of explaining its behaviour using the values of the other variables, as in (5). Intuitively, however, changes in values (such as the amount of a certain drug that it is administered) may have a delayed effect on the temperature; thus, it is possible that the behaviour of the temperature is, in actuality, better explained by an instance of (6).

As a different example, consider an environmental study scenario. In it, we have a number of observation points, let us say underground water wells, from which, at given times, water samples are extracted. Each sample is analyzed from the chemical-physical point of view, and the amount of interesting elements is registered. Since each observation point is sampled many times during the observation period, it may be seen as a multivariate time series. As before, one particular characteristics of the samples may be of interest, for example the amount of some pollutant, and we may want to search, if it exists, for the mathematical relationship that links the amount of pollutant to the amount of the other values of each sample, possibly towards a geological explanation of its presence. In some cases, the presence of chemical elements in the water has a delayed effect on the concentration of pollutant(s), so that such a mathematical relation may be modelled by an instance of (7).

## 3 An Optimization Model for Lag Regression

A *multi-objective optimization problem* (see, e.g. [4]) can be formally defined as the optimization problem of simultaneously minimizing (or maximizing) a set of $k$ arbitrary functions:

$$\begin{cases} \min / \max \ f_1(\bar{x}) \\ \min / \max \ f_2(\bar{x}) \\ \dots \\ \min / \max \ f_k(\bar{x}), \end{cases} \tag{8}$$

where $\bar{x}$ is a vector of decision variables. A multi-objective optimization problem can be *continuous* or *discrete* (*combinatorial*). In combinatorial problems, we look for objects from a countably (in)finite set, typically integers, permutations, or graphs. Maximization and minimization problems can be reduced to each other, so that it is sufficient to consider one type only. A set $\mathcal{F}$ of solutions is *non dominated* (or *Pareto optimal*) if and only if for each $\bar{x} \in \mathcal{F}$, there exists no $\bar{y} \in \mathcal{F}$ such that *(i)* there exists $i$ $(1 \leq i \leq n)$ that $f_i(\bar{y})$ improves $f_i(\bar{x})$, and *(ii)* for every $j$, $(1 \leq j \leq n, j \neq i)$, $f_j(\bar{x})$ does not improve $f_i(\bar{y})$. In other words, a solution $\bar{x}$ *dominates* a solution $\bar{y}$ if and only if $\bar{x}$ is better than $\bar{y}$ in at least one objective, and it is not worse than $\bar{y}$ in the remaining objectives. We say that $\bar{x}$ is *non-dominated* if and only if there is not other solution that dominates it. The set of non dominated solutions from $\mathcal{F}$ is called *Pareto front*.

Consider, as before, a multi-variate time series $A_1(t), \ldots, A_n(t), B(t)$ with $m$ distinct observations, and a vector $\bar{x} = (x_1, \ldots, x_n)$ of decision variables with domain $[0, \ldots, m]$. Let $M$ be the maximum of $\bar{x}$ (called *maximum lag* of $\bar{x}$). The vector $\bar{x}$ entails a lag transformation of (4) into a new data set with $m - M$ observations, in which the feature (time series) $A_i$ is lagged (i.e., delayed) of the amount $x_i$:

$$A(\bar{x}) = \begin{bmatrix} t_M & a_{(M-x_1)1} & a_{(M-x_2)2} & \cdots & a_{(M-x_n)n} & b_M \\ t_{M+1} & a_{(M+1)-x_1)1} & a_{((M+1)-x_1)2} & \cdots & a_{((M+1)-x_1)n} & b_{M+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ t_{m-M} & a_{((m-M)-x_1)1} & a_{((m-M)-x_1)2} & \cdots & a_{((m-M)-x_1)n} & b_{m-M} \end{bmatrix} \tag{9}$$

The resulting data set can be used to train a classical linear regression algorithm, with the effect of learning a model as in (6). This model can be used to explain the time series $B(t)$; a more complex mechanism would be required to optimize the coefficients in order to perform forecasting, also. Let $f_1(\bar{x})$ in (8) be any performance measure of the learned model after the transformation $\bar{x}$; depending on the particular application, we can instantiate $f_2, f_3, \ldots$ as necessary, in order to optimize not only the performance of the model but also any other characteristics. For example, we can slightly improve our original formulation by allowing each $x_i$ to take values in $[-1, 0, \ldots, m]$, and interpret $x_i = -1$ as discarding completely the $i$-th column (so to embed a feature selection mechanism). In this case we can instantiate $f_2()$ as:

$$CARD(\bar{x}) = \sum_{i=1}^{n} \begin{cases} 0 \text{ if } x_i \neq -1 \\ 1 \text{ otherwise} \end{cases} \tag{10}$$

In this case (8) becomes:

$$\begin{cases} \min / \max \ f_1(\bar{x}) \\ \min \ CARD(\bar{x}) \end{cases} \tag{11}$$

It is worth observing that (11) could be improved by letting $\bar{x}$ spanning over $B$ as well (in that case, the value $-1$ would be forbidden in $B$, obviously). Solving

the problem in this version, would entail searching for a linear equation similar to (6), but with the addition of terms of the type $B(t-k)$ $(k \geq 1)$, in the spirit of auto regressive models. The main drawback of such a choice is the reduced interpretability of the resulting explanation model, which would include past values of the independent variable as part of the explanation of the current one. For this reason, in this first proposal we did not include this feature.

## 4 Implementation and Test

### 4.1 Evolutionary Algorithms

*Multi-objective evolutionary algorithms* are known to be particularly suitable to perform multi-objective optimization, as they search for multiple optimal solutions in parallel. In this experiment, in order to solve (11) we have chosen the well-known NSGA-II (Non-dominated Sorted Genetic Algorithm) [5] algorithm, which is available as open-source from the suite *jMetal* [6]. NGSA-II is an elitist Pareto-based multi-objective evolutionary algorithm that employs a strategy with a binary tournament selection and a rank-crowding better function, where the rank of an individual in a population is the non-domination level of the individual in the whole population. As black box linear regression algorithm, we used the class *linearRegression* from the open-source learning suite *Weka* [15], run in 5-fold *cross-validation* mode, with standard parameters and no embedded feature selection. We have represented each individual solution $\bar{x}$ as an array:

$$x_1, x_2, \ldots, x_n$$

with values in $[-1, \ldots, m]$, where $m$ is the number of observations of the data set. As performance measure for the underlying linear regression algorithm we used:

$$f_1(\bar{x}) = 1 - |CORR(\bar{x}, \bar{y}, \bar{z})|$$

where $CORR$ measures the correlation between the stochastic variable obtained by the observations and the linear variable obtained by *linearRegression* on the data set after the transformation indicated by $\bar{x}$, as explained in the previous section. The correlation varies between $-1$ (perfect negative correlation) to 1 (perfect positive correlation), being 0 the value that represents no correlation at all. Thus, we have designed the evolutionary computation to optimize the correlation only. We have used the standard mutation and crossover operations (suitably adapted to correctly deal with our solution representation), with probabilities (tuned with an initial experiment) of 0.3 and 0.7, respectively. Our population is composed by 100 individuals; we have set the algorithm for a total of 1000 evaluations in a single execution, and launched 5 independent executions.

**Table 1.** Features in the original data set.

| **feature** | **description** |
|---|---|
| air_temp | hourly recording of the air temperature |
| solar_rad | hourly amount of solar irradiation |
| wind_speed | hourly recording of the wind speed |
| rel_humidity | hourly recording of the relative air humidity |
| air_pressure | hourly recording of the air pressure |
| traffic | hourly sum of vehicle numbers at the considered intersection |
| NO2_conc | hourly recording of the $NO_2$ concentration level |
| NOX_conc | hourly recording of the $NO_X$ concentration level |
| PM25_conc | hourly recording of the $PM_{2.5}$ concentration level |

**Table 2.** Results of the experiment.

| **correlation coefficient** | | | | **lags** |
|---|---|---|---|---|
| **original (c.v.)** | **lagged (c.v)** | **optim. (c.v)** | **optim. (test)** | |
| 0.6250 | 0.7749 | 0.7180 | 0.7378 | 14,7,2,10,0,0 |
| 0.6251 | 0.7752 | 0.7184 | 0.7382 | 14,0,2,8,10,0 |
| 0.6250 | 0.7752 | 0.7187 | 0.7297 | 21,5,2,9,23,0 |
| 0.6251 | 0.7748 | 0.7208 | 0.7363 | 20,0,2,7,19,0 |
| 0.6252 | 0.7750 | 0.7039 | 0.7243 | 21,0,3,8,7,0 |

### 4.2 Data Origin and Preparation

The first environmental study that relates air pollution and meteorological variables and traffic conditions in Wrocław (Poland) is presented in [10]. The overall goal of the study was determining how the levels of specific pollutants, namely, $NO_2$, $NO_X$, and $PM_{2.5}$, are related to the values of other attributes, such as weather conditions and traffic intensity, with the purpose of building an explanation model. In it, the value of a pollutant at a certain time instant is linked to the value of the predictor attributes from the same time instant. In [10], a non-linear, non-interpretable, atemporal model has been used; the fitting ability of a non-interpretable model compensates, partially, for not using the historical values of the predictor, giving rise to a relatively good explanation model. The considered data set spans over the years 2015–2017, and it records information at one-hour granularity. The structure of reduced data set, obtained from the original one after eliminating the explicit temporal attributes (by interpreting data as a time series, the notion of time becomes implicit), and the categorical ones, can be seen in Tab. 1. The attribute *traffic* refers to the number of vehicle crossings recorded at a large intersection equipped with a traffic flow measurement system. The air quality information has been recorded by a nearby measurement station.

In this experiment we considered only one pollutant, namely $NO_2$, and we interpreted the data as a multivariate time series. For efficiency reasons, we

**Table 3.** Coefficients of the linear functions (best individuals).

| **air_temp** | **solar_rad** | **wind_speed** | **rel_humidity** | **air_pressure** | **traffic** |
|---|---|---|---|---|---|
| -0.4037 | 12.6468 | -4.5834 | -0.3840 | -0.0844 | 0.0083 |
| -0.2362 | -6.3024 | -4.5848 | -0.4956 | 0.0615 | 0.0089 |
| -0.3769 | 9.9758 | -4.5378 | -0.4206 | 0.067 | 0.0085 |
| -0.1853 | -6.3731 | -4.6274 | -0.4733 | 0.1049 | 0.0087 |
| -0.2011 | -7.0963 | -4.2235 | -0.5036 | 0.0502 | 0.0093 |

considered only the 10% of the entire data set, and we have split it into a training and test set, operating the optimization on the former one only. We have used the training set in all 5 executions of the optimization model (11), and selected the best (in terms of correlation) element from each final population. Our maximum lag allowed in the optimization model is 24 hours.

### 4.3 Results

The first reference result is the correlation that can be obtained by training a *linearRegression* model on the training data with standard parameters and no embedded feature selection, and executing on the test data: 0.6652. Also, we consider the correlation coefficients on the training data only, in 5 experiments, varying the seed (1 to 5), in 5-folds cross validation, again in the original configuration, as shown in Tab. 2, leftmost column. Even if our data are temporal, learning (as base reference) an atemporal model (such as (5)) makes sense in some problems. Indeed it may be the case that the delayed effect of the independent variables on the dependent one falls below the temporal granularity of the data (for us, one hour), and that, at the same time, the dependent variable presents a quasi-constant behaviour in such a small interval. Should that be the case, a model such as (5) would have a relatively high performance (that is, it would be an acceptable approximation of the physical reality); in our case this is not true, which justifies the resort to temporal lag regression.

The second reference results emerges from creating a lagged version of the data set in the standard way, using WEKA's *TSLagMaker*. Because of the dimensions of the problem, we created a lagged version of each variable (excluding the class) up to 12 hours only, for a total of 79 attributes. The correlation coefficient that resulted from training a *linearRegression* model on the lagged version of the training set, and executing it on the lagged version of the test set is 0.8066, which is quite high. Unfortunately, observing the resulting model, the interpretation limits of this technique emerge clearly. For example, the resulting function shows a positive factor for the temperature at the same time, 4,5,6,7,8, and 10 hours before the observation, but negative for the temperature 1,2,3,9,11, and 12 hours before the observation. A similar behaviour is shown in almost all other variables. This makes it very hard to identify, if it exists, a cause-effect phenomenon, on top of the fact that the expert should be able to interpret a 79-variables linear function to extract a meaningful environmental model. An

intermediate step of feature selection does not solve the interpretation problem; as a matter of fact, the effect of feature selection is that of selecting the best features (with an absolute measure, in the case of filters, and relatively to a learning task, in the case of wrappers), and, again, selecting, for example, the temperature at 4, 6, and 10 hours before the observation would make it very hard to construct a concrete explanation model. The results in cross-validation of the training data only, in the same conditions as before, are shown in Tab. 2, second column.

In Tab. 2, third column we can see the correlation coefficient of the best individual for each of the five execution, in cross-validation mode (that is, on training data only). Compared with those in the first column, it is possible to appreciate an improvement of about 8 points, in average. When each best individual is executed on the test set, we obtain the results shown in the fourth column, which again, compared with the original training-test experiment, show an average improvement of about 7 points. The loss in correlation coefficient of these individuals with respect to the extended (lagged) version of the data set is compensated by the intrinsic greater interpretability of the former over the latter.

### 4.4 Discussion

Observe, first, the coefficients of the linear functions that correspond to each individual (Tab. 3): as for five variables out of six, the coefficients present the same sign and a very similar module across the individuals (this is an indication that our proposed models are stable), and when both positive and negative coefficient appear (that is, in the variable that measures the solar radiation), the change coincides with a change in the amount of lag, maybe indicating two different physical processes.

Let us focus now on the chosen lags in each individual. Observe, to start with, that the variable that measures the hourly traffic has always lag zero: in other words, all models coincide that the amount of $NO_2$ is influenced by the amount of (car) traffic with no delay. This could be explained by the small distance between the point of pollution concentration measurement and the intersection where the main emission source (the cars) is located. Similarly, four out of five models agree that the speed of the wind influences the amount of pollutant with two hours of delay. This may be due to the distance between the meteorological station and the intersection. In Wrocław, North-West winds prevails; therefore, the wind generally blows from the meteorological station towards the intersection. The distance, in a straight line, is about 10km and the average wind speed is 3m/s. Taking into account the porosity of the city development area and the time needed to evacuate pollution from the built-up area around the intersection, a delay of about 2 hours in the reaction of pollution concentration to the measured wind speed is reasonable. Moreover, observe that in three out of five models the lag for the solar radiation is zero, with negative coefficient in the corresponding equation, while the remaining two is between 5 and 7, with positive

coefficient. This opens the possibility of two different explanation models: for negative correlation (increasing solar radiation corresponding to a decrease in $NO_2$ with no delay), the physical process may be related to an intensification of photochemical reactions, while positive correlations take place with 5 to 7 hours of delay, and may indicate the reverse process.

In conclusion, our learning model produces individuals that are easier to interpret, because they identify the most relevant delays for the explanation task, so that devising a meaningful environmental model becomes possible.

## 5 Conclusions

In this paper we have proposed, and tested, a novel optimization model for temporal lag regression. Lag variables can be very important for the task of single multivariate time series explanation and prediction, as they allow a model to take into account possible delays in the effect of an independent variable on the dependent one. The standard approach for lag variable using consists of creating predetermined lagged artificial variables, and then using standard learning techniques on the obtained, extended data set; in a sense, this can be seen as a *brute force* approach. We proposed in this work an optimization model in which the amount of lag for each variable is decided dynamically, and we implemented it with a multi-objective evolutionary algorithm. Our learning model, that implicitly includes a feature selection mechanism, chooses the best lag for each variable, effectively providing a more interpretable, yet accurate enough, explanation model for a multivariate time series. Our schema, with minimal adaptations, can be used for multivariate time series forecasting as well. We tested our model on real data taken from a larger project that aims to construct an explanation model for the study of atmospheric pollution in the city of Wrocław (Poland).

Our model can be improved in several ways. In certain applications, for example, the same independent variable can influence the dependent one with a prolonged delay that spans more than one observation. A possible generalization, therefore, would aim to optimize the number of consecutive observations to take into account, and their algebraic combinations.

## Acknowledgments

Estrella Lucena-Sánchez and Guido Sciavicco acknowledge the partial support by the project *Artificial Intelligence for Improving the Exploitation of Water and Food Resources*, founded by the University of Ferrara. Estrella Lucena-Sánchez acknowledges, moreover, the support by the Emilia-Romagna (Italy) regional project *New Mathematical and Computer Science Methods for the Water and Food Resources Exploitation Optimization*. Finally, this work was also partially funded by the Spanish Ministry of Science, Innovation and Universities under the SITUS project (Ref: RTI2018-094832-B-I00), and by the European Fund for Regional Development (EFRD, FEDER).

## References


1. Al-Douri, Y., AL-Chalabi, H., Lundberg, J.: Time series forecasting using a two-level multi-objective genetic algorithm: A case study of maintenance cost data for tunnel fans. Algorithms **11**, 1–19 (2018)
2. Bijan, P.: Some applications of nonlinear regression models in forestry research. The Forestry Chronicle **59**(5), 244–248 (1983)
3. Box, G., Jenkins, G., Reinsel, G., Ljung, G.: Time Series Analysis: Forecasting and Control. Wiley (2016)
4. Collette, Y., Siarry, P.: Multiobjective Optimization: Principles and Case Studies. Springer Berlin Heidelberg (2004)
5. Deb, K.: Multi-objective optimization using evolutionary algorithms. Wiley, London, UK (2001)
6. Durillo, J., Nebro, A.: jMetal: a Java framework for multi-objective optimization. Avances in Engineering Software **42**, 760 – 771 (2011)
7. Hall, M.: Time series analysis and forecasting with WEKA (2014), *https://wiki.pentaho.com*, last accessed: May, 2019
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735—-1780 (1997)
9. Holt, C.: Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting **20**(1), 5–10 (2004)
10. Kamińska, J.: The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. Journal of environmental management **217**, 164–174 (2018)
11. Poulos, L., Kvanli, A., Pavur, R.: A comparison of the accuracy of the box-jenkins method with that of automated forecasting methods. International Journal of Forecasting **3**, 261–267 (1987)
12. Salimi, A., Rostami, J., Moormann, C., Delisio, A.: Application of non-linear regression analysis and artificial intelligence algorithms for performance prediction of hard rock tbms. Tunnelling and Underground Space Technology **58**, 236 – 246 (2016)
13. S.V Archontoulis, S., Miguez, F.: Nonlinear regression models and applications in agricultural research. Agronomy Journal **107**(2), 786 – 798 (2013)
14. Winters, P.: Forecasting sales by exponentially weighted moving averages. Management Science **3**(6), 324–342 (1960)
15. Witten, I., Frank, E., Hall, M.: Data mining: practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, Elsevier (2011)

# Stochastic dimension reduction techniques for time-point forecasting data

Shrikant Pawar[1, 2]* & Aditya Stanam[3*]

[1]Department of Computer Science, Georgia State University, 34 Peachtree Street, 30303, Atlanta, GA, USA.
[2]Department of Biology, Georgia State University, 34 Peachtree Street, 30303, Atlanta, GA, USA.
[3]College of Public Health, The University of Iowa, UI Research Park, #219 IREH, 52242-5000, Iowa City, Iowa, USA.

*Contributed equally
Correspondence: spawar2@gsu.edu

**Abstract:**

Dimension reduction techniques are essential for forecasting datasets. The reduction can be achieved through vector processes, probabilistic approaches to modeling macroeconomic uncertainties, nonstationarity, model integration, forecasting theory and adjustment, rnsemble forecasting, forecasting performance evaluation, interval forecasting, data decomposition etc. Using stochastic modeling techniques, we can retrieve accurate probabilities about a given system or event. This article compares five ("direct", "multitaper", "mvspec", "pgram"; and "wosa") different reduction techniques for stochastic distributed forecast data, while there are other parametric dimension reduction techniques, "direct" and "mvspec" were found to be the most effective reduction techniques (lowest entropies) for M4 Forecasting Competition dataset.

**Keywords:** Forecasting, Dimension Reduction, Principal Component Analysis

**Introduction:**

Dimension reduction techniques are essential for forecasting datasets [1, 2]. Application of forecasting can vary from nonparametric and functional methods to atmospheric science, telecommunication, hydrological, traffic, tourism, marketing, modelling and forecasting in power markets, energy, climate, financial forecasting and risk analysis, forecasting electricity load and prices and forecasting and planning systems [3-11]. The reduction can be achieved through vector processes, probabilistic approaches to modeling macroeconomic uncertainties, nonstationarity, model integration, forecasting theory and adjustment, rnsemble forecasting, forecasting performance evaluation, interval forecasting, data decomposition, seasonal adjustment, singular spectrum analysis and detrending methods [12-14]. Further, recent advances in machine learning considering adaptivity for stochastic models, on-line machine learning for forecasting, aggregation of predictors, hierarchical forecasting, computational intelligence and integration of system dynamics and forecasting models can be implied on the reduced data. The amount of data is increasing and there seems a difficulty of handling these data. One of the problems most frequently faced is the difficult to detect information among such large amounts of data [15-20]. Knowledge discovery in databases (KDD) is one of the databases targeted to overcome the problem of data mining and processing the extracted information. Stochastic modeling or Markovian models to predict states in complex scenarios have been efficiently utilized to address these problems. Using stochastic modeling techniques, we can retrieve accurate probabilities about a given system or

event. However, the "state space explosion" makes it hard to scale models to real application size problems. One interesting technique which has proven itself to be efficient on different datasets is dimensionality reduction, which will be tested in this paper for different time series forecasting data points. A univariate time series yt=(y1,…,yT), yt=(y1,…,yT), can be turned into a multivariate time series by embedding its lagged (p+1)(p+1) dimensional feature space as Xt=(yt,yt−1,…,yt−p)Xt=(yt,yt−1,…,yt−p). This is a common technique in non-linear time series analysis. One important aspect is that this transformation requires the multivariate spectrum of a KK-dimensional time series with TT observations, which is stored in a T×K×KT×K×K array and a symmetry/Hermitian property can be used to half the size of this array. In the current analysis we take the multi-dimensional time series of M3-COMPETITION - 3003 series data with 6 categories (MICRO, INDUSTRY, MACRO, FINANCE, DEMOGRAPHIC, and OTHERS) and try to find a 6-dimensional subspace that has interesting patterns that can be easily forecasted.

**Materials and Methods:**

**i. Data Collection:** The M4 Forecasting Competition is the next step in the evolution of the Makridakis or M-Competitions, which aims to identify the most accurate forecasting method for different types of predictions. This competition is organized by the Institute For the Future (IFF) at the University of Nicosia (UNIC), with the support of the Forecasting & Strategy Unit at the National Technical University of Athens (NTUA). The observations are divided into 6 different categories, MICRO, INDUSTRY, MACRO, FINANCE, DEMOGRAPHIC, and OTHERS. All the analysis code repositories and raw datasets are deposited on authors GitHub account which can be found at: https://github.com/spawar2/Forcasting-pipelines

**ii. Reduction Analysis:** Dimension reduction techniques for multivariate time series Xt can be applied with Forecastable Component Analysis (ForeCA) in R [1]. It finds a linar combination yt = Xtv that is easy to forecast. The measure of forecastability Ω(yt) (Omega) is based on the entropy of the spectral density fy(λ) of yt, higher entropy means less forecastable, lower entropy is more forecastable. The main function foreca runs on a multivariate time series Xt. Dimension reduction can be performed on Xt – a K-dimensional time series with T observations. Other parameters like foreca.one_weightvector is a wrapper around several algorithms that solve the optimization problem for a single weightvector wi and whitened time series Ut while foreca.multiple_weightvectors applies foreca.one_weightvector iteratively to Ut in order to obtain

multiple weightvectors that yield most forecastable, uncorrelated signals. There are several techniques of estimation, for our analysis we have compared "direct", raw periodogram; "multitaper", tapering the periodogram; "mvspec" smoothing estimate using mvspec; "pgram"; and "wosa" which is Welch overlapping segment averaging (WOSA) technique.

**Results and Discussion:**

Figure 1 shows the distribution of all the observations of the forecast data for all the 5 categories for years 1811-1975. There is a random stochastic distribution for our query data. The biplot shows that for the first component all points are in the same direction, stating it to be the overall/average pattern (Figure 2A-E). The barplots on the right show how the forecastable components (ForeCs) have indeed decreasing forecastability, and the first component is more forecastable than the original series (Figure 2A-E). The first component is more forecastable than the original series, it is less noisy. The remaining series also show very interesting patterns that are not visible in the original series. We found that all ForeCs are orthogonal to each other, i.e., they are uncorrelated (Figure 3A-E). Amongst, "direct", "multitaper", "mvspec", "pgram"; and "wosa", "direct" and "mvspec" had lowest entropies Ω(yt) (Omega) (Figure 4A-E). This article compares five different reduction techniques for stochastic distributed forecast data, while there are other parametric dimension reduction techniques, "direct" and "mvspec" are most effective for M4 Forecasting Competition dataset. Further validations of our observations with detrending methods would be necessary for conclusive results.

**Acknowledgments**

No external funding was utilized for the analysis of this paper.

**Author contributions**

AS and SP contributed to the conception and design as well as the drafting of the manuscript. All authors read and approved the final paper.

**Disclosure**

The author reports no conflicts of interest in this work.

**Supplementary files**

1. Dimension reductionality.xls: Raw M4 Forecasting Competition dataset.

**Figures**

**Figure 1 shows the distribution of all the observations of the forecast data for all the 5 categories for years 1811-1975.**

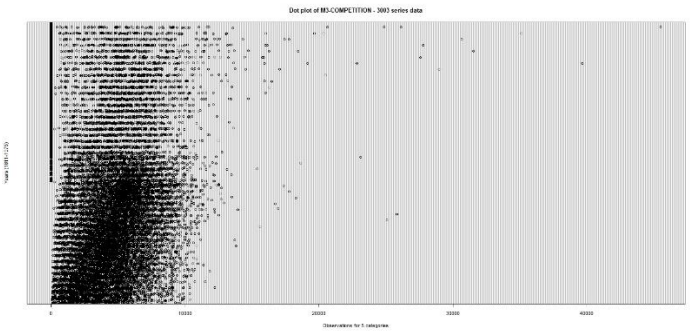



**Figure 2 shows the distribution of forcastability and noise measured in terms of P value for "direct", "multitaper", "mvspec", "pgram"; and "wosa" (A-E).**

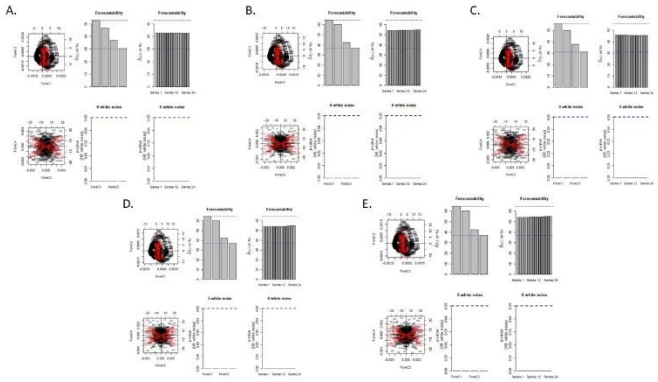



**Figure 3 shows the forcast scores for 4 components with "direct", "multitaper", "mvspec", "pgram"; and "wosa" techniques (A-E).**

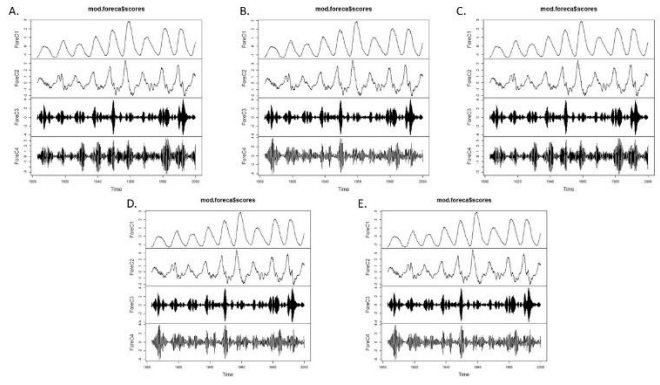
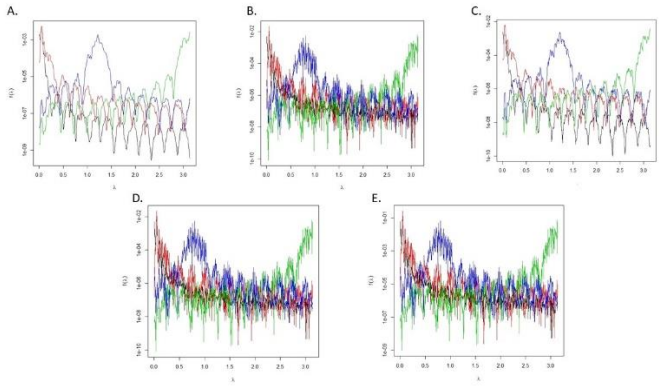


**Figure 4 shows the distribution of entropies Ω(yt) (Omega) for "direct", "multitaper", "mvspec", "pgram"; and "wosa" (A-E).**



**References**


1. Goerg, G. M. (2013). "Forecastable Component Analysis". Journal of Machine Learning Research (JMLR) W&CP 28 (2): 64-72, 2013. Available at jmlr.org/proceedings/papers/v28/ goerg13.html.
2. J. BERMAN, Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information, Elsevier Science, 2013.
3. Pawar, S, et al. Statistical analysis of microarray gene expression data from a mouse model of toxoplasmosis. BMC Bioinform. 12(Suppl. 7), SA19 (2011)
4. Ashraf M, et al. (2018) A side-effect free method for identifying cancer drug targets. Sci Rep.
5. G. E. BOX, et al. Time series analysis: forecasting and control, Wiley. com, 2013.
6. L. BRENNER, et al. Modelling Grid5000 point availability with SAN, Electronic Notes in Theoretical Computer Science (ENTCS), 232 (2009), pp. 165–178.

7. L. BRENNER, et al. Stochastic Automata Networks Software Tool, in Proceedings of the 4th International Conference on Quantitative Evaluation of SysTems (QEST 2007), Edinburgh, UK, S
8. L. BRENNER, et al. The Need for and the Advantages of Generalized Tensor Algebra for Kronecker Structured Representations, International Journal of Simulation: Systems, Science & Technology (IJSIM), 6 (2005), pp. 52–60.
9. K. CHAKRABARTI, et al. Locally adaptive dimensionality reduction for indexing large time series databases, ACM Trans. Database Syst., 27 (2002), pp. 188–228.
10. Lahiri, C, et al. Interactome analyses of Salmonella pathogenicity islands reveal SicA indispensable for virulence. J. Theor. Biol. 363, 188–197 (2014)
11. Lahiri, C, et al. Identifying indispensable proteins of the type III secretion systems of Salmonella enterica serovar Typhimurium strain LT2. BMC Bioinform. 13(Suppl. 12), SA10 (2012)
12. C. CHATFIELD, Time-series forecasting, Chapman and Hall/CRC, 2002.
13. R. M. CZEKSTER, et al. Split: a flexible and efficient algorithm to vectordescriptor product, in International Conference on Performance Evaluation Methodologies and tools (ValueTools'07), vol. 321 of ACM International Conferences Proceedings Series, ACM Press, 2007, pp. 83–95.
14. Pawar, S, et al. In silico identification of the indispensable quorum sensing proteins of multidrug resistant Proteus mirabilis. Front. Cell. Infect. Micro-Biol. 8, 269 (2018)
15. R. M. CZEKSTER, et al. Efficient vector-descriptor product exploiting time-memory trade-offs, ACM SIGMETRICS Performance Evaluation Review, 39 (2011), pp. 2–9. doi: 10.1145/2160803.2160805.
16. DATAMARKET!, The open portal to thousands of datasets from leading global providers. http://datamarket.com/, 2013.
17. Pawar, S, et al. Computational identification of indispensable virulence proteins of Salmonella typhi CT18. Curr. Top. Salmonella Salmonellosis (2017)
18. Pawar S, et al. (2011) Statistical analysis of microarray gene expression data from a mouse model of toxoplasmosis. BMC Bioinform 12(Suppl 7):A19
19. C. ENGEL, Can the markov switching model forecast exchange rates?, Journal of International Economics, 36 (1994), pp. 151–165.

20. P. FERNANDES, et al. Analysis of exponential reliable production lines using kronecker descriptors, Int. Journal of Production Research, 51 (2013), pp. 2511–2528.

# Towards a Better Forecasting and Nowcasting of Tunisian Economic Growth: The Relevance of Sovereign Rating Data

Adel KARAA
Professor in Quantitative Methods at High Institute of Management of Tunis, Tunisia
Corresponding author, adelkaraa@yahoo.fr
Azza BEJAOUI
Assistant Professor in Finance at High Institute of Management of Tunis, Tunisia
bjaouiazza2@yahoo.fr

**Short Abstract**

In this paper, we propose a formal and unified statistical framework in order to help performing reliable nowcasts, short and long-term forecasts of the real GDP growth in the Tunisian context. To do so, we use a set of available monthly macro-financial data and takes into account the sovereign ratings assigned to Tunisia by the four biggest Credit Rating Agencies (i.e. R&I, Fitch, Moody's and Standard & Poor's) since 1994. These data are used to estimate an appropriate multivariate unobserved componants times series model. The empirical results clearly show that combining of macro-financial indicators and sovereign ratings data seems to produce valid and consistent latent factors which track well GDP growth realizations throughout the estimation period. The short and long-term factor-based forecasts of the real GDP growth obtained by applying the AutoRegessive Distributed lag (ARDL) bound testing approach to cointegration by Pesaran et al. (2001) are also more accurate than those produced by other classical benchmark forecasting models. This framework, which we consider as a first step in setting up a real-time monitoring system for the macro-financial situation in Tunisia, already allows us to highlight the main flash indicators capturing the dynamics of the Real GDP growth.

**Keywords:** Nowcasting; Forecasting; GDP growth; flash indicators; factors' valtidity; PLS method; Multivariate unobserved componants time series model; ARDL model; pro/counter-cyclical effects; Sovereign ratings data.

**Long Abstract**

Overall, the economies throughout the world have experienced periods of economic instability and severe upheaval in growth, historically as well as recently. In such circumstances, and even during the upswing periods, knowing more information on future real activity is helpful for different economic agents. For instance, policymakers who seek to know more about future economic conditions, can build better policies. Timely and accurate estimates of the state of the economy are therefore essential for policymakers to provide reliable and early analysis of the ongoing economic situation. As well, further information about the future real activity can help business entrepreneurs to plan better their business projects.

In this regard, and formally, performing a proper real-time forecasting exercise using macroeconomic data therefore seems to be greatly needed. Not surprisingly, a particular attention, is amply paid to Gross Domestic Product[1] (GDP) which widely analyzed indicator of the economy's

[1]The growing importance of GDP is increasingly attributed to its distinguishable features. Indeed, GDP is synthetic, statistical and economic indicator and determines as the output of gathering different frequencies data (monthly, quarterly and annual). In this regard, and from a functional perspective, GDP can be considered as a common factor without compiling it from factor model (Cuevas and Quilis, 2012).

state and seems to be the proxy of the whole business cycle. Many central banks and other international institutions are already focusing on designing different business cycles indicators based on GDP.

Nevertheless, various problems arise with using GDP. Obviously, it is well known that GDP is released at quarterly frequency and even with a delay of weeks or months and revised repeatedly (Modugno et al., 2016). This spawns an impediment to track very short-term economic evolution (monthly). To overcome these shortfalls, several composite indexes based on a set of representative monthly indicators are designed by researchers[2] either for studying and analyzing business cycles or to nowcast[3] GDP growth in real time. Such indicators, which are also known as hard indicators,can include data on industrial production, prices of goods, services, expenditures, unemployment and financial variables (e.g. term spread, stock returns) which can be useful predictors of economic activity since they carry information expectations of future economic developments.

In this respect, various sources of information are employed to assess the current state of the economy (Galli et al., 2017) or to design composite index. Fortunately, innovation in computer technology and especially cheap computer power make it possible to readily gather and store large dataset. The ever-growing amount of data raises new salient challenges for researchers to enhance the nowcasting performance of their models. One issue linked to richer dataset is the large size of the information set, i.e. the number of variables retained by researchers. For instance, what makes the nowcasting of GDP a very difficult task is the selection of relevant variable among larger vintages of real-time data. According to Jansen et al. (2016), the datasets employed in the nowcasting literature vary increasingly in size and can include more than 300 variables, while restricting the length of time series used can make over-parametrization a real issue.

Searching for indicators with particular relevance still remains a debated question in the literature. Not surprisingly, some recent studies tend to investigate the feasibility of using other available data to nowcast the real GDP, for example credit data (Ermisoglu et al., 2013), firm-level data (Fornaro, 2016) and payments system data (Galbraith and Tkacy, 2016). The so-called 'soft' information such as surveys can be also handy valuable data due to their timeliness. In this respect, a potential use of sovereign ratings can be encouraging track, in particular in the emerging economies for which the quality and reliability of data are questionable (Liu et al., 2012).The sovereign rating, that is primarily developed to provide a forward-looking estimate of country risk, can be also used to retrieve an accurate information about the state and evolution of the countries' economies.

By delving deeper into the literature on the sovereign rating determinants, we notice that the rating agencies base their rating globally on the following key factors: Macroeconomic indicators (e.g. GDP growth, inflation); public finance indicators (e.g. the government deficit, the current of public deficit); monetary and external indicators (e.g. the current of account balance, foreign reserves).

Already widely used in nowcasting, this intricate set of indicators is completed in the sovereign ratings context by a series of soft data such as government effectiveness, track record of default, domestic political risk, effectiveness, stability, predictability and transparency of policymaking, geopolitical and external security risk and debt payment culture (Ligeti and Szőrfi, 2016). As well, during the review process (i.e. the credit watch process), the CRAs are involved in collecting further information and monitoring the rated government (Hill and Faff, 2007). In this respect, dialogue would be between the CRAs analysts and key policymakers and senior representatives of country's different public institutions (e.g. central banks, finance ministry). The CRAs are uniquely at an informational advantage over the credit watch process and the private

[2]Noticeable examples are the indexes built by Chauvet (1998) for the US and Brazil economies, Camacho and Perez-Quiros (2009a, 2009b) for the Spanish and Eurozone economies. From methodological perspective, researchers constructed composite indexes based on sophisticated statistical methods of extracting a common latent dynamic factor from the coincident indicators (e.g. Stock and Watson, 1991, 2002; Geweke, 1977; Sargent and Sims, 1977; Cuevas and Quilis, 2012).

[3]Nowcasting, which is defined as the prediction of the present, the very near future and the very recent past, implies providing a projection of a variable of interest on the available information set (Banbura et al., 2013). Nowcasting is pioneered by Evans (2005) using a restricted number of time series and developed by Giannone et al. (2008) for a great number of series to yield real-time GDP predictions, accounting for the idea of relating high-frequency indicators to low-frequency GDP data and the idea of employing real-time within a single statistical framework (Lamprou, 2016).

information of the end of the credit watch through the rating change (Boot et al., 2006). So, there is now much to indicate that the sovereign ratings can provide information advantage given they carry information expectations of future economic and financial evolution.

From the foregoing, a potential use of sovereign ratings in the Tunisian context could be specially interesting. Indeed, the country has not ceased to be recursively downgrated since January 2011 (the date of Tunisian revolution) by the four international rating agencies (Standard & Poor's, Moody's, Fitch and Rating and Investment Information, Inc). Accordingly, Tunisia was downgraded from the investment grade rating with positive outlook in January 2011 (the date of Tunisian revolution) to the speculative category with a negative outlook at the end of 2013.

At the same time, in its 2017 annual report, the Tunisian Court of Auditors questioned the quality of the data published by the National Institute of Statistics, reproaching it of not adopting methodologies complying with international standards when calculating some indicators.

Additionally, the main ratings agencies recognize that Tunisian national statistical system has been drastically shaken since the revolution and recommend the setting-up of more reliable and credible statistical system based on technical and legislative reforms. Without meaningful indicators, policymakers do not obviously have much foresight. The lack of visibility on the political and economic levels hinders Tunisia to have access to international financing sources and seriously undermine the investors' confidence and the country's credibility in the eyes of the international financial institutions (IMF, the World Bank, the European Bank for Reconstruction and Development (EBRD) and the European Commission). The challenge for Tunisia is to deal with its economic transition and reforms that aim for heightening its long-term growth potential by establishing a better business environment to attract more foreign direct investment and private investments, such as public-private partnership laws, competition laws, bankruptcy laws, tax reforms. So, there is no doubt that restoring international investor confidence requires a lot of efforts from the Tunisian authorities to enhance the short-term outlook for the economy and implement major reforms.

Facing this difficult situation, forecasting of macro-financial series is of utmost importance for the economic policymakers and general public to gain greater visibility on the state and the evolution of the financial and economic conditions in Tunisia. This article is in this perspective and strives to help in setting up reliable nowcasts, short-term and long-term forecasts of the real GDP growth in Tunisia based on a bunch of available monthly data (financial, economic and trade), the set of sovereign ratings posted by the international rating agencies and the quarterly Gross Domestic Product (GDP). For the nowcasting purpose, we propose an appropriate dynamic multivariate unobserved components time series model which analyzes the target variable (GDP growth) and the remaining variables simultaneously. More precisely, the proposed model consists of two parts. The first one deals with the set of the monthly macro-financial indicators and involves the extraction of a latent common factor and the estimation of its dynamic properties by means of an autoregressive equation. This factor is extracted according to the Dynamic Factor Models approach (DFM) to sum up the information contained in the monthly indicators dataset. According to Stock and Watson (1989), Chauvet and Pigger (2008), this factor can be considered as a monthly composite business-cycle index which measures comovements between different indicators. The second part of the model is used to forecast the monthly GDP growth[4] (the target variable) based on a second dynamic latent factor calibrated accordingly with respect to GDP using a second autorregression equation which expresses this latter in terms of its own past realizations and those of the first factor. This latent factor-oriented GDP, capturing the part of monthly GDP growth which is impacted by economic activity, can have the potential to yield favorable nowcasting properties and starts laying the underpinnings of forecasting GDP growth. In our model, the idiosyncratic part of GDP growth is explicitly modeled and jointly treated with dynamic factors. All the parameters of the proposed model are estimated by the maximum likelihood method after putting the model in a space-state form. The common latent factors are estimated by means of the Kalman filter.

[4] The monthly GDP growth series is obtained by applying the method of augmenting data (data imputation process) using the Multiple Imputation by Chained Equations (MICE) method on quarterly GDP.

The overall performance of our suggested approach is apprehended by studying how useful the extracted factors are for the quarterly economic growth forecasting[5]. In this perspective, we opt for the AutoRegessive Distributed lag (ARDL) bound testing approach to cointegration by Pesaran et al. (2001) to examine the short-run and long-run relationships between the quarterly real GDP growth and quarterly counterparts of the dynamic common factors resulting from the model estimation. The forecasting performance of our ARDL model is compared to a univariate classic benchmark models.

In order to highlight the contribution of sovereign ratings data in improving the predictive ability of the model, the aforementioned factors are again designed by incorporating impulse functions which signal the arrival of new sovereign rating change and assess its incidence on economic activity growth. In our case, the arrival dates of the ratings are indicated by the set of breakpoints recorded at a numerical composite index which designed on the occasion to retrace the whole history of the sovereign rating of Tunisia. Formally, this index corresponds to the first common factor extracted by applying a NonLinear Principal Components Analysis to the series of ratings that have been granted to Tunisia by the four international rating agencies, Standard & Poor's, Moody's, Fitch and Rating & Investment Information, Inc, respectively.

The paper is organized as follows: Section 2 provides an overview of the evolution of Tunisia's sovereign rating and highlights its interdependence with the country's evolution of macro-financial conditions. Section 3 deals with the econometric approach, specifying the dynamic intreactions of a selected set of monthly macro-financial indicators, the sovereign rating change announcements and the real GDP growth. Section 4 presents the data description and the procedure of selecting variables. Section 5 reports the results and interpretation of estimated factors. Section 6 addresses the in-sample performance of the econometric model through the factors' convergent validity and historical reconstruction. Section 7 examines the short-term and long-term relationships between the real GDP growth and the latent factors as well as the factors' predictive validity. Section 8 provides a summary and discussion. Finally, a set of appendices describes the technical details of the model and some additional empirical results.

[5] This approach is often adopted by researchers due to the difficulties posed by the non-observability of extracted factors in assessing the quality of different estimates.

# How Well Does Economic Uncertainty Forecast Economic Activity?*

John Rogers
International Finance Division
Federal Reserve Board

Jiawen Xu
Shanghai University of
Finance and Economics

April 2019

*The views expressed here are solely our own and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

# How Well Does Economic Uncertainty Forecast Economic Activity?

**Abstract**

It is difficult to overstate the reach and influence of the literature on economic and economic policy uncertainty over the last decade (www.policyuncertainty.com). On-going research relates uncertainty to macroeconomic phenomena such as inflation and GDP growth, microeconomic issues concerning firm-level investment and export market entry and exit, and finance considerations like corporate strategy and equity returns. In this paper, we address one surprisingly under-researched topic: what is the forecasting performance of economic uncertainty measures? We examine both in-sample and out-of-sample forecasting, both real and financial outcome variables, sub-sample stability, and real-time considerations. Our measures of uncertainty include U.S. Economic Policy Uncertainty (Baker, Bloom, & Davis, 2016), U.S. macroeconomic uncertainty (Jurado, Ludvigson, & Ng, 2015), and U.S. financial uncertainty (Ludvigson, Ma, & Ng, in press). We begin by showing that there is substantial explanatory power both in-sample and out-of-sample over the 128 outcome variables in the McCracken and Ng (2016) data set. Next, we document that uncertainty sometimes has additional predictive content over the widely-used excess bond premium (Gilchrist & Zakrajšek, 2012) and the Chicago Fed's National Financial Conditions Index (NFCI). We then use quantile regressions to examine whether the explanatory power of the above measures varies over different parts of the GDP growth distribution. There is good reason to expect that uncertainty would forecast recessionary conditions better than expansions (Adrian, Boyarchenko, & Giannone, 2019). We find that both in sample and out of sample, the EPU and MU measures of uncertainty show strong predictive power, especially at lower quartiles, consistent with the prior. Finally, we construct a *real-time* measure of macroeconomic uncertainty, using vintages of data back to 1999. We compare the real-time uncertainty measure to the original measure that is ex-post in its construction, in terms of their forecasting performances both in sample and out of sample.

## 1 Introduction

Research on economic uncertainty over the last decade has been ubiquitous. As made plain from a glance at www.policyuncertainty.com, research on uncertainty is devoted to macroeconomic phenomenon such as inflation and GDP growth, microeconomic issues concerning firm-level investment and export market entry and exit, and finance topics such as corporate strategy and equity returns. New measures reflect uncertainty in the minds of consumers, traders, managers, and policymakers about possible futures, and cover events like terrorism, natural disasters, war and climate change. It is difficult to overstate the reach and influence of this literature. As of this writing, Google scholar citation counts for four prominent articles in this recent literature are approaching ten thousand, with (Bloom, 2009), (Baker et al., 2016), (Bloom, Bond, & Van Reenen, 2007), and (Bloom, Floetotto, Jaimovich, Saporta-Eksten, & Terry, 2018) registering 3099, 2559, 1118, and 944, respectively.

Surprisingly, little work has focused on the *forecasting* performance of the various measures of economic uncertainty.[1] We fill that gap in the literature in this paper. We consider both in-sample and out-of-sample forecasting, both real and financial outcome variables, and sub-sample stability. We also devote attention to real-time considerations, both in terms of how the uncertainty measures themselves are constructed and in terms of the series that are being forecast, such as GDP.

Our measures of uncertainty include the newspaper-based index of U.S. Economic Policy Uncertainty (EPU) from (Baker et al., 2016), U.S. macroeconomic uncertainty (MU) from (Jurado et al., 2015), and U.S. financial uncertainty (FU) from (Ludvigson et al., in press).[2] To answer the question "How well relative to what?", we benchmark the forecasting performance of the uncertainty measures by comparing it to the excess bond premium (EBP) (Gilchrist & Zakrajšek, 2012) and the Chicago Fed's National Financial Conditions Index (NFCI), which have been shown to have high predictive power over a range of key macroeconomic outcome variables.[3] None of these measures contains values that were, strictly speaking, available in real time. MU, FU, and EBP are all regression based. Their magnitudes, at each point it time, are resid-

[1] There are, of course, many papers identifying shocks to measures of uncertainty in VARs and estimating their transmission effects in-sample. In addition, papers like (Caldara, Fuentes-Albero, Gilchrist, & Zakrajsek, 2016), for example, examine the interaction between financial conditions and economic uncertainty, also using VARs to trace out the impact of these two types of shocks. These authors document the importance of uncertainty shocks and show that they have an especially negative economic impact in situations where they elicit a concomitant tightening of financial conditions. This type of estimation strategy is quite different from the forecasting exercises we perform.

[2] We also examine all sub-indexes of EPU: monetary policy, fiscal policy, taxes, government spending, health care, national security, entitlement programs, regulation, financial regulation, trade policy and sovereign debt (currency crises).

[3] The NFCI provides a comprehensive weekly update on U.S. financial conditions in money markets, debt and equity markets, and the traditional and shadow banking systems. The index is constructed to have an average value of zero and a standard deviation of one over a sample period extending back to 1971. Positive values of the NFCI have been historically associated with tighter-than-average financial conditions, while negative values have been historically associated with looser-than-average financial conditions.

uals derived through estimation using a full-sample-period data set. Furthermore, the data set includes many series that are themselves continuously revised. The NFCI, an index constructed from 46 weekly, 33 monthly, and 26 quarterly indicators, is also subject to revisions. The newspaper-based EPU measure is closest to a real-time series. We level the playing field by constructing real-time uncertainty measures, and compare the performance of these real-time measures to those typically analyzed, i.e., those constructed ex-post from revised time series data.

We examine the marginal explanatory power over a baseline forecast from a dynamic factor model of the type used extensively in the literature with success (Bai & Ng, 2002). We begin by casting a wide net, examining how well our (aggregate) uncertainty measures forecast each of the 128 variables in the updated (McCracken & Ng, 2016) data set. We show that there is substantial explanatory power, both in-sample and out-of-sample, based on comparison of the baseline dynamic factor model to the model augmented alternately with one of the uncertainty measures. Digging into the EPU sub-categories, we find that uncertainty concerning monetary policy, regulation, and financial regulation have similar in-sample performance as does the general EPU index. In out-of-sample forecasting, sub-categories such as monetary policy, regulation, financial regulation, and trade policy perform even better than overall EPU.

Following these initial explorations, we next examine whether uncertainty has any additional predictive content over the widely-used excess bond premium of Gilchrist and Zakrajsek (2012). We find that there is. Adding EPU to the regressions used by Gilchrist-Zakrajsek, we show that in the sub-period before the financial crisis (1985-2007), EPU has the expected sign and is statistically significant in regressions for employment, unemployment, industrial production, non-residential investment, and inventories, even controlling for EBP (which remains highly significant). The performance of EPU declines noticeably after 2008, however, while the predictive content of EBP and NFCI remains high.

Next we use quantile regressions to examine whether the forecasting performance of uncertainty measures varies over different parts of the GDP growth *distribution*. For example, does uncertainty forecast recessionary conditions better than expansions? There is good reason to expect that it might, in light of important recent work on "Growth at Risk". Adrian et. al. (2019) model the distribution of future U.S. GDP growth as a function of current financial and economic conditions. They show that the estimated lower quantiles exhibit strong variation as a function of current financial conditions, while the upper quantiles are stable over time. (Adrian, Grinberg, Liang, & Malik, 2018) extend this analysis to 11 advanced and 10 emerging market economies, and show that growth-at-risk (the lower 5th percentile of the distribution) is more responsive to financial conditions than the median or upper percentiles. We find both in sample and

out of sample that the EPU and MU measures of uncertainty show strong predictive power, especially at lower quintiles and at short horizons. The MU measure out-performs all competitors, including EBP and the NFCI measure emphasized by (Adrian et al., 2019) in this exercise.

In the next section, we describe the data used in the paper and the in-sample predictive exercises, and follow that with a description of the out-of-sample forecast tests. In section 3, we estimate the marginal predictive content of EPU and NFCI when added to the Gilchrist-Zakrajsek regressions. In section 4, we estimate quantile regressions, that allow us to compare predictability across the GDP growth distribution. In the final section we devote to comparison of the predictive content of the results above to those using uncertainty measures based on real-time vintage data.

## 2 Predictive Power of Uncertainty Indexes over a Large Macroeconomic Data Set

We begin with the "kitchen sink", examining the predictive power of our uncertainty measures over a large number of time series, specifically the 128 monthly macroeconomic and financial time series from the (updated) data set of McCracken and Ng (2016). We define "predictability" of a particular uncertainty measure as its marginal contribution to the dynamic factor model represented by equation (1):

### 2.1 In-sample predictive regression

$$y_{i,t+h} = \alpha_i + \phi_i^y(L)y_{i,t} + \beta_i\varphi^F(L)\hat{F}_t + \gamma_i' Z_t + \epsilon_{i,t+h}^y \tag{1}$$

where $y_{i,t}$ refers to the transformed variable of interest, one of the time series from the (McCracken & Ng, 2016) data set. Similarly, we transform $y_{i,t+h}$, the $h$-step ahead forecast, also according to the McCracken-Ng code.[4] The $\hat{F}_t$ are estimated factors from the dynamic factor model, with the number of factors selected using criteria proposed by (Bai & Ng, 2002). Our benchmark, workhorse dynamic factor model is a formidable one, as the literature has shown it to have great success in forecasting (Stock and Watson (2006) provide an early survey). The $Z_t$ term contains, alternately, one of the auxiliary measures: the economic policy uncertainty index (EPU) of (Baker et al., 2016); 12-month horizon macroeconomic uncertainty (MU) (Jurado et al., 2015); 12-month horizon financial uncertainty (FU) (Ludvigson et al., in press); the excess bond premium (EBP) (Gilchrist & Zakrajšek, 2012); and National Financial Conditions Index constructed by Federal Reserve Bank of Chicago (NFCI).

[4]This is an unbalanced monthly data set spanning 1959:1-2018:12. We apply specific transformations to the raw series before estimation and construct the factors according to the transformation code provided in the data file. For example, real personal income (RPI), the first variable in the monthly data set, is transformed by $\triangle \ln(x_t)$. $y_{t+h}$ is defined as $y_{t+h} = \frac{C}{h}(\ln(x_{t+h}) - \ln(x_t))$, with $C = 1200$ for monthly data and $C = 400$ for quarterly data. For details, see the data appendix of (McCracken & Ng, 2016).

The predictive regression (1) is estimated by OLS, with 4 lags of $y'_{i,t}s$ and 2 lags of $\hat{F}_t$.[5] The in-sample predictive content of the aforementioned uncertainty indexes are shown by the t-statistics of $\gamma_i$ computed using HAC standard errors. Table 1 summarizes the number of series with significant indexes for $h = 1, 3, 12$. The t-statistics for all of the 128 series are relegated to appendix Table A-1 to save space. Each column in table 1 reports the number of significant series for different time periods since uncertainty indexes have different time spans. In the first column, we use all data available for each index. Among them, EPU is the shortest series, which starts from 1985:1. Therefore, in the second column we use data from 1985:1 to 2018:12 for all five indexes. We also run sub-sample regressions using data pre and post 2008, the onset of the great financial crisis. As we can see from the first two columns, EPU has relatively less predictive power than other indexes, but it does improve as the horizon increases. MU and FU are quite stable across all horizons, but especially good at short horizons. EBP, on the other hand, performs well at the long horizon, while NFCI has an average but robust performance.

### 2.2 Out-of-sample forecasting

We design an out-of-sample forecasting exercise, in which we use data from 1985:1-1994:12 for in-sample estimation and model selection, and the rest of the data for out-of-sample forecast accuracy evaluation. We compute the $h$-step ahead mean squared forecast error (MSFE) for each model $j$ and series $i$.

$$MSFE^h_{i,j} = \frac{1}{T_2 - T_1 - h + 1} \sum_{t=T_1}^{T_2-h} (y_{i,t+h} - \overset{\wedge}{y}{}^{j}_{i,t+h|t})^2$$

where $\overset{\wedge}{y}{}^{j}_{i,t+h|t}$is the h-step ahead forecast of $y_{i,t}$ in model $j$ computed using the direct approach. Parameter estimation, factor estimation and model selection are fully recursive. The first simulated out of sample forecast is made in 1994:12. To construct this forecast, we use only data available from 1985:1. Thus regressions were run for t = 1985:1,...,1994:12$-h$, then the values of the regressors at t = 1994:12 were used to forecast $y_{1994:12+h}$. All parameters, factors, and so forth were then re-estimated, information criteria were recomputed, and models were selected using data from 1985:1 through 1995:1, and forecasts from these models were then computed for $y_{1995:1+h}$. The final simulated out of sample forecast is made in 2018:12$-h$ for $y_{2018:12}$.

[5] We always keep 4 lags of $y_{i,t}$'s in the regression and leave out those insignificant regressors in $F_t$ and its lag. We report t-statistics of $Z_t$ in the screened regression.

Forecast accuracy is evaluated via an out-of-sample $R^2$, which is computed as:

$$R^2 = 1 - \frac{\frac{1}{T_2-T_1-h+1}\sum_{t=T_1}^{T_2-h}(y_{i,t+h} - \overset{\wedge}{y}{}^{j}_{i,t+h|t})^2}{\frac{1}{T_2-T_1-h+1}\sum_{t=T_1}^{T_2-h}(y_{i,t+h} - \overset{\wedge}{y}{}^{0}_{i,t+h|t})^2}$$

where $\overset{\wedge}{y}{}^{0}_{i,t+h|t}$ is the h-step ahead forecast of $y_{i,t}$ using the factor-based benchmark model (2). The competing model $j$ is a nested model with additional uncertainty index $j$, $j \in$ {EPU, MU, FU, EBP, NFCI}. The out-of-sample $R^2$ can go negative if the benchmark model offers a better forecast than the competing model $j$.

$$y_{i,t+h} = \alpha_i + \phi_i^y(L)y_{i,t} + \beta_i\varphi^F(L)\hat{F}_t + \epsilon^y_{i,t+h} \tag{2}$$

We choose the same forecasting horizons as above for the in-sample predictive regressions $(h = 1, 3, 12)$. In Table 2, we report the number of series with positive out-of-sample $R^2$ (see appendix Table A-2 for the out-of-sample $R^2$ for the individual series).

Similar to Table 1, we also compute out-of-sample forecasts during time periods pre and post 2008 and summarize the number of positive out-of-sample $R^2$ in column 2 and 3. Column 1 is the result for out-of-sample forecasts from 1985:1-2018:12. As for short horizons, MU, EBP and NFCI have strong forecasting power. They perform better than the benchmark in nearly half of the 128 series. The factor-based or diffusion index forecasting model has been shown to perform quite well and hard to beat in many existing papers. As the forecasting horizon increases, EPU tends to perform better and can beat the benchmark in approximately 1/3 of the 128 series.

## 3 Marginal predictability of EPU and NFCI over EBP

In this section, we examine if there is marginal predictive power of EPU over EBP and NFCI. Specifically, we estimate the Gilchrist-Zakrajsek regressions (their regression 2, Table 6) with EPU and NFCI as additional regressors. We are motivated to include NFCI based on the strong evidence presented in (Adrian et al., 2019) of its predictive power, including over EBP.

The in-sample predictive regression is:

$$\bigtriangledown^h Y_{t+h} = \alpha + \sum_{i=1}^{p}\beta_i \bigtriangledown Y_{t-i} + \gamma_1 TS_t + \gamma_2 RFF_t + \gamma_3 \overset{\wedge}{S}{}^{GZ}_t + \gamma_4 EBP_t + \gamma_5 EPU_t + \gamma_6 NFCI_t + \epsilon_{t+h}$$

where $\bigtriangledown^h Y_{t+h} \equiv \frac{C}{h+1}\ln(\frac{Y_{t+h}}{Y_{t-1}})$, $h \geq 0$ is the forecast horizon. Here $TS_t$ denotes the "term spread"—that is, the slope of the Treasury yield curve, defined as the difference between the three-month constant-maturity

Treasury yield and the ten-year constant-maturity yield; $RFF_t$ denotes the real federal funds rate. The credit spread index is decomposed into two parts: a component that captures systematic movements in default risk of individual firms and a residual component—the excess bond premium, we denote $\overset{\wedge}{S}_t^{GZ}$ and $EBP_t$ respectively.

The full sample data is from 1973:1-2018:12. We use historical data from the policy uncertainty website for EPU to construct the index from 1973 to 2018. The complete results are in tables 3 and 4, in which we report both the estimated coefficients and t-statistics for six regressors (four are included in (Gilchrist & Zakrajšek, 2012), while EPU and NFCI are new). The $Y_t$ in monthly regressions are EMP, UER and IPM, representing private non-farm payroll employment; civilian unemployment rate; and index of manufacturing industrial production. From the first panel in table 3, we see that NFCI has marginal predictive power over EBP for all three series at all horizons $(h = 1, 3, 12)$, while EPU is not significant for all series and all horizons. In the first panel of table 4, we run regression (2) using quarterly data for GDP and its main components. In the table, C-D is PCE on durable goods; C-NDS is PCE on non-durable goods and services; I-RES is residential investments; I-NRS is business fixed investment in structures. The full sample is from 1973:Q1 to 2008:Q4, and forecast horizon is 4 steps. EBP is significant in GDP and I-NRS. EPU is significant in C-D and I-RES. NFCI is significant in GDP, C-D, I-RES and I-NRS.

## 4 GDP Growth and Uncertainty Indexes: a quantile regression perspective

We hypothesize that measures of uncertainty forecast changes in GDP better at lower ends of the GDP growth distribution than at higher ends. In this section, we estimate quantile regressions to assess the correlation and predictive content of uncertainty indexes with GDP growth at different quantiles. First, examine Figure 1, where we display the unconditional correlations between GDP growth at different quantiles with five uncertainty indexes. On the horizontal axis, we display the average annualized quarterly GDP growth rates at $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$; on the vertical axis, we show the mean value for each uncertainty index in those quarters when GDP growth is in that particular quantile. Consistent with our hypothesis, when GDP growth is low and even negative ($\tau = 0.1$), all uncertainty indexes are quite high, and conversely, when GDP growth is high, the uncertainty indexes are typically low. This negative relationship is monotonically so only EPU and MU.

Next, we further analyze whether uncertainty indexes can provide additional predictive power, both in-sample and out-of-sample, over factors estimated from a large quarterly macroeconomic dataset. In order

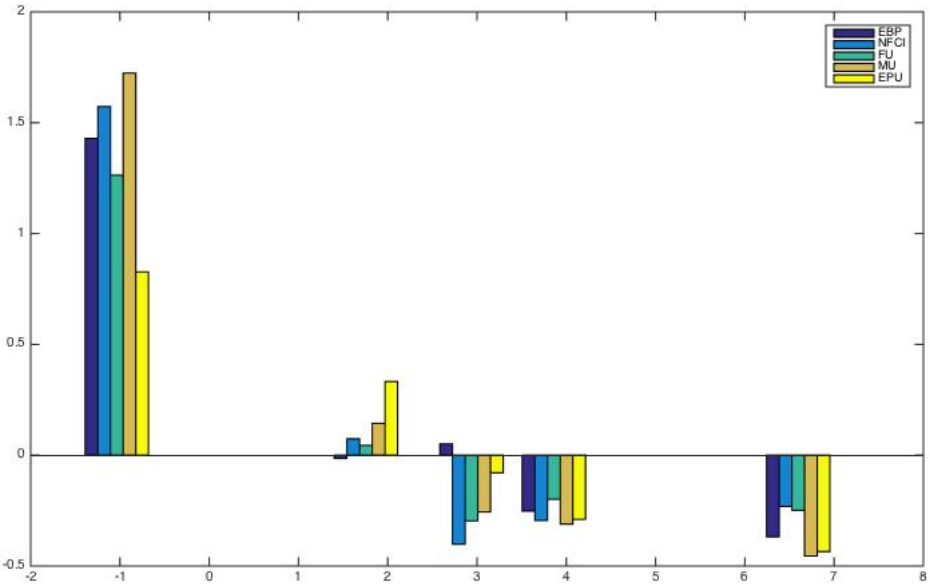



Figure 1: Uncertainty and GDP Growth, by GDP growth quintile

to do so, we run predictive quantile regression of $y_{t+h}$ on $x_t$, where $x_t$ is a vector containing a constant, current and lagged values of $y_t$, estimated factors $\hat{F}_t$, and uncertainty indexes. The quantile coefficients $\beta_\tau$ are chosen to minimize the quantile weighted absolute value of errors:

$$\hat{\beta}_\tau = \arg\min_{\beta_\tau \in R^k} \sum_{t=1}^{T-h} \left( \tau \cdot 1_{(y_{t+h} \geq x_t\beta)} |y_{t+h} - x_t\beta_\tau| + (1-\tau) \cdot 1_{(y_{t+h} < x_t\beta)} |y_{t+h} - x_t\beta_\tau| \right)$$

where $1_{(.)}$ denotes the indicator function. We use FRED-QD for factor estimation in this section.[6] There are in total 248 series, out of which 125 series are used for factor estimation. We exclude EPU from the dataset for factor estimation, and so use 124 series for factor estimation. The $\hat{F}_t$ are estimated using the complete unbalanced panel from 1959:I to 2018:IV. The in-sample quantile regressions are estimated from 1973:I to 2018:IV for all uncertainty indexes.

In Table 5, we report the quantile regression coefficients and t-statistics for each of the uncertainty indexes at $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ and for $h = 1, 4, 8$. In general, the above-stated hypothesis holds quite well, for all indexes especially at short (in-sample) forecast horizons. When $h = 1$, all uncertainty series are significantly and negatively related to 1-quarter ahead GDP growth rate at the lower quantiles; when $h = 4$, the coefficients on MU and EBP remain negative, significant, and largely with a downward effect at

[6]FRED-QD can also be downloaded at http://research.stlouisfed.org/econ/mccraken/. It is updated every quarter.

higher quintiles. When $h = 8$, however, only MU remains negatively related to GDP growth; it continues to exhibit the hypothesized stronger effect at lower quintiles. Curiously, EPU and NFCI become positively related to GDP growth.

Next, we use the same quantile predictive regression to calculate out-of-sample forecast values for GDP growth. We use 20 years of data beginning in 1973:I for in-sample estimation and forecast from 1993:I to 2018:IV. We compare the forecast accuracy evaluated via a quantile $R^2$ based on the loss function $\rho_\tau = x(\tau - I_{x<0})$,

$$R^2 = 1 - \frac{\frac{1}{T}\sum_t [\rho_\tau(y_{t+h} - \hat{\alpha} - \hat{\beta} X_t^j)]}{\frac{1}{T}\sum_t [\rho_\tau(y_{t+h} - \hat{\alpha} - \hat{\beta} X_t^B)]}$$

where $X_t^B$ denotes the vector containing the benchmark regressors, which include a constant, current and lagged values of $y_t$ and estimated factors $\hat{F}_t$. Here $X_t^j$ denotes the vector of the competing model $j$, which includes all the benchmark regressors and one additional uncertainty index. The out-of-sample quantile $R^2$ is reported in Table 6 for $\tau = 0.2, 0.4, 0.6$ and $h = 1, 4, 8$. From the full sample results, we see that out-of-sample $R^2$ is positive in 11 out of 45 cases (5 indexes$\times$3 quantiles$\times$3 horizons). As noted above, the benchmark model is a strong competitor. In general, forecast accuracy decreases as forecast horizon increases. Furthermore, in constrast to the in-sample results, there is no clear pattern about how accuracy changes over different quantiles of the growth distribution.

## 5 Summary of the Sub-sample Analysis

Macroeconomic time series cover a long time span and when it comes to forecast evaluation, it is usually crucial to consider time variation in parameters. This often leads to improved performance in sub-samples (see (Clements & Hendry, 1999) and (Hendry & Mizon, 2005)). Stock and Watson (2009) split the data into pre and post 1984 sub-samples and found substantial in-sample predictive fit improvements in sub-periods after 1984 (Stock & Watson, 2009). In this section, we investigate sub-sample results both before and after the 2008 beginning of the financial crisis. Most of the results are reported in the tables presented above.

In Table 1, EPU performs particularly well in the pre-2008 period but has less predictive content after 2008. The other four indexes perform better in the post-crisis period. The number of series with significant indexes even increases as the forecast horizon increases. When $h = 12$, MU, FU and NFCI are significant in over 80 out of 128 regressions. The out-of-sample results are mostly consistent with in-sample results: EPU

performs better before 2008 while FU and EBP perform afterward. The best performing indexes pre and post 2008, respectively, are EPU and EBP. EPU improves upon the benchmark in about 50 out of 128 series. EBP outperforms the benchmark in 67 out of 128 cases.

We report in the lower parts of Table 3 (Gilchrist & Zakrajšek, 2012) regressions for two sub-samples: 1985:1-2007:12 and 2008:1-2018:12. EPU has significant predictive power and largely displaces that of EBP and NFCI especially for $h = 1, 3$ during 1985:1-2007:12. We also replicate the GZ Table 7 (quarterly series) for sub-samples 1985:Q1-2007:Q4 and 2008:Q1-2018:Q4. EPU is statistically significant and of the correct sign for I-NRS in the pre-2008 sub-period, but overall does not appear to have much predictive content. The predictive power of NFCI decreases in this period. In the post-2008 crisis period, all three indexes lose their predictive power compared to the full sample.

We report sub-sample results of quantile regressions in tables 7 and 8. Table 7 shows in-sample quantile regression for GDP growth during 1973:I-2007:IV. In general, the results are quite similar to the full sample results. Slight differences exist at $h = 8$. EPU is positively related to GDP growth at quantiles lower than 0.5; EBP is positively related at the lowest quantile. MU and EBP at other quantiles are negatively related to GDP growth. Table 8 shows in-sample quantile regression for GDP growth during 2008:I-2018:IV. When $h = 1$, MU, EBP and NFCI are all significantly and negatively related to GDP growth; when $h = 4$, all five indexes have negative coefficients at all quantiles; when $h = 8$, MU, FU and EBP remain negative, but EPU and NFCI become significantly positively related to GDP growth especially at lower quantiles. For out-of-sample $R^2$, we have 14 and 17 out of 45 cases positive $R^2$ for the pre and post 2008 periods, respectively. Forecast accuracy is stable at $h = 1, 4$ and deteriorates as horizon increases to 8. For the pre-2008 period, forecast accuracy is better at lower quantiles. For the post-2008 period, accuracy is good at $\tau = 0.2, 0.6$. Overall, the performance of MU is the best: 11 out of 31 forecast gains are due to MU. Next best are EPU and NFCI, to which we attribute 7 and 6 out of 31 forecast gains.

## 6 Uncertainty in Real-time

As noted above, although none of the uncertainty measures contains values that are strictly speaking available in real time, EPU is by far the closest. Our analysis above indicates that MU has stronger predictive content than EPU. In this section, we level the playing field by constructing a real-time MU index and comparing its performance to the one analyzed above, i.e., that constructed ex-post from revised data.

We begin by using vintages of the McCracken-Ng data set beginning in 2004:01 and ending in 2019:01.

Since financial data are never revised, we just use financial dataset updated to 2018:12.[7]. All macro and financial series except for 'MZMSL', 'DTCOLNVHFNM', 'DTCTHFNM', 'INVEST' are used for factor estimation. Due to data availability, we construct a balanced panel starting from 1978:06, which includes 120 out of 132 macroeconomic series used in (Jurado et al., 2015). [8] We use the Matlab and R code posted on Serena Ng's website to construct the Macro Uncertainty index.[9] The estimation and construction procedure are repeated every month. We collect the last observation of each MU series, estimated vintage by vintage starting with 2004:01, to form a real time MU series. In Figure 2, we plot the 1-step, 3-step, and 12-step ahead real time MU (top panel) together with the ex-post MU updated in 2019:02 (bottom panel).The original MU series from JLN (2015) looks much smoother than real time MU.

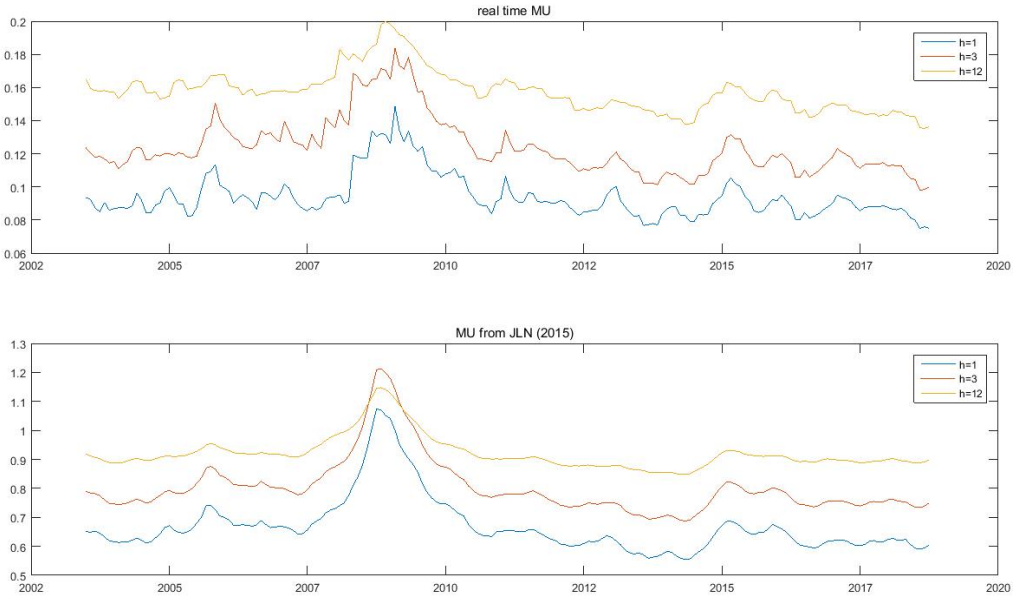



Figure 2: Real time MU v.s. MU in JLN (2015)

Next, we compare the predictive content of real-time and ex-post MU. In Table 9, we report comparison results of the in-sample predictive regressions. Clearly, the real time MU measure has less predictive power than its ex-post version in both full sample and sub-sample regressions. The pattern is similar to that

[7]Special thanks to Sai Ma for providing us the updated financial dataset used in (Ludvigson et al., in press)

[8]We exclude 'HWI', 'HWIURATIO', 'NAPMPI', 'NAPMEI', 'NAPM', 'NAPMNOI', 'NAPMSDI', 'NAPMII', 'NAPMPRI','VXOCLSx', 'Agg wkly hours', 'Currency' from the original raw data set for various reasons. 'VXOCLSx' is excluded from macro dataset but included in financial dataset to calculate financial uncertainty. In historical vintage data before 2014:12, all 'HWI' and "NAPM' related series are not reported. Also 'Agg wkly hours' and 'Currency' are not found in FRED-MD dataset.

[9]The R code is used to get estimates of the error terms following stochastic volatility processes.

displayed in Table 1, where MU (both real-time and ex-post versions) shows stronger predictive power in the post-crisis period. In table 10, we report out-of-sample forecasting summary results. We use the first five years data from 2003:7-2008:6 for in-sample estimation. The rest ten years are used for out-of-sample forecasting evaluation. We also divide the out-of-sample period into two sub-samples. In general, real time MU has less predictive power than ex-post MU from JLN (2015).

**References**


Adrian, T., Boyarchenko, N., & Giannone, D. (2019). Vulnerable growth. *American Economic Review*, *109*(4), 1263–89.

Adrian, T., Grinberg, F., Liang, N., & Malik, S. (2018). *The term structure of growth-at-risk*. International Monetary Fund.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593–1636.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, *77*(3), 623–685.

Bloom, N., Bond, S., & Van Reenen, J. (2007). Uncertainty and investment dynamics. *The review of economic studies*, *74*(2), 391–415.

Bloom, N., Floetotto, M., Jaimovich, N., Saporta-Eksten, I., & Terry, S. J. (2018). Really uncertain business cycles. *Econometrica*, *86*(3), 1031–1065.

Caldara, D., Fuentes-Albero, C., Gilchrist, S., & Zakrajsek, E. (2016). The macroeconomic impact of financial and uncertainty shocks. *European Economic Review*, *88*, 185–207.

Clements, M. P., & Hendry, D. F. (1999). *Forecasting nonstationary economic time series*. MIT Press.

Gilchrist, S., & Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, *102*(4), 1692–1720.

Hendry, D. F., & Mizon, G. E. (2005). Forecasting in the presence of structural breaks and policy regime shifts. In J. H. Stock & D. W. K. Andrews (Eds.), *Identification and inference for econometric models: Essays in honor of thomas j. rothenberg* (p. 481-502). Cambridge University Press.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, *105*(3), 1177–1216.

Ludvigson, S. C., Ma, S., & Ng, S. (in press). Uncertainty and business cycles: exogenous impulse or endogenous response? *Journal of Finance*.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Stock, J. H., & Watson, M. (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, *173*, 205–261.

Table 1: Summary table of in-sample predictive regression

| | full sample | 1985:1-2018:12 | 1985:1-2007:12 | 2008:1-2018:12 |
|---|---|---|---|---|
| | | h=1 | | |
| EPU | 17 | 17 | 47 | 5 |
| MU | 50 | 43 | 41 | 65 |
| FU | 41 | 24 | 14 | 64 |
| EBP | 44 | 46 | 38 | 59 |
| NFCI | 47 | 46 | 36 | 74 |
| | | h=3 | | |
| EPU | 24 | 24 | 54 | 5 |
| MU | 43 | 52 | 57 | 85 |
| FU | 48 | 31 | 20 | 76 |
| EBP | 62 | 54 | 46 | 70 |
| NFCI | 42 | 49 | 35 | 79 |
| | | h=12 | | |
| EPU | 23 | 23 | 28 | 23 |
| MU | 50 | 58 | 68 | 83 |
| FU | 47 | 22 | 27 | 84 |
| EBP | 64 | 52 | 55 | 72 |
| NFCI | 27 | 45 | 43 | 81 |

Note: The table summarizes the number of series with significant indexes. Full sample means we use the complete data span for each uncertainty index. EPU index is from 1985:1-2018:12; MU and FU index is from 1960:7-2018:12; EBP is from 1973:1-2018:12; NFCI is from 1971:1-2018:12. The macroeconomic data used in estimating factors and predictive regressions in all three sub-samples (1985:1-2018:12, 1985:1-2007:12 and 2008:1-2018:12) are the same for each uncertainty index.

Table 2: Summary table of out-of-sample forecasting

| | 1995:1-2018:12 | 1995:1-2007:12 | 2008:1-2018:12 |
|---|---|---|---|
| | | h=1 | |
| EPU | 23 | 50 | 26 |
| MU | 50 | 35 | 55 |
| FU | 48 | 32 | 58 |
| EBP | 60 | 39 | 67 |
| NCFI | 51 | 47 | 57 |
| | | h=3 | |
| EPU | 38 | 54 | 34 |
| MU | 44 | 49 | 48 |
| FU | 41 | 39 | 58 |
| EBP | 63 | 51 | 67 |
| NCFI | 46 | 56 | 48 |
| | | h=12 | |
| EPU | 39 | 53 | 40 |
| MU | 26 | 54 | 28 |
| FU | 19 | 27 | 33 |
| EBP | 39 | 30 | 67 |
| NCFI | 20 | 42 | 29 |

Note: This table reports the number of series with positive out-of-sample R2 relative to the benchmark model. The pseudo out-of-sample forecasting values are computed from 1995:1 to 2018:12. Data starting from 1985:1 to 1994:12 are used for in-sample estimation. The parameter estimation, model selection, and lag orders are estimated recursively. The multiple steps ahead forecasts are computed using direct approach. Forecasting performance before and after 2008 financial crisis are evaluated separately.

Table 3: In-sample predictive regression in Gilchrist & Zakrajsek (2012): monthly data

| | h=1 | | | h=3 | | | h=12 | | |
|---|---|---|---|---|---|---|---|---|---|
| | EMP | UER | IPM | EMP | UER | IPM | EMP | UER | IPM |
| | | | | 1973:1-2018:12 | | | | | |
| Term Spread | -0.06 | 1.37* | -0.71** | -0.12* | 1.65** | -0.97*** | -0.38*** | 3.83*** | -1.38*** |
| | (-0.81) | (1.31) | (-1.88) | (-1.49) | (1.89) | (-2.58) | (-3.54) | (4.29) | (-3.78) |
| Real FFR | -0.01 | -0.42 | 0.31 | -0.02 | -0.26 | 0.36 | -0.11** | 0.60 | 0.20 |
| | (-0.16) | (-0.45) | (1.02) | (-0.29) | (-0.33) | (1.17) | (-1.70) | (1.30) | (0.97) |
| Predicted GZ spread | -0.64*** | 3.82 | -1.40 | -0.84*** | 4.26** | -1.42* | -1.24*** | 4.49*** | -1.38** |
| | (-2.49) | (1.18) | (-1.20) | (-3.01) | (1.69) | (-1.31) | (-4.14) | ?2.61) | (-1.75) |
| EBP | -0.63*** | 12.52*** | -3.09*** | -0.82*** | 12.25*** | -3.32*** | -1.01*** | 9.98*** | -1.59** |
| | (-3.85) | (5.34) | (-3.59) | (-4.14) | (6.16) | (-3.77) | (-4.05) | ?4.99) | (-1.94) |
| EPU | 0.001 | -0.004 | 0.003 | 0.002 | -0.033 | 0.003 | 0.003 | -0.037* | 0.002 |
| | (0.30) | (-0.12) | (0.36) | (0.97) | (-1.13) | (0.28) | (1.14) | (-1.54) | (0.24) |
| NFCI | -0.47*** | 8.47*** | -2.50*** | -0.61*** | 8.71*** | -2.95*** | -0.43** | 4.63*** | -2.05*** |
| | (-2.79) | (3.99) | (-2.87) | (-3.00) | (4.22) | (-3.36) | (-2.19) | (3.13) | (-3.59) |
| | | | | 1985:1-2007:12 | | | | | |
| Term Spread | -0.26*** | 3.80** | -0.84* | -0.25*** | 3.65** | -0.68 | -0.26** | 2.22** | -0.07 |
| | (-2.56) | (1.94) | (-1.38) | (-2.34) | (2.11) | (-1.13) | (1.92) | (1.81) | (-0.17) |
| Real FFR | 0.09 | -0.92 | 0.05 | 0.04 | -0.66 | -0.06 | -0.09 | 0.85 | -0.32 |
| | (0.91) | (-0.49) | (0.09) | (0.39) | (-0.43) | (-0.10) | (-0.89) | (0.77) | (-0.92) |
| Predicted GZ spread | 0.01 | -1.63 | -0.30 | -0.11 | -1.68 | -0.76 | -0.46* | 2.08 | -1.73** |
| | (0.05) | (-0.30) | (-0.20) | (-0.37) | (-0.35) | (-0.54) | (-1.39) | (0.59) | (-1.84) |
| EBP | -0.10 | 3.95 | -1.56* | -0.23 | 6.71** | -1.92** | -0.73*** | 10.41*** | -2.24*** |
| | (-0.56) | (0.94) | (-1.60) | (-1.10) | (1.76) | (-1.97) | (-2.68) | (3.61) | (-2.91) |
| EPU | -0.01*** | 0.21*** | -0.03*** | -0.01*** | 0.16*** | -0.02 | -0.001 | 0.023 | 0.005 |
| | (-4.48) | (4.24) | (-2.61) | (-3.68) | (4.13) | (-1.77) | (-0.34) | (0.94) | (0.69) |
| NFCI | -0.39 | 3.21 | -1.51 | -0.37 | 1.57 | -2.00 | -0.36 | 1.63 | -2.61* |
| | (-0.81) | (0.31) | (-0.51) | (-0.66) | (0.18) | (-0.67) | (-0.61) | (0.26) | (-1.53) |
| | | | | 2008:1-2018:12 | | | | | |
| Term Spread | -0.11 | 2.26 | -2.34* | -0.03 | 0.60 | -1.15 | 0.00 | -0.69 | -0.64 |
| | (-0.66) | (0.52) | (-1.59) | (-0.15) | (0.23) | (-0.93) | (-0.02) | (-0.39) | (-0.55) |
| Real FFR | -0.41** | 7.58 | -1.23 | -0.81*** | 11.84*** | -2.60* | -1.95*** | 20.12*** | -6.49*** |
| | (-2.03) | (1.21) | (-0.59) | (-3.75) | (2.88) | (-1.31) | (-7.11) | (6.47) | (-2.96) |
| Predicted GZ spread | -0.03 | 5.21 | -4.67 | -0.35 | 8.08 | -2.32 | -0.33 | 4.64 | -2.04 |
| | (-0.06) | (0.50) | (-0.88) | (-0.90) | (1.07) | (-0.55) | (-0.96) | (1.19) | (-0.67) |
| EBP | -0.26 | 11.88** | -8.45*** | -0.41* | 9.30** | -6.49*** | -0.29 | 4.08* | -1.90 |
| | (-1.06) | (1.71) | (-3.19) | (-1.42) | (2.00) | (-3.34) | (-1.19) | (1.46) | (-1.18) |
| EPU | 0.004*** | -0.038 | 0.023** | 0.004*** | -0.042** | 0.002 | 0.005*** | -0.034** | 0.010 |
| | (3.34) | (-1.09) | (1.81) | (2.94) | (-1.92) | (0.15) | (2.55) | (-2.12) | (0.95) |
| NFCI | -1.27** | 20.68 | 2.52 | -1.30** | 16.64* | 1.57 | -1.37*** | 11.55** | 1.90 |
| | (-1.82) | (1.20) | (0.43) | (-1.77) | (1.39) | (0.31) | (-3.16) | (2.18) | (0.58) |

Table 4: In-sample predictive regression in Gilchrist & Zakrajsek (2012): quarterly data

| 1973:Q1-2018:Q4 (h=4) | | | | | |
|---|---|---|---|---|---|
| | GDP | C-D | C-NDS | I-RES | I-NRS |
| Term Spread | -0.49*** | -1.28*** | -0.34*** | -4.29*** | 2.06** |
| | (-4.38) | (-3.04) | (-4.37) | (-3.46) | (2.54) |
| Real FFR | 0.29* | 1.99 | 0.18 | 1.88 | -0.65 |
| | (1.58) | (1.12) | (0.40) | (-0.55) | (-0.29) |
| Predicted GZ spread | 0.63* | 5.31 | 0.33** | 3.19** | 0.67 |
| | (-1.28) | (0.00) | (-1.89) | (-2.38) | (-1.35) |
| EBP | -0.83*** | 0.11 | -0.52 | 2.88 | -5.02*** |
| | (-2.37) | (-0.25) | (-0.79) | (0.50) | (-5.64) |
| EPU | 0.00 | 0.05* | 0.00 | 0.09*** | -0.12 |
| | (0.41) | (1.31) | (-1.15) | (2.46) | (0.14) |
| NFCI | -1.35*** | -14.50*** | -0.46 | -15.34** | -1.84* |
| | (-3.40) | (-2.31) | (-0.30) | (-2.00) | (-1.27) |
| Adjusted R2 | 0.44 | 0.29 | 0.48 | 0.48 | 0.48 |
| 1985:Q1-2007:Q4 (h=4) | | | | | |
| Term Spread | -0.49* | -1.28* | -0.34** | -4.29*** | 2.06* |
| | (-1.61) | (-1.41) | (-1.74) | (-2.91) | (1.51) |
| Real FFR | 0.29* | 1.99*** | 0.18 | 1.88** | -0.65 |
| | (1.34) | (3.30) | (1.23) | (1.78) | (-0.44) |
| Predicted GZ spread | 0.63 | 5.31*** | 0.33 | 3.19 | 0.67 |
| | (1.01) | (3.11) | (0.79) | (1.15) | (0.20) |
| EBP | -0.83** | 0.11 | -0.52* | 2.88* | -5.02*** |
| | (-1.80) | (0.09) | (-1.63) | (1.60) | (-2.51) |
| EPU | 0.00 | 0.05*** | 0.00 | 0.09*** | -0.12*** |
| | (0.28) | (2.67) | (-0.16) | (3.45) | (-3.17) |
| NFCI | -1.35 | -14.50*** | -0.46 | -15.34*** | -1.84 |
| | (-1.19) | (-4.70) | (-0.53) | (-2.61) | (-0.28) |
| Adjusted R2 | 0.25 | 0.34 | 0.18 | 0.41 | 0.52 |
| 2008:Q1-2018:Q4 (h=4) | | | | | |
| Term Spread | -0.23 | -0.31 | -0.48** | 3.73* | -1.90 |
| | (-0.67) | (-0.40) | (-1.92) | (1.59) | (-0.74) |
| Real FFR | -1.56*** | -5.41*** | -0.30* | -14.27*** | -1.28 |
| | (-9.80) | (-8.45) | (-1.39) | (-5.21) | (-0.47) |
| Predicted GZ spread | -0.08 | 0.35 | 0.64 | -0.33 | -4.20 |
| | (-0.07) | (0.13) | (0.81) | (-0.05) | (-0.45) |
| EBP | -0.48 | -0.36 | 1.23*** | -1.43 | -5.65 |
| | (-0.88) | (-0.32) | (2.60) | (-0.39) | (-1.17) |
| EPU | 0.00 | 0.01 | 0.00 | 0.01 | 0.13*** |
| | (0.31) | (0.51) | (-1.21) | (0.55) | (3.25) |
| NFCI | -1.27 | -4.44* | -2.82*** | -5.86 | -5.26 |
| | (-0.89) | (-1.58) | (-2.79) | (-0.64) | (-0.45) |
| Adjusted R2 | 0.79 | 0.85 | 0.76 | 0.84 | 0.64 |

Note: This table reports the predictive regression coefficients and t-statistics. Statistical significance at the 10%, 5% and 1% levels are denoted by *, ** and ***, respectively.

Table 5: In-sample predictive quantile regression for GDP growth

| $\tau$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| | | | h=1 | | |
| EPU | -0.55** | -0.18* | -0.06 | -0.18 | -0.07 |
| | (-2.27) | (-1.32) | (-0.64) | (-1.22) | (-0.57) |
| MU | -1.65*** | -1.50*** | -0.88*** | -0.35 | 0.60* |
| | (-4.42) | (-5.05) | (-2.46) | (-1.25) | (1.54) |
| FU | -0.43* | -0.42** | -0.42*** | -0.06 | 0.25 |
| | (-1.36) | (-1.96) | (-2.73) | (-0.54) | (1.27) |
| EBP | -0.94*** | -0.76*** | -0.83*** | -0.45*** | -0.24 |
| | (-4.34) | (-3.09) | (-2.55) | (-2.86) | (-0.89) |
| NFCI | -1.24*** | -1.09*** | -0.67*** | -0.48** | -0.18 |
| | (-3.83) | (-4.93) | (-3.56) | (-1.75) | (-0.56) |
| | | | h=4 | | |
| EPU | -0.01 | -0.01** | -0.03** | -0.01 | -0.01* |
| | (-0.95) | (-2.17) | (-2.10) | (-0.93) | (-1.34) |
| MU | -0.48*** | -0.45** | -0.64** | -0.27** | 0.25 |
| | (-2.89) | (-2.18) | (-2.04) | (-1.86) | (1.47) |
| FU | 0.11** | 0.05 | 0.18 | 0.13 | 0.06** |
| | (1.71) | (0.81) | (1.27) | (1.20) | (1.79) |
| EBP | -0.14** | -0.25** | -0.03* | -0.14* | -0.08** |
| | (-1.71) | (-1.69) | (-1.33) | (-1.32) | (-2.10) |
| NFCI | -0.02 | -0.01* | -0.02 | -0.02 | -0.02** |
| | (-0.89) | (-1.62) | (-1.16) | (-0.90) | (-1.90) |
| | | | h=8 | | |
| EPU | 0.34*** | 0.17* | 0.13* | 0.19* | 0.05 |
| | (3.15) | (1.33) | (1.31) | (1.36) | (0.76) |
| MU | -0.58*** | -0.39*** | -0.43** | -0.19** | -0.01 |
| | (-4.34) | (-4.21) | (-2.29) | (-2.11) | (-0.67) |
| FU | 0.15 | 0.06** | 0.04 | 0.13* | -0.02** |
| | (0.73) | (1.93) | (0.78) | (1.48) | (-2.20) |
| EBP | 0.06 | 0.05 | 0.08 | 0.12 | 0.07* |
| | (1.07) | (0.83) | (1.15) | (1.23) | (1.50) |
| NFCI | 0.22** | 0.11* | 0.25** | 0.38** | 0.48*** |
| | (1.84) | (1.62) | (1.67) | (2.08) | (3.53) |

Note: This table reports the quantile regression coefficients and t-statistics for five uncertainty indexes, adding one index to the benchmark model individually. Statistical significance at the 10%, 5% and 1% levels are denoted by *, ** and ***, respectively.

Table 6: Out-of-sample $R^2$ for GDP growth forecasts

| $\tau$ | | 0.2 | | | 0.4 | | | 0.6 | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | h=1 | | | | | |
| EPU | -0.50 | -0.07 | -1.01 | -2.81 | 0.52 | -7.03 | 0.47 | 2.91 | -2.51 |
| MU | 1.84 | 1.69 | 2.01 | -0.89 | -0.37 | -1.54 | 1.75 | -0.78 | 4.84 |
| FU | -0.39 | -1.12 | 0.48 | -0.65 | 0.11 | -1.62 | -2.77 | -2.75 | -2.79 |
| EBP | -0.27 | -5.71 | 6.12 | -3.30 | -2.94 | -3.75 | -3.72 | -5.61 | -1.41 |
| NFCI | 5.63 | 3.93 | 7.62 | 0.19 | 0.97 | -0.81 | 0.67 | -2.85 | 4.95 |
| | | | | h=4 | | | | | |
| EPU | 3.31 | 6.97 | 0.43 | -3.74 | 1.04 | -8.05 | -2.02 | 2.23 | -6.08 |
| MU | 5.79 | 5.28 | 6.18 | 2.95 | 0.26 | 5.37 | 3.04 | 3.96 | 2.16 |
| FU | -9.02 | -15.06 | -4.26 | -3.90 | -2.90 | -4.80 | -1.18 | -1.42 | -0.95 |
| EBP | -0.63 | -6.08 | 3.66 | -8.17 | -12.96 | -3.86 | -7.25 | -17.01 | 2.07 |
| NFCI | -1.69 | -11.97 | 6.41 | -3.69 | -3.33 | -4.01 | -1.11 | -0.32 | -1.87 |
| | | | | h=8 | | | | | |
| EPU | -15.07 | 6.39 | -40.86 | -12.29 | -4.07 | -23.26 | -5.57 | -6.81 | -3.94 |
| MU | -1.56 | -0.24 | -3.15 | 0.35 | -1.09 | 2.28 | -0.10 | -1.72 | 2.03 |
| FU | -12.35 | -6.76 | -19.07 | -6.19 | -6.01 | -6.45 | -3.97 | -8.01 | 1.34 |
| EBP | -15.11 | -17.52 | -12.22 | -7.36 | -11.83 | -1.39 | -1.40 | -5.91 | 4.54 |
| NFCI | -6.48 | 0.88 | -15.32 | -8.00 | -2.36 | -15.54 | -2.26 | -3.92 | -0.09 |

Note: The table reports out-of-sample R2 (in percentage) relative to the benchmark model.

Table 7: In-sample predictive quantile regression for GDP growth: 1973:I-2007:IV

| $\tau$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| | | | h=1 | | |
| EPU | -0.38 | -0.49*** | -0.30** | -0.19 | 0.03 |
| | (-1.21) | (-3.93) | (-2.01) | (-1.06) | (0.50) |
| MU | -1.22*** | -0.94*** | -0.77** | -0.61** | -0.28 |
| | (-5.31) | (-7.33) | (-2.27) | (-1.80) | (-1.03) |
| FU | -0.51*** | -0.29 | -0.20 | 0.01* | -0.16 |
| | (-2.61) | (-0.97) | (-1.00) | (1.33) | (-0.98) |
| EBP | -0.78*** | -0.52** | -0.84*** | -0.70** | -0.64*** |
| | (-2.66) | (-2.08) | (-3.12) | (-1.89) | (-3.01) |
| NFCI | -0.72* | -0.45 | -0.46 | -0.33 | -0.34 |
| | (-1.51) | (-1.02) | (-1.26) | (-0.97) | (-1.14) |
| | | | h=4 | | |
| EPU | -0.04 | 0.07 | -0.21* | -0.24** | -0.22** |
| | (-0.58) | (0.71) | (-1.45) | (-1.88) | (-1.84) |
| MU | -1.43*** | -0.97*** | -0.98*** | -0.90*** | -0.53*** |
| | (-6.04) | (-9.48) | (-9.49) | (-9.68) | (-5.84) |
| FU | -0.08 | -0.06 | -0.01 | -0.02 | -0.09 |
| | (-0.75) | (-0.91) | (-0.78) | (-0.92) | (-0.62) |
| EBP | -0.62** | -0.50*** | -0.38** | -0.22* | -0.12* |
| | (-2.20) | (-3.34) | (-1.98) | (-1.44) | (-1.53) |
| NFCI | -0.80*** | -0.43** | -0.15 | -0.05 | -0.14* |
| | (-3.97) | (-2.12) | (-0.88) | (-0.48) | (-1.45) |
| | | | h=8 | | |
| EPU | 0.21 | 0.32** | 0.21** | 0.02 | -0.01 |
| | (1.11) | (2.03) | (1.82) | (0.58) | (-1.07) |
| MU | -1.03*** | -0.84*** | -0.68*** | -0.76*** | -0.56*** |
| | (-4.53) | (-7.39) | (-5.15) | (-7.17) | (-4.30) |
| FU | 0.03 | 0.04 | -0.05 | -0.02 | 0.04 |
| | (1.18) | (0.63) | (-0.60) | (-0.52) | (0.54) |
| EBP | 0.03* | -0.09* | -0.17** | -0.07 | -0.15 |
| | (1.53) | (-1.43) | (-1.92) | (-0.97) | (-0.92) |
| NFCI | -0.17 | -0.13 | -0.19** | -0.10 | -0.37*** |
| | (-0.90) | (-1.23) | (-1.91) | (-0.92) | (-4.25) |

Note: This table reports the quantile regression coefficients and t-statistics for five uncertainty indexes, adding one index to the benchmark model individually. Statistical significance at the 10%, 5% and 1% levels are denoted by *, ** and ***, respectively.

Table 8: In-sample predictive quantile regression for GDP growth: 2008:I-2018:IV

| $\tau$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| | | | h=1 | | |
| EPU | -0.17 | -0.16 | 0.16 | 0.03 | -0.14 |
| | (-0.46) | (-0.55) | (0.73) | (0.80) | (-0.55) |
| MU | -0.51 | -1.58** | -0.61** | -0.71** | -0.86* |
| | (-0.61) | (-2.05) | (-1.96) | (-2.04) | (-1.53) |
| FU | 0.01 | -0.24 | -0.17* | -0.46* | -0.06 |
| | (0.52) | (-0.82) | (-1.29) | (-1.30) | (-0.51) |
| EBP | -1.73*** | -0.94** | -1.77** | -2.00*** | -1.86*** |
| | (-2.95) | (-1.97) | (-2.07) | (-2.60) | (-2.96) |
| NFCI | -3.08*** | -2.72*** | -3.34*** | -2.61*** | -3.29*** |
| | (-3.74) | (-5.07) | (-7.50) | (-3.98) | (-7.40) |
| | | | h=4 | | |
| EPU | -0.14** | -0.18*** | -0.06*** | -0.07** | -0.15*** |
| | (-2.19) | -(2.97) | (-2.92) | (-2.18) | (-2.40) |
| MU | -0.56*** | -0.55** | -0.85*** | -1.04*** | -1.20*** |
| | (-2.37) | (-1.90) | (-3.08) | (-4.67) | (-3.44) |
| FU | -0.67*** | -0.71*** | -0.63*** | -0.69*** | -0.75*** |
| | (-3.67) | (-5.76) | (-4.80) | (-3.66) | (-4.39) |
| EBP | -0.52*** | -0.48** | -0.45*** | -0.57** | -0.65*** |
| | (-3.58) | (-1.90) | (-2.35) | (-2.16) | (-2.74) |
| NFCI | -1.55*** | -1.63*** | -1.56*** | -1.52*** | -1.57*** |
| | (-14.59) | (-9.18) | (-7.99) | (-8.71) | (-3.12) |
| | | | h=8 | | |
| EPU | 0.08*** | 0.08*** | 0.09*** | 0.12*** | 0.07*** |
| | (5.18) | (2.97) | (3.17) | (2.86) | (2.78) |
| MU | -0.30*** | -0.54*** | -0.70*** | -0.66*** | -0.61*** |
| | (-3.91) | (-3.20) | (-6.24) | (-5.08) | (-4.89) |
| FU | -0.24*** | -0.27*** | -0.28*** | -0.27*** | -0.27*** |
| | (-4.97) | (-3.91) | (-3.03) | (-2.61) | (-2.90) |
| EBP | -0.17*** | -0.18*** | -0.22** | -0.22** | -0.18*** |
| | (-2.96) | (-2.55) | (-1.68) | (-2.19) | (-6.39) |
| NFCI | 0.05*** | 0.06*** | 0.06 | 0.07* | 0.07 |
| | (3.04) | (3.78) | (0.80) | (1.31) | (1.00) |

Note: This table reports the quantile regression coefficients and t-statistics for five uncertainty indexes, adding one index to the benchmark model individually. Statistical significance at the 10%, 5% and 1% levels are denoted by *, ** and ***, respectively.

Table 9: Summary table of in-sample predictive regression

| | 2003:7-2018:10 | 2003:7-2007:12 | 2008:1-2018:10 |
|---|---|---|---|
| | h=1 | | |
| real time MU | 22 | 23 | 39 |
| MU in JLN (2015) | 52 | 20 | 66 |
| | h=3 | | |
| real time MU | 41 | 29 | 64 |
| MU in JLN (2015) | 64 | 25 | 85 |
| | h=12 | | |
| real time MU | 62 | 69 | 69 |
| MU in JLN (2015) | 72 | 61 | 83 |

Table 10: Summary table of out-of-sample forecasting

| | 2008:7-2018:10 | 2008:7-2013:6 | 2013:7-2018:10 |
|---|---|---|---|
| | h=1 | | |
| real time MU | 66 | 69 | 57 |
| MU in JLN (2015) | 71 | 72 | 57 |
| | h=3 | | |
| real time MU | 45 | 45 | 53 |
| MU in JLN (2015) | 68 | 66 | 59 |
| | h=12 | | |
| real time MU | 37 | 39 | 65 |
| MU in JLN (2015) | 29 | 29 | 66 |

Table 11: Comparison between real-time MU and ex-post MU: in-sample quantile regression

| $\tau$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| *2003:Q3-2018:Q3* | | h=1 | | | |
| real time MU | 0.06 | 0.02 | 0.06 | 0.09 | 0.03* |
| | (0.58) | (0.97) | (0.75) | (0.78) | (1.57) |
| MU from JLN | 0.01 | -0.28 | -0.44 | -0.38 | -0.12 |
| | (0.87) | (-0.94) | (-1.04) | (-0.99) | (-0.77) |
| | | h=4 | | | |
| real time MU | 0.07 | -0.10** | -0.22* | -0.18* | -0.26** |
| | (0.97) | (-1.75) | (-1.28) | (-1.34) | (-1.93) |
| MU from JLN | -0.88*** | -0.55** | -0.48** | -0.32** | -0.44** |
| | (-3.99) | (-1.84) | (-1.78) | (-2.04) | (-1.94) |
| | | h=8 | | | |
| real time MU | 0.01** | -0.23* | -0.08 | -0.56*** | -0.33*** |
| | (1.81) | (-1.48) | (-1.22) | (-4.65) | (-3.81) |
| MU from JLN | -0.39** | -0.39** | -0.33* | -0.42** | -0.52*** |
| | (-1.72) | (-1.67) | (-1.58) | (-2.26) | (-4.42) |
| *2008:Q1-2018:Q3* | | h=1 | | | |
| real time MU | 0.71 | 0.30 | 0.64** | 0.75* | 0.13 |
| | (1.27) | (0.93) | (1.69) | (1.64) | (0.88) |
| MU from JLN | -1.08* | -1.13** | -1.73*** | -1.39*** | -0.60 |
| | (-1.52) | (-1.84) | (-2.74) | (-2.44) | (-1.19) |
| | | h=4 | | | |
| real time MU | 0.01 | 0.01* | 0.01 | 0.01* | 0.01 |
| | (1.23) | (1.49) | (1.24) | (1.48) | (1.09) |
| MU from JLN | -0.48*** | -0.74*** | -0.97*** | -0.88*** | -0.70*** |
| | (-2.53) | (-2.51) | (-3.90) | (-4.27) | (-3.52) |
| | | h=8 | | | |
| real time MU | -0.33*** | -0.32*** | -0.32*** | -0.33*** | -0.32*** |
| | (-5.27) | (-2.62) | (-3.26) | (-2.34) | (-3.15) |
| MU from JLN | -0.91*** | -1.04*** | -1.02*** | -1.06*** | -1.08*** |
| | (-15.97) | (-8.17) | (-7.02) | (-8.94) | (-12.02) |

Note: This table reports the quantile regression coefficients and t-statistics (numbers in the parenthesis) for real time MU and ex-post MU constructed in JLN (2015). Statistical significance at the 10%, 5% and 1% levels are denoted by *, ** and ***, respectively.

# The Impact of Oil Prices on Products Groups Inflation: is the Effect Asymmetric?

Ligia Topan[1], Miguel Jerez[1,2], and Sonia Sotoca[1]

[1] Universidad Complutense de Madrid
[2] Fundación Ramon Areces

**Abstract.** In this paper we assess the oil price pass-through into both, the global inflation in Spain and the inflation derived from the non-deterministic prices of the standard European classification of product groups, during the period 2002-2018. To this end we fit a transfer function to inflation in each group, extended to allow for an asymmetry in the transmission of positive/negative oil cost shocks, that is, a "rockets and feathers effect". Our results show that most often there is a significant asymmetry, which can be explained by the degree of competition in each market. Even more, we show that allowing for asymmetric effects yields a remarkable improvement in the precision of inflation forecasts.

**Keywords**: Oil price, Products groups inflation, Asymmetric effects, Transfer function, Forecasting.

## 1 Introduction

Many studies test for asymmetric effects of oil price shocks. Some of them investigate their effect on macroeconomic and financial activity, while others concentrate in the pass-through of oil cost into gasoline price. The presence of asymmetry in the latter case is known as "rockets and feathers" effect.

The effect of these shocks on macroeconomic and financial activity has been investigated by Dhaoui et al. (2018), who show an asymmetric long-run impact of oil prices on the stock markets of Poland, the US and Austria. Huang et al. (2017) discuss whether an oil price shock could have an asymmetric response on the stock market in China. They conclude that there is no such effect. Gately and G. Huntington (2001) estimated the effects on energy and oil demand of changes in income and oil prices, for 96 of the largest countries in the world. They found that oil demand often reacts more to increases in oil prices than to decreases. Rahman and Serletis (2010) find that oil price volatility is a major determinant of the US macroeconomic activity, with a stronger effect on output growth in the high-volatility regime of oil price than in the low volatility regime. Donayre and Wilmot (2016) show that the reduction in inflation due to a negative oil price shock is larger than the increase in inflation after a positive innovation in Canada. Finally, Alvarez et al. (2011) show that the inflationary effect of oil price changes on Spanish and euro area is limited, even though crude oil price fluctuations are a major driver of inflation variability.

The literature about the effect of oil price shocks on gasoline prices builds on the seminal paper by Bacon (1991), who coined the term "rockets and feathers". This expression means that gasoline prices tend to shoot up "as rockets" when oil prices increase, but usually fall "like feathers" when crude costs go down. Kristoufek and Lunackova (2015) re-investigated this effect for seven developed countries, finding no statistical evidence of asymmetry. Radchenko (2005), detected a significant asymmetric transfer of oil price variations on gasoline prices, perhaps due to the market power of large retailers in U.S. Tappata (2009) and Lewis and Marvel (2011) focus on the demand side of the market. They argue that the explanation of the "rockets and feathers" effect is that the consumers search "the best deal" less intensively when the gasoline price is going down than when is raising. Last, Borenstein and Shepard (2002) argue that wholesale gasoline prices respond with a lag to cost shocks because it is costly for firms to adjust production and inventory.

Despite this large literature, studies about the sensitivity of product group prices to crude costs are lacking. In this paper we will analyze this sensitivity and will test whether the response to positive and negative shocks is roughly the same.

Our main objectives are: (a) building econometric models for the total inflation and products groups inflation in Spain as a function of oil prices; (b) obtaining a quantitative measure of the potential asymmetries between positive and negative shocks in oil prices, and (c) using the estimated models to check the forecasting power of the models built.

The methodology we employ is based on transfer function models (Box et al., 2015). There are two reasons for this choice. First, we will be working with seasonal time series, for which transfer function models with ARIMA errors are better suited. Second, the transfer function assumes unidirectional causality, which is adequate in our case because Spanish inflation and oil prices show unidirectional dynamic (Granger) causality from the former variable to the latter, with no significant feedback.

Our basic hypothesis is that a positive shock in oil prices may have a different effect than a negative one over product inflation. To define product groups we use the European Classification of Individual Consumption by Purpose (hereafter ECOICOP) disaggregation of the Consumer Price Index provided by the Spanish Institute of Statistics.

The structure of the paper is as follows: Section 2 describes the dataset and the econometric methodology employed. Section 3 presents and discusses the positive vs. negative oil shocks effects for the general inflation rate. Section 4 does the same for the inflation in each product group. In Section 5 we test the forecasting capacity of the models and, finally, Section 6 summarize the main conclusions of this work.

## 2 Data and Methods

### 2.1 Dataset and Variables

The dataset employed in this work includes the general and ECOICOP Consumer Price Index provided by INE, as well as the Brent[3] price published by the U.S. Energy

[3]Brent oil price per barrel in US Dollars (USD).

Administration (hereafter EIA). The ECOICOP Consumer Price Index is a functional disaggregation of the general Consumer Price Index (hereafter CPI). For that purpose, the shopping basket products are classified in 12 groups: Aliments and non-alcoholic drinks; Alcoholic drinks and tobacco; Clothing and footwear; Dwelling and supplies; Furniture and household goods; Health; Transport; Communications; Entertainment and culture; Education; Restaurants and hotels; and Others goods and services. We excluded four groups (Alcoholic drinks and tobacco, Health, Communications and Education) because, in Spain, the corresponding prices are essentially determined by the government and are therefore deterministic.

As crude oil prices are originally quoted in US Dollars (USD), we also used the USD/EURO exchange rate published by the European Central Bank (hereafter ECB). All the time series are observed in a monthly frequency from January 2002 to November 2018, for a total of 203 observations. Table 1 provides further details about this dataset.

**Table 1.** Definition of the dataset.

| Notation | Variable | Source |
|---|---|---|
| $P_t^i$ | General and ECOICOP CPI | Spanish Instutute of Statistics, INE |
| $O_t^{USD}$ | Brent Oil Price in USD | US Energy Administration, EIA |
| $ER_t$ | EUR/USD exchange rate | European Central Bank, ECB |
| $O_t^{EUR}$ | Brent Oil Price in EURO | EIA and ECB |

The original values of these variables were transformed to annual percent rates, which are the actual variables to be analyzed. To denote this transformation we consider that, for any variable, $X_t$, $r^{12}(X_t)$ is the corresponding annual rate, defined as:

$$r^{12}(X_t) = (\frac{X_t}{X_{t-12}} - 1) \times 100$$

Figure 1 displays the general inflation rate and its first-order difference. It can be interpreted as the monthly change in annual inflation and, therefore, can be interpreted as a monthly acceleration , if positive, or deceleration, if negative. The annual rates are non-stationary and requires an additional difference to show a stable mean. The series $r^{12}(O_t^{EUR})$, $r^{12}(O_t^{USD})$ and $r^{12}(ER_t)$, not shown here for brevity, have the same properties.

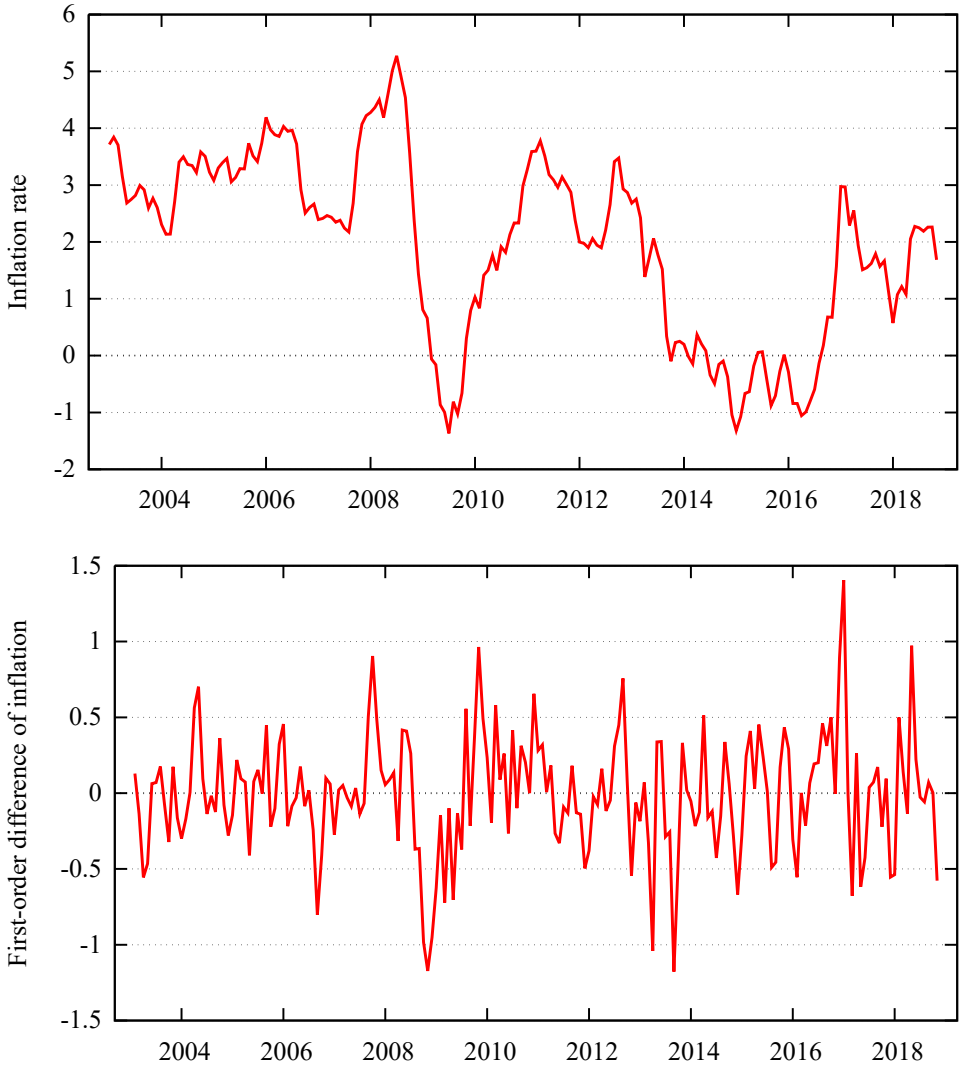



**Fig. 1.** General inflation rate $r^{12}(P_t^G)$ and its stationary series (acceleration) $\nabla r^{12}(P_t^G)$ .

The Dickey-Fuller (ADF) and Kwiatkowski, Phillips, Schmidt y Shin (KPSS) tests, see (Dickey and Fuller, 1981) and (Kwiatkowski et al., 1992), confirm that the first-order difference of these variables are stationary in the mean (Table 2 ). Note that, the null hypothesis of ADF test is that the series has a unit root, while KPSS test assumes that it is stationary. Statistical testing is more decisive when rejecting the null and, because of this, these tests supplement each other. In particular, ADF and KPSS are more decisive when the series is stationary and non-stationary respectively.

**Table 2.** Unit-root tests for the first order difference series of: inflation $\nabla r^{12}(P_t^G)$; Brent price per Barrel in euros $\nabla r^{12}(O_t^{EUR})$ and dollars $\nabla r^{12}(O_t^{USD})$; and exchange rates EUR/USD $\nabla r^{12}(ER_t)$.

| | $\nabla r^{12}(P_t^G)$ | $\nabla r^{12}(O_t^{EUR})$ | $\nabla r^{12}(O_t^{USD})$ | $\nabla r^{12}(ER_t)$ |
|---|---|---|---|---|
| ADF | -4.9200 | -6.1905 | -6.7559 | -10.2662 |
| | (<0.01) | (<0.01) | (<0.01) | (<0.01) |
| KPSS | 0.0473 | 0.0256 | 0.0296 | 0.2183 |
| | (>0.10) | (>0.10) | (>0.10) | (>0.10) |

### 2.2 Univariate Analysis and Transfer Function

Our basic model is a transfer function (Box et al., 2015). A transfer function model is a flexible and efficient representation for a unidirectional causal relationships, allowing for instantaneous and lagged effects, seasonal autocorrelation and intervention variables could be easily added if were required. In our particular case, the transfer functions considered link different inflation series with oil prices. In this way, the relationship model captures the influence of oil price changes to inflation, while the part of inflation explained by other unspecified factors is represented by the error term model.

To parameterize the transfer function, we employed the Box *et al.* (2015) methodology as follows:

1. We first performed an univariate analysis of the inflation and oil price series,
2. ...to filter them using the univariate model for the input (oil price),
3. ...and we computed the sample cross correlation function between the series prewhitened in this way, and finally,
4. ...the error term was modelled with the ARIMA structure of inflation.

## 3 Empirical Results for General Inflation

### 3.1 Symmetric and Asymmetric Transfer Function Estimations

Following Box et al. (2015) the standard univariate identification analysis, not shown here for simplicity, suggest an ARIMA $(1,1,0) \times (0,0,1)_{12}$ specification for the series $r^{12}(P_t^G)$, $r^{12}(O_t^{USD})$ and $r^{12}(O_t^{EUR})$. These models are the base for the transfer function construction and forecasting.

To build the transfer function, first we filtered the series $\nabla r^{12}(P_t^G)$ to shocks in $\nabla r^{12}(O_t^{EUR})$, using the model for the later, and then we computed the sample cross-correlation function (CCF) between both series, which is shown in Figure 2. This cross-correlation function:

($i$) ...has no significant values in the negative lags, which means that there is no inverse causality relationship between oil prices and inflation.

($ii$) ...suggests that a shock in $\nabla r^{12}(O_t^{EUR})$ has a positive and significant effect over $\nabla r^{12}(P_t^G)$ and $\nabla r^{12}(P_{t+1}^G)$.

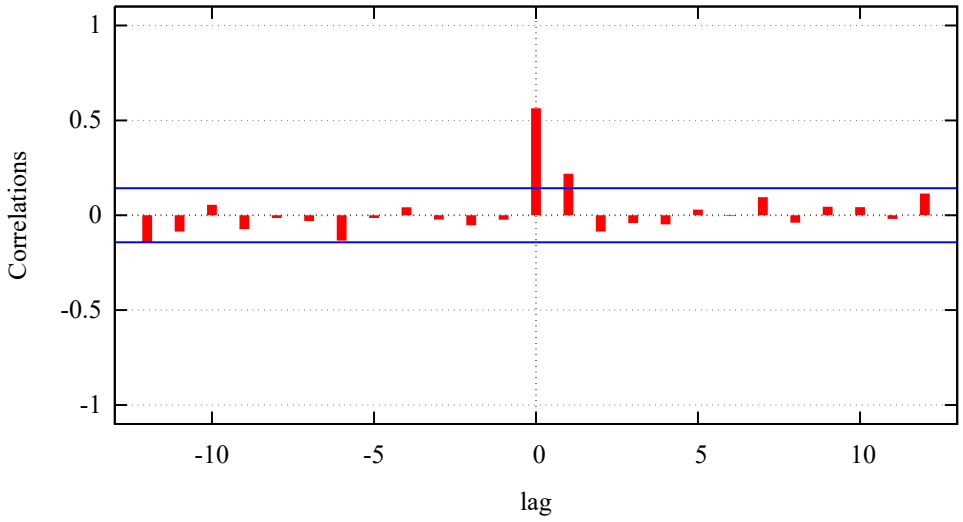



**Fig. 2.** Cross correlations between the prewhitened series of inflation in Spain, $\nabla r^{12}(P_t^G)$ and the lagged annual variation rate of Brent prices in euros. Note that negative lags are actually leads for $\nabla r^{12}(O_t^{EUR})$.

Previous results suggest a transfer function specification relating inflation with the contemporary and first lagged values of the annual variation of oil prices and an error term with the ARIMA $(1,1,0)\times(0,0,1)_{12}$ structure of the endogenous variable. This transfer function corresponds to the symmetric approach, since this specification assumes that the magnitude of the effects on inflation are equals (in absolute values), both for negative and positive shocks in oil prices. The main estimation results for this transfer function are:

$$r^{12}(P_t^G) = \underset{(<0.01)}{(0.0148} + \underset{(<0.01)}{0.0088L)} r^{12}(O_t^{EUR}) + \hat{N}_t \tag{1}$$

$$\underset{(<0.01)}{(1-0.2631L)}\nabla\hat{N}_t = \underset{(<0.01)}{(1-0.6465L^{12})}\hat{a}_t \tag{2}$$

$$\hat{\sigma}_a = 0.2239 \qquad log-lik = 11.3802$$

$$AIC = -12.7606$$

where $L$ denotes the lag operator, $log-lik$ is the (log) value of the Gaussian likelihood function on convergence and AIC stands for the Akaike (1974) Information Criterion. The figures in parentheses are the corresponding $p$-values.

To take into account the potential existence of asymmetric effects on inflation, we build an alternative transfer function that allows positive shocks in oil prices to have a different effect than negative ones:

$$r^{12}(P_t^G) = \underset{(<0.01)}{(0.0135} + \underset{(<0.01)}{0.0075L)} r^{12}(O_t^{EUR}) + \underset{(<0.01)}{(0.0105} + \underset{(<0.01)}{0.0109L)} r^{12}_{neg}(O_t^{EUR}) + \hat{N}_t \tag{3}$$

$$\underset{(<0.01)}{(1-0.2057)}\nabla\hat{N}_t = \underset{(<0.01)}{(1-0.7803L^{12})}\hat{a}_t \tag{4}$$

$$\hat{\sigma}_a = 0.2038 \quad log-lik = 26.7775$$

$$AIC = -39.5549$$

where we define a new variable as follows:
$r^{12}_{neg}(O^{EUR}_t) = r^{12}(O^{EUR}_t)$ if $r^{12}(O^{EUR}_t)$ is less or equal to 0 or,
$r^{12}_{neg}(O^{EUR}_t) = 0$ otherwise

All the parameters in (1)-(2) and (3)-(4) are significant and the residuals do not show relevant autocorrelations, so we consider them statistically adequate. But, in (3)-(4) the parameters corresponding to negative shocks are statistically significant, so there is a significant asymmetric effect. Furthermore, the AIC[4] values are consistently smaller than those in the symmetric model, so (3)-(4) fits better than (1)-(2). The LR test[5], shown in Table 3, also confirms that the goodness of the asymmetric model improves considerably.

The symmetric transfer function implies that:

(*i*) the value of inflation in any month is affected by the annual change in Brent price in the same and previous month, so the effect is transient;
(*ii*) the expected total response of inflation to a 1 percentage point (p.p.) increase in $r^{12}O^{EUR}_t$ would be $\hat{g} = 0.0146 + 0.0088 = 0.0236$ p.p. Obviously this total response, which is known in the time series literature as the transfer function gain, provides a measure of the sensitivity of the inflation level to changes in oil prices.
(*iii*) the total response of inflation to a 1 p.p. decrease in $r^{12}O^{EUR}_t$ would be $-0.0236$ p.p., which is the same magnitude as in the case of a positive increase in absolute values.

The asymmetric transfer function implies that:

(*i*) the value of inflation in any month is affected by the annual change in Brent price in the same and previous month, so the effect is transient
(*ii*) the expected total response of inflation to a 1 p.p. increase in $r^{12}O^{EUR}_t$ would be $\hat{g} = 0.0135 + 0.0075 = 0.0210$ p.p.
(*iii*) the total response of inflation to a 1 p.p. decrease in $r^{12}O^{EUR}_t$ would be $\hat{g} = -(0.0135 + 0.0075 + 0.0105 + 0.0109) = -0.0424$ p.p. In absolute value, this gain doubles the one corresponding to a positive increase. Therefore, inflation is more sensitive to negative shocks in oil prices.

[4]The same conclusion is supported by both, Schwarz (1978) and Hannan and Quinn (1979) Information Criteria, but we do not show the values for simplicity.

[5]The *LR*-test is a likelihood ratio test, computed to compare the fit of the symmetric and asymmetric models. The null hypothesis of this test, in our models, is that the asymmetric model fit as well as the symmetric one.

## 4 Empirical Results for ECOICOP Inflation

As explained in previous sections, one of our main objectives is to test for asymmetric effects of oil prices on various products categories. Such an asymmetric behavior would be evidence of the different degrees of market competition. Our hypothesis is that on industries with higher level of competitiveness, a negative shock in oil prices produce a higher effect than a positive one, in absolute values. This means that if crude oil is a raw material for a very competitive industry, producers will find themselves forced to reduce prices more when a negative shock occurs than raise them when oil price increases, to ensure their permanence on the market. In the case of industries with lower level of competitiveness, producers will not translate their costs reductions to the prices of their products.

### 4.1 Symmetric and Asymmetric transfer Function Estimations

Per products groups we use the same specification as in the general inflation rate, (1)-(2) for the symmetric model and (3)-(4) for the asymmetric one.

Table 3 displays a summary of the results for the groups where we found asymmetric effects. We show the corresponding long-term gain, both for positive and negative shocks and the LR goodness test that confirms in each case that de asymmetric model fits better than the symmetric one. Groups are sorted from more to less sensitive to negative shocks.

**Table 3.** Summary of sensitivity and goodness results for the groups with asymmetric effects.

| Products groups | Gain "+" shock | Gain "-"shock | $LR$-test ($p$-value) |
|---|---|---|---|
| Transport | 0.1003 | -0.1868 | 64.9784 (<0.01) |
| Dwelling and Supplies | 0.0343 | -0.0582 | 10.3422 (<0.01) |
| **Global Inflation** | **0.0210** | **-0.0424** | **30.7944 (<0.01)** |
| Restaurants and Hotels | 0.0000 | -0.0070 | 6.4969 (0.0388) |
| Clothing and Footwear | 0.0000 | -0.0059 | 14.8378 (<0.01) |
| Entertainment and culture | 0.0078 | 0.0071 | 5.1798 (0.0750) |

**Note**: The figures in parentheses are the corresponding $p$-values.

Note that:

($i$) ...the groups "Transport" and "Dwelling and Supplies" display the higher asymmetrical responses and both of them are more sensitive to negative oil shocks than the total inflation.

($ii$) ...the Transport group is the most sensitive to oil shocks, with a 0.100 percentage points increase and a $-0.188$ percentage points decrease as a reaction of positive and negative shocks of 1 percentage point.

($iii$) ...the groups: "Restaurants and Hotels" and "Clothing and Footwear" show asymmetric effects, but their prices are less sensitive to negative oil shocks than the

total inflation. In these three cases, a positive shock in crude oil price has a null effect on their own prices. This suggest that the producers avoid translate to their products prices the costs increases, so we could conclude that there is a high level of competitiveness on this industries.

(*iv*) ..."Entertainment and Culture" prices seems to be asymmetrically affected by oil price variations. Note that in this case the sign of a negative shock is positive, contrary to the total inflation behavior when oil prices drops. This particular response may be due to the heterogeneity of the products included in this group, but this issue deserves further research.

The group "Aliments and Non-alcoholic drinks"is affected by oil shocks symmetrically. "Other goods and services" and "Furniture and household goods"are not affected by oil price variations.

## 5 Out-of-sample Forecasting and Predictive Quality

As shown in previous subsection, fluctuations in oil prices have significant effects on inflation and are therefore relevant to forecast short-term inflation. One way to check out the predictive quality of the models is to compare the out-of-sample forecasting of both, symmetric and asymmetric specifications. Using both models we computed out-of-sample forecasts for a twelve months ahead time horizon. Figure 3 show the inflation out-of-sample forecasts of the symmetric and asymmetric transfer function models and the observed variable real values. Note that:

(*i*) the dynamic of the symmetric model tend to spread more uncertainty than the asymmetric one because of its larger confidence interval.

(*ii*) the symmetric model has a lower level of sensitiveness to shocks, this means that the negatives shocks are smoother in the symmetric model respect to the asymmetric one.

We also computed some of the common measures to evaluate the predictions of both specifications. Results are shown in Table 4. Note that all the measures to evaluate the

**Table 4.** Measures to evaluate the predictions of the symmetric and asymmetric specification

| | Transfer Function Models | |
|---|---|---|
| Measure | Symmetric | Asymmetric |
| Root mean squared error | 0.2899 | 0.1619 |
| Mean absolute error | 0.2739 | 0.1272 |

predictions improve considerably when we allow for asymmetric effects. Furthermore, the root mean squared error of predictions is reduced a $44,15\%$ in the case of the asymmetric model. This fact leads to a remarkable enhancement in the precision of inflation forecasts.

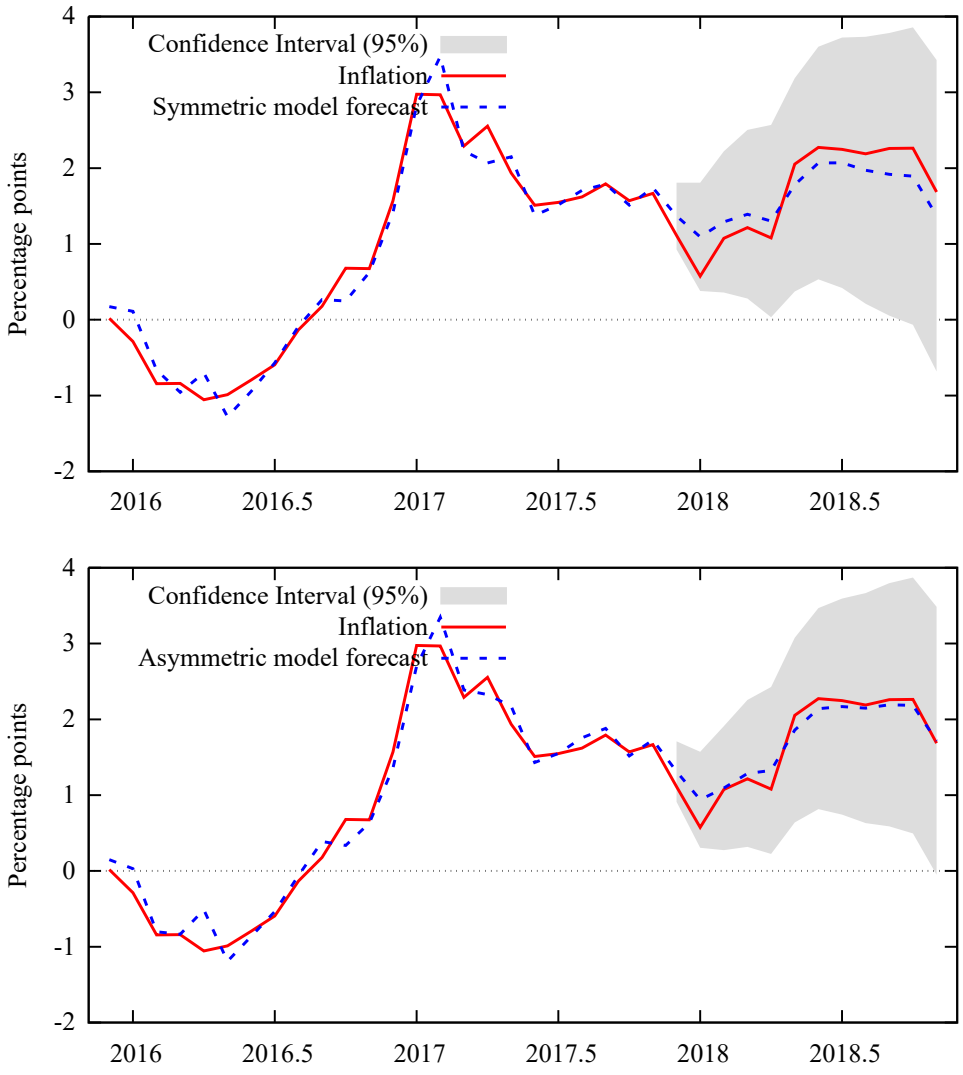



**Fig. 3.** Out-of-sample inflation forecasts of the symmetric and asymmetric models

## 6 Concluding Remarks

The results in previous sections show that a shock in oil prices in a given month creates a transient effect of the same sign on the inflation in the current and next month. In the long-term, the total expected effect of an increase in oil price of one percentage (p.p.) point is about 0.021 p.p. in inflation, while a decrease of the same magnitude brings

down inflation by −0.042 p.p. Therefore, the sensitivity of inflation to negative shocks in oil price is almost twice than the corresponding response to positive innovations.

Per product groups, there are important differences in the effects of oil price shocks into inflation. For example, in "Aliments and non-alcoholic drinks" the effect is symmetric. On the other hand, prices of "Furniture and household goods" and "Other goods and services" are not sensitive to changes in oil price. All the other groups display asymmetric effects, being "Transport" the one receiving a larger impact, with a 0.100 p.p. increase and a 0.187 p.p. decrease as a reaction of positive and negative shocks of 1 p.p.

Besides providing a more accurate description of the pass-through effect, allowing for asymmetric effects in the inflation models improves remarkably the forecast accuracy in comparison with the analogue symmetric effects model. In particular, in the case of total inflation and a twelve months ahead time horizon for predictions: (a) the root mean squared error is reduced by 44.15% and (b) the confidence intervals decrease substantially.

## Acknowledgements

This research was supported by Complutense University of Madrid and Santander Bank programme CT17/17 - CT18/17. We thank these institutions for providing funding for our project. We would also like to show our gratitude to The Economic Analysis and Quantitative Economy Department and Complutense Institute of Economic Analysis, for all the support.

# Bibliography


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723.

Alvarez, L. J., Hurtado, S., Sánchez, I., and Thomas, C. (2011). The impact of oil price changes on spanish and euro area consumer price inflation. *Economic Modelling*, 28(1):422 – 431.

Bacon, R. W. (1991). Rockets and feathers: the asymmetric speed of adjustment of uk retail gasoline prices to cost changes. *Energy Economics*, 13(3):211 – 218.

Borenstein, S. and Shepard, A. (2002). Sticky prices, inventories, and market power in wholesale gasoline markets. *The RAND Journal of Economics*, 33(1):116–139.

Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics.

Dhaoui, A., Goutte, S., and Guesmi, K. (2018). The asymmetric responses of stock markets. *Journal of Economic Integration*, 33(1):1096–1140.

Dickey, D. and Fuller, W. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4):1057–1072.

Donayre, L. and Wilmot, N. (2016). The asymmetric effects of oil price shocks on the canadian economy. *International Journal of Energy Economics and Policy*, 6(2):167–182. cited By 2.

Gately, D. and G. Huntington, H. (2001). The asymmetric effects of changes in price and income on energy and oil demand. *C.V. Starr Center for Applied Economics, New York University, Working Papers*, 23.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41(2):713–723.

Huang, S., An, H., Gao, X., and Sun, X. (2017). Do oil price asymmetric effects on the stock market persist in multiple time horizons? *Applied Energy*, 185:1799 – 1808. Clean, Efficient and Affordable Energy for a Sustainable Future.

Kristoufek, L. and Lunackova, P. (2015). Rockets and feathers meet joseph: Reinvestigating the oil–gasoline asymmetry on the international markets. *Energy Economics*, 49:1 – 8.

Kwiatkowski, D., Phillips, P., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178.

Lewis, M. and Marvel, H. P. (2011). When do consumers search? *Journal of Industrial Economics*, 59(3):457–483.

Radchenko, S. (2005). Oil price volatility and the asymmetric response of gasoline prices to oil price increases and decreases. *Energy Economics*, 27(5):708–730.

Rahman, S. and Serletis, A. (2010). The asymmetric effects of oil price and monetary policy shocks: A nonlinear var approach. *Energy Economics*, 32(6):1460 – 1466.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Tappata, M. (2009). Rockets and feathers: Understanding asymmetric pricing. *The RAND Journal of Economics*, 40(4):673–687.

# Forecasting macroeconomic processes with missing or hidden data

John Mashford

School of Mathematics and Statistics
University of Melbourne, Victoria 3010, Australia
E-mail: mashford@unimelb.edu.au

June 4, 2019

**Abstract**

An approach to utilizing Bayesian methodology to model and forecast macroeconomic processes over a region of interest and multiple planning horizons in the case where some of the data is unavailable is proposed. The generalized state space associated with such a model is defined. The stochastic model considered is autoregressive in time and has a simple standard covariance structure for the spatial component. It is proposed simply to fill in missing values by utilising standard forecasting techniques on complete data blocks to extend the block up to just before the next complete data block for any set of data corresponding to any given spatial index.

## 1 Introduction

Consider a system $\mathcal{S}$ whose task it is to analyse given financial data over a region $\mathcal{R}$ of spacetime, i.e. the system has access to a collection of time and place stamped financial data and is tasked to analyse the data and produce a report or summary such as Dow Jones $= x$ or "bottom line" $= y$ (to give two very simplistic examples) and, furthermore, to predict the state of some $\mathcal{R}' \supset \mathcal{R}$.

Suppose that $\mathcal{R}$ can be represented as a product

$$\mathcal{R} = \mathcal{R}_{\text{time}} \times \mathcal{R}_{\text{space}}, \tag{1}$$

where

$$\mathcal{R}_{\text{time}} = [a, b], \text{ for some } a, b \in \mathbf{R} \cup \{-\infty, \infty\}, a < b, \tag{2}$$

and

$$\mathcal{R}_{\text{space}} \subset \{rx \in \mathbf{R}^3 : x \in S^2\}, \tag{3}$$

where $S^2$ denotes the unit sphere in 3D space and $r$ is the mean radius of the earth relative to its centroid.

We suppose that the *possible* input data to $\mathcal{S}$ is located at some discrete set of spatio-temporal points $S = \{t_1, \ldots, t_T\} \times \{s_1, \ldots, s_n\} \subset \mathcal{R}_{\text{time}} \times \mathcal{R}_{\text{space}}$ with $t_i < t_{i+1}, \forall i = 1, \ldots, T-1$ and $s_i \neq s_j, \forall i, j \in \{1, \ldots, n\}$ with $i \neq j$. Define $\omega : \{1, \ldots, T\} \times \{1, \ldots, n\} \to \{0, 1\}$ be defined by

$$\omega(i, j) = \begin{cases} 1 \text{ if data is available at time } t_i \text{ and place } s_j \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

It is reasonable to believe that the processes determining the distribution of economic activity over the region under consideration over the period of time under consideration, and in particular at the $n$ observation points at the $T$ time periods, is independent of whether or not we have data at these or other points in space and time. However, if we are to use Bayesian modelling then we need to have data specified over the configuration space (or state space) of the model that we are using.

In a previous paper (Mashford *et al.*, 2018) we described a technique for imputing missing values by generating the next value in the MCMC process by sampling from a truncated normal distribution based on the current value. In the present paper we sketch an approach where missing values are imputed by applying standard Bayesian forecasting technique within each block of contiguous data. We will consider this approach for the rest of this paper.

## 2 Definition of the generalized configuration space

Let $\mathcal{D}$ be the data type under consideration. E.g.

$$\mathcal{D} = \mathbf{R}, \tag{5}$$

when the data represents the net profit of company $i$ over year $t$ (to give a simple example) (company $i$ is "incorporated" at a particular head office address) or, more elaborately,

$$\mathcal{D} = \text{D\&B (Dun and Bradsreet) company (or individual) report structure.} \tag{6}$$

Then the data domain space is

$$S = \{(t, i) \in \{1, \ldots, T\} \times \{1, \ldots, n\} : \omega(t, i) = 1\}. \tag{7}$$

The state space of the general model that we will define is

$$\mathcal{F}(S, \mathcal{D}) = \{f : S \to \mathcal{D}\}. \tag{8}$$

For the rest of this paper we will assume that Eq. 5 holds and that the economic activity that we are interested in is represented by the profitability of all the companies in our region of interest. An element of $\mathcal{F}(S, \mathcal{D})$ is a function $f$ which assigns a datum $f(t, i) \in \mathbf{R}$ to each $(t, i) \in S$.

In the simple case when $\omega(t, i) = 1, \forall t = 1, \ldots, T, i = 1, \ldots, n$ then $\mathcal{F}(S) = \mathbf{R}^{T \times n}$ is the set of $T \times n$ real matrices. In the classical forecasting problem we have

$$(S' \subset \mathcal{R}') = \{1, \ldots, T + 1\} \times \{s_1, \ldots, s_n\}, \tag{9}$$

i.e. one would like to forecast one time step (e.g. year, month, 5 years, etc.) into the future.

Define

$$a_i = \{\min\{t \in \{1, \ldots, T\} : \omega(t, i) = 1\}, \tag{10}$$

$$b_i = \max\{t \in \{1, \ldots, T\} : \omega(t, i) = 1\}, \tag{11}$$

$$I_i = \{a_i, a_i + 1, \ldots, b_i - 1, b_i\}, \tag{12}$$

forall $i = 1, \ldots, n$. Thus for all $i \in \{1, \ldots, n\}$ we have data for some of the points in the interval $I_i$ but nothing outside $I_i$. Note that $I_i$ may be all of $\mathbf{N} = \{1, 2, \ldots\}$ (the natural numbers). One may choose to discard intervals which are too short by introducing a censoring threshold or else assume that the points in them have been amalgamated with nearby points in $S$. Note that if $\omega(t, i) = 0, \forall t = 1, \ldots, T$ then $I_i = [\infty, -\infty] = \emptyset$ and we may assume that the location $s_i$ can be safely ignored and the remaining $s_i$ locations have been renumbered accordingly.

Let

$$S_i = \{t \in I_i : \omega(t, i) = 1\}. \tag{13}$$

We have data at the points in $S = (S_1, \ldots, S_n)^T$ which is a subset of the "ragged lattice" $I = (I_1, \ldots I_n)^T$. The state space for the model that we are proposing is $\mathcal{F}(S)$ which is a subset of the ragged array $\mathcal{F}(I)$.

## 3 The data process and parameter models for the BHAM model (with full information)

In this case there exists $T \in \{1, 2, \ldots\}$ such that $S_i = I_i = \{1, \ldots, T\}, \forall i = 1, \ldots, n$. We may consider the following BHAM model.

Data model:

$$\begin{aligned} \mathbf{H}_t &= \boldsymbol{\mu} + \mathbf{Z}_t + \mathbf{v}_t, \\ \mathbf{v}_t &\sim \text{iid } N(0, \sigma_v^2 \mathbf{I}). \end{aligned} \tag{14}$$

Process model:

$$\begin{aligned} \mu_i &= \mu_\mu + f(\xi, \rho_i) + u_i, \\ \mathbf{u} &= (u_1, \ldots, u_n)' \\ \mathbf{u} &\sim N(0, \sigma_\mu^2 \mathbf{M}(\psi_\mu)), \\ \mathbf{Z}_t &= \phi \mathbf{Z}_{t-1} + \mathbf{w}_t, \\ \mathbf{w}_t &\sim \text{iid } N(0, \sigma_w^2 \mathbf{M}(\psi)), \end{aligned} \tag{15}$$

One might consider modeling the economy in $\mathcal{R}$ as a spatio-temporal stochastic process $\mathbf{H}$ driven, in a hierarchical fashion (Berger *et al.* 2001, Cressie and Wikle 2011), by an underlying process $\mathbf{Z}$ and also a process $\boldsymbol{\Lambda}$ for the mean $\boldsymbol{\mu}$ of $\mathbf{H}$. Here $\mathbf{H}_t$ is a random vector representing the actual current profit status of all the companies $\{H_{ti}\}$. (It may be a transformation $y \to h$ of the original data $y$ if one wants to make the data more manageable, this is an easy exercise (see Mashford, *et al.*, 2018)). $\mathbf{Z}$ represents the underlying process model driving $\mathbf{H}$, $\boldsymbol{\mu}$ is a random vector providing a measure of central tendency (e.g. mean) for the random matrix $\mathbf{H}$, $'$ denotes transpose $T$. Also we assume that the process vector $\mathbf{Z}$ is autoregressive as a function of time.

One may model the covariance structure of the spatial part of the process $\mathbf{Z}$ by the standard exponential covariogram $\boldsymbol{\Sigma} : (rS^2)^n \times (0, \infty) \times (0, \infty) \to \mathbf{R}^{n \times n}$ (where $\mathbf{R}$ denotes the real numbers) defined by

$$(\boldsymbol{\Sigma}(\mathbf{s}_1, \ldots, \mathbf{s}_n, \sigma, \psi))_{ij} = \sigma^2 C_\psi(d(\mathbf{s}_i, \mathbf{s}_j)), \tag{16}$$

where $d(\mathbf{s}_i, \mathbf{s}_j)$ is the (geodesic) distance between station number $i$ and station number $j$ (by "station" we mean the address of the home office of the company under consideration),

$C_\psi : [0, \infty) \to (0, \infty)$ is the standard isotropic exponential covariogram function defined by

$$C_\psi(d) = \exp(-\frac{d}{\psi}), \tag{17}$$

$\sigma > 0, \psi > 0$ and $\mathbf{s}_i \in \mathbf{R}^2$ is the location of station $i$.

When $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are given and fixed we may write

$$\mathbf{\Sigma}(\mathbf{s}_1, \ldots, \mathbf{s}_n, \sigma, \psi) = \mathbf{\Sigma}(\sigma, \psi) = \sigma^2 \mathbf{M}(\psi), \tag{18}$$

for $\sigma > 0, \psi > 0$.

$\rho_i$ is the credit rating of the $i^{\text{th}}$ company (as, for example, determined by the Standard and Poors corporation). The $\rho_i$ are given fixed data for the model. $f$ describes a model of the relationship between credit rating and mean annual profit. $\xi$ is a variable ranging over some space, e.g. $\mathbf{R}^k$ (and we will assume this for the rest of this paper). $\xi$ parametrises the non-linear regression modelling the relationship between annual profit and credit rating. One might be able to determine $\xi$ using non-linear regression using simulated annealing fitting a function to the scatter data

$$(\rho_i, \mu_\mu + f(\xi, \rho_i)), \tag{19}$$

to the observed mean profit / credit rating data. $\xi$ may not determined as part of the MCMC (Monte Carlo Markov Chain) procedure because the profit / credit rating model is a stand-alone independent submodule of the BHAM model. It is time independent and does not depend on the detailed financial data, only the profit means and credit ratings.

Note that in the present paper we are assuming that companies can go into the red as much as they like (e.g. if they are funded by the defense budget). One might consider a less fortunate example where companies are closed down when their net profit dips below some figure in $(-\infty, 0)$ as deemed by their bank or banks.

To complete the BHAM framework, we need to specify the prior distributions of parameters from the previous stages. For simplicity, we may consider basic conjugate parameter

models. The following prior distributions might be used for the unknown parameters:

$$\begin{aligned}
\pi_{\mu}(\mu_{\mu}) &= (N(\mu_{\mu_{\mu}}, \sigma^2_{\mu_{\mu}}))(\mu_{\mu}), \\
\pi_{\sigma_v}(\sigma_v) &= (\text{Inverse-gamma}(q_v, r_v))(\sigma^2_v), \\
\pi_{\sigma_w}(\sigma_w) &= (\text{Inverse-gamma}(q_w, r_w))(\sigma^2_w), \\
\pi_{\sigma_{\mu}}(\sigma_{\mu}) &= (\text{Inverse-gamma}(q_{\mu}, r_{\mu}))(\sigma^2_{\mu}), \\
\pi_{\tau}(\tau) &= (N(\mu_{\tau}, \sigma^2_{\tau}))(\tau), \\
\log(\psi) &\sim \text{Uniform}[a_{\psi}, b_{\psi}], \\
\log(\psi_{\mu}) &\sim \text{Uniform}[a_{\psi_{\mu}}, b_{\psi_{\mu}}],
\end{aligned} \tag{20}$$

$\phi = \tanh(\tau), \tau \in (-\infty, \infty)$.

Here the values of hyper-parameters were chosen such that the priors are relatively vague and so that Gibbs sampling could be used as much as possible..

One would use these priors if one were overly concerned about computational efficiency. However with modern powerful computers one is not constrained to a great extent by computational efficiency. Therefore we will assume for the rest of this paper that the priors have any suitable form, not neccessarily congugate and Metropolis algorithm can be used if need be.

Sensitivity analyses showed (in a different but related context (Mashford *et al.*, 2018)) that the quality of the model performance was found to be robust to choice of hyper-parameters.

The parameters for the model are $\mu, v_1, \ldots, v_T, \phi, w_1, \ldots, w_T$. The hyper-parameters are $\sigma_v$ (for $v$), $\sigma_{\mu}$ (for $\mu$), $\rho_i$ (for $\mu_i$), $u_i$ (for $\mu_i$), $\sigma_w$ (for $w$). The hyper-hyper-parameters are $\sigma_{\mu}$ (for $u$) and $\psi_{\mu}$ (for $u$).

Thus the parameter space is

$$\begin{aligned}
\Pi = \{&\mu, v_1, \ldots, v_T, \sigma_v, \sigma_w, \sigma_{\mu}, \phi, w_1, \ldots, w_T, u_i, \psi_{\mu} : \mu \in \mathbf{R}^n, v_1, \ldots, v_T \in \mathbf{R}^n, \\
&\phi \in (-1.1), w_1, \ldots, w_T \in \mathbf{R}^n, \sigma_v, \sigma_w, \sigma_{\mu} > 0, \zeta \in \mathbf{R}^k\}
\end{aligned}$$

The $\rho_i$ are fixed given data provided by e.g. Standard and Poors. $\phi$ is smpled from $(-1, 1)$ and therefore the process is stable.

The model defined by Eq. 14 and Eq. 15 is a stochastic model and to view the consequences of the model we generate an ensemble of solutions. Each solution in the ensemble is defined by an element $\theta \in \Pi$. The ensemble of such parameter values may be obtained

using Bayes theorem for desities as follows.

$$[\Theta|Z](\theta, z) = [Z|\Theta](z, \theta)[\Theta](\theta)/[Z](z). \tag{21}$$

$[\Theta](\theta)$ is the prior which we may choose in a way that we think appropriate, $[Z|\Theta](z, \theta)$ is called the likelihood function and $[Z](z)$ represents the evaluation of the density function generated by the process model evaluated on the real data that we are fitting our Bayesian model to.

It is straightforward to generate an ensemble for **H** given an ensemble for **Z**.

The likelihood function can be computed as follows.

$$\begin{aligned}
\zeta(z,\theta) &= [Z|\Theta](z,\theta) \\
&= p(Z = z|\Theta = \theta) \\
&= p(Z_1 = z_1, \ldots, Z_T = z_T|\Theta = \theta) \\
&= p(Z_1 = z_1, \ldots, Z_{T-1} = z_{T-1}|\Theta = \theta) pr(Z_T = z_T|Z_1 = z_1, \ldots, Z_{T-1} = z_{T-1}, \Theta = \theta) \\
&= p(Z_1 = z_1, \ldots, Z_{T-1} = z_{T-1}|\Theta = \theta) p(Z_T = z_T|Z_{T-1} = z_{T-1}, \Theta = \theta) \\
&\vdots \\
&= p(Z_1 = z_1|\Theta = \theta) \prod_{t=2}^{T} p(Z_t = z_t|Z_{t-1} = z_{t-1}, \Theta = \theta)
\end{aligned}$$

## 4 The generalized (ragged) BHAM model

For the ragged model the stochastic model is the same as that in the model with full information. The matrix valued function $\psi \mapsto M(\psi)$ is knwn. **M** only depends on the location of the companies under considerstion and not on their prrofit value. Therefore **M** is fully known. Only some of the data values $z_{ti}$ are not known.

In (Mashford *et al.*, 2018) the missing $z_{ti}$ values were filled in by imputation in which the missing value was drawn from a truncated normal distribution based on its last value in the iterative MCMC procedure.

We now propose an approach to filling in the missing $z_{ti}$ values. Let $i \in \{1, \ldots, n\}$ be the index of some company of interest. The domain for data for that company, that is, $S_i$ is a union of full intervals,

$$S_i = \bigcup_{j=1}^{J} S_{ij} = \bigcup_{j=1}^{J} [a_{ij}, b_{ij}], \tag{22}$$

for some $J \in \{1, 2, \ldots\}$, where $a_1 < b_1 < a_2 < b_2 < \ldots < b_{K-1} < a_K < b_K$.

We propose that, rather than filling in the values of, for example, $z_{ti}$ for $t$ between $b_{1i}$ and $a_{2i}$ by imputation drawing from a truncated normal distribution, one can first generate

the Bayesian model for the data in $\{a_{1i}, \ldots, b_{1i}\}$ as described in the previous section and then fill in the values from $b_{1i}+1$ up to $a_{2i}$ by using the model obtained for $S_{1i}$ to forecast these values.

Continuing in this fashion one can generate an ensemble of predictions for the performance of company $i$ at arbitrary times in the future.

More elaborately, one may predict the unknown future starting at a given block for company $i$ by forecasting ahead from all the preceding blocks and taking a weighted sum of all these predictions. This procedure can be carried out for all the companies from company 1 to company $n$ thereby giving a prediction of the economic activity over arbitrary times and for all corporate locations. Such a procedure can be executed within the iterative MCMC process.

This may be the subject of future research.

## Acknowledgements

The author would like to thank Yong Song for very helpful discussions and help with this work.

## References

Berger, J. O., de Oliveira, V., and Sansó, B., *Objective Bayesian analysis of spatially correlated data*. J. Am. Stat. Assoc. 96, pp. 1361-1374, 2001.

Cressie, N. and Wikle, K., *Statistics for Spatio-Temporal Data*. Wiley; 2011.

Gelman, A., Carlin, B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, Chapman & Hall/CRC, 2004.

Mashford, J., Song, Y., Wang, Q. J. and Robertson, D., *A Bayesian hierarchical spatio-temporal rainfall model*, Journal of Applied Statistics, DOI: 10.1080/02664763.2018.1473347, 2018.

Song, Y., Wang, Q. J., Robertson, D. and Mashford, J., *Bayesian hierarchical model for estimating subcatchment rainfall at daily time steps*, Hydrology and Water Resources Symposium 2014, p. 1002.

# Imputing monthly values for quarterly time series. An application performed with Swiss business cycle data

Klaus Abberger[1], Michael Graff[1], Oliver Müller[1] and Boriss Siliverstovs[2]

[1] ETH Zürich, KOF Swiss Economic Institute, Zurich
[2] Latvijas Banka, Riga

**Abstract.** This paper documents an applied investigation into strategies and algorithms to deal with the problem of missing higher frequency data. We refer to Swiss business tendency survey (BTS) data, in particular the KOF manufacturing surveys, which are conducted in both monthly and quarterly frequency. As a result, some information is available at quarterly frequency only. There is a wide range of ways to address this problem comprising univariate and multivariate approaches. We resort to different multivariate imputation algorithms and apply them to generate monthly series out of quarterly series from the KOF BTS in the Swiss manufacturing sector and compare the results. Our strategy to compare the suitability of the different approaches is to make sure that we do possess adequate reference series for the model selection stage. To this end, we apply our procedures to series that are monthly, from which we create artificial quarterly data by deleting the same two out of three data points from each quarter. The candidate series for the imputation of the missing (i.e. deleted) observations are given by the entire set of time series that are resulting from the monthly KOF manufacturing BTS survey. In this way, we resort to a set of indicators that share the common theme, which is a reflection of the Swiss business cycle. With this set of potential indicators, we conduct the different imputations. On this basis, we then run standard tests of forecasting accuracy by comparing the imputed monthly series to the original monthly series. Descriptive statistics like the correlation and the absolute mean or root square forecast error allow ranking the algorithms; statistical tests like the encompassing test reveal whether the different methods are significantly superior/inferior.

**Keywords:** Frequency transformation, Business tendency surveys.

# Hybrid Method Forecasting Stock Market Data

S. Al Wadi
Department of Risk Management and Insurance, The University of Jordan, Jordan.
E-mail: sadam_alwadi@yahoo.co.uk

Ahmad M. Awajan
Department of Mathematics, Faculty of Science, Al-Hussein Bin Talal University, Jordan.
Email: awajanmath@yahoo.com

**Abstract.**
In this article events of productivity of the insurance data in Jordan will be explored and forecasted using some of traditional model which is Exponential model (EM) compound with Wavelet transform (WT) in order to improve the forecasting accuracy. The decomposed dataset will be collected form Amman Stock Exchange (ASE) from Jordan. As a result the forecasting accuracy will be improved by using EM.
Key words: Wavelet Transform, Exponential Model, Insurance Data, Forecasting.

## 1- Introduction

It well known that financial experts find it difficult to make accurate predictions, because industrial stock market trends tend to be nonlinear, uncertain, and non-stationary. No consensus exists among experts as to the effectiveness of forecasting industrial time series [1].
WT is a sufficient model that attempts to forecast stock prices by proposing a method that uses the WT combined with EM.
WT is a relatively new field in signal processing [2]. Wavelets are mathematical functions that decompose data into different frequency components, after which each component is studied with a resolution matched to its scale, where a scale denotes a time horizon [3]. WT is closely related to the volatile and time varying characteristics of the real-world time series and is not limited by the stationary assumption [4]. WT decomposes a process into different scales, making it useful in distinguishing seasonality, revealing structural breaks and volatility clusters, and identifying local and global dynamic properties of a process at specific timescales [5]. WT has been shown to be particularly useful in analyzing, modeling, and predicting the behavior of financial instruments as diverse as stocks and exchange rates [6,7].
This study applies WT functions which are (Haar, Daubechies) to decompose the insurance time series then combine the approximation coefficients with EM in order to make improve the forecasting accuracy then select the best WT function in forecasting. Then finally the authors compare the forecasting using the combined method with the forecasting using EM directly also. The framework combines several statistical methods and soft computing techniques using MATLAB and MINTAB programs. The data used will be presented in next section.

## 2- Methodology and Mathematical Models

In this section the research framework will be presented with the mathematical equations used.

### 2.1. Research Framework

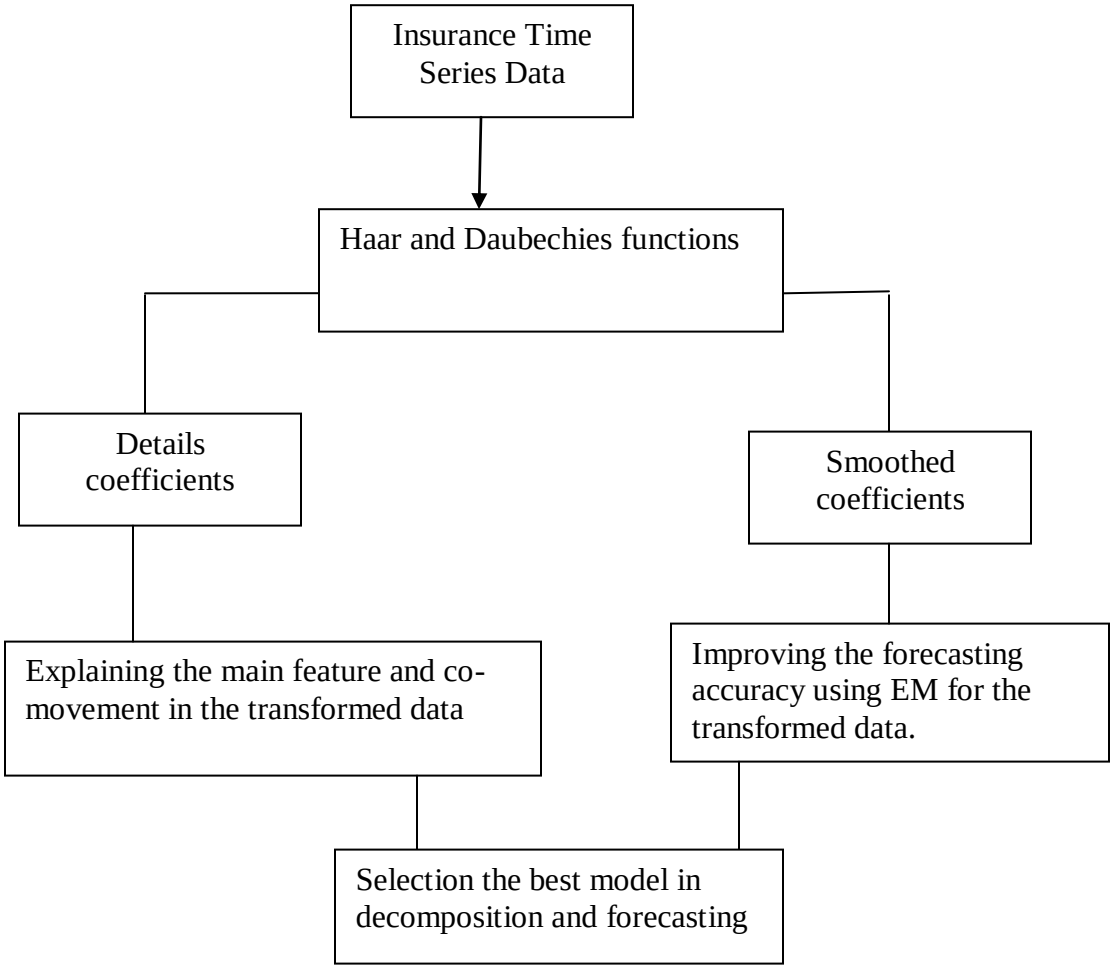



Figure 1. Research Framework

2.2 Wavelet Transform

WT is based on Fourier analysis, which represents any function as the sum of the sine and cosine functions. A wavelet is simply a function of time t that obeys a basic rule, known as the wavelet admissibility condition [8]:

$$C_{\varphi} = \int_{0}^{\infty} \frac{|\varphi(f)|}{f} df < \infty \qquad (1)$$

Where $\varphi(f)$ is the Fourier transform and a function of frequency f, of $\varphi(t)$. The wavelet transform (WT) is a mathematical tool that can be applied to numerous applications, such as image analysis and signal processing. It was introduced to solve problems associated with the Fourier transform as they occur. This occurrence can take place when dealing with nonstationary signals, or when dealing with signals that are localized in time, space, or frequency. Depending on the normalization rules, there are two types of wavelets within a given function/family. Father wavelets describe the smooth and low-frequency parts of a

signal, and mother wavelets describe the detailed and high-frequency components. In the following equations, (2a) represents the father wavelet and (2b) represents the mother wavelet, with j=1, ..., J in the J-level wavelet decomposition: [7]

$$\phi j,k = 2^{-j/2}\phi(t - 2^j k / 2^j)$$
$$\varphi j,k = 2^{-j/2}\varphi(t - 2^j k / 2^j) \qquad (2)$$

Where J denotes the maximum scale sustainable by the number of data points and the two types of wavelets stated above, namely father wavelets and mother wavelets, and satisfies:

$$\int \phi(t)dt = 1 \quad \text{and} \int \varphi(t)dt = 0 \qquad (3)$$

**2.3. Exponential Function**

Exponentials are often used when the rate of change of a quantity is proportional to the initial amount of the quantity. If the coefficient associated with *b* and/or *d* is negative, *y* represents exponential decay. If the coefficient is positive, *y* represents exponential growth. This model has expanded by Taylor (2003) [10] to include methods with multiplicative damped trend. He suggested fifteen exponential smoothing methods. Where each method has two components, seasonality (None, Additive, Multiplicative) and trend (None, Additive, Additive damped, Multiplicative, Multiplicative damped).

**2.4. Mathematical criteria**

The author used some criteria in order to make fair comparison between ARIMA and ARIMA-WT that can be presented in this section. Some types of accuracy criteria have used root means squared error (RMSE), percentage root mean absolute percentage error (MAPE), and mean absolute error (MAE). For the mathematical formulas, refer to [9].

**4) Conclusion**

The goal of this study is to show estimating and forecasting of closed price stock market data. EM is the most general way of forecasting since there is no need for any assumptions and it is not limited to specific type of pattern. This model can be fitted to any set of time series data (stationary or non-stationary). In this study, firstly, the dataset is modelled based on WT. Secondly, we compared EM directly and EM + WT. Then the researchers were found that EM+WT is better than EM directly in forecasting accuracy.

**References**


1. Y.S. Abu-Mostafa, A.F. Atiya, "Introduction to financial forecasting", Applied Intelligence (1996) 205-213.

2. A.Cohen, I.Daubechies, P.Vial, "Wavelets on the interval and fast wavelet transform", Applied and Computational Harmonic (1993)54–81.

3. J.B.Ramsey, Z.Zhang, "The analysis of foreign exchange data using wave form dictionaries", Journal of Empirical Finance (1997)341–372.

4. A.Popoola, K.Ahmad, "Testing the suitability of wavelet preprocessing for TSK fuzzy models", in: Proceeding of FUZZ-IEEE: International Conference Fuzzy System Network (2006), pp.1305–1309.

5. R.Gencay, F.Selcuk, B.Whitcher, "Differentiating intraday seasonalities through wavelet multi-scaling", PhysicaA (2001)543–556.

6. J.B.Ramsey, "The contribution of wavelets to the analysis of economic and financial data", Philosophical Transactions of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences (1999)2593–2606.

7. K.Papagiannaki, N.Taft, Z.-L.Zhang, C.Diot, "Long-term forecasting of internet backbone traffic", IEEE Transactions on Neural Networks (2005)1110–1124.

8. R.Gencay, F.Selcuk, B.Whitcher, "An Introduction to Wavelets and Other Filtering Methods in Finance and Economics", Academic Press, NewYork (2002).

9. F. M. Al-Rawashdi, S. Al Wadi, and M. Hasan Saleh, “Wavelet methods in forecasting for insurance companies listed in Amman Stock Exchange, European Journal of Economics, Finance and Administrative Sciences, (82), 2015.

10. Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend, International journal of Forecasting 19(4): 715–725.

# Measuring the Effect of Unconventional Policies on Market Volatility

Demetrio Lacava[1] and Edoardo Otranto[1]

[1] University of Messina
dlacava@unime.it
[2] University of Messina
eotranto@unime.it

**Abstract.** During the great recession, with interest rate closed to the zero lower bound, many central banks resorted to unconventional monetary policy measures in order to stimulate real economy. These policies consist of central bank's balance sheet expansion which affects real economy by modifying inflation rate expectation during period in which the so-called liquidity trap makes conventional policy, i.e. further cuts of interest rate, no longer effective.
Following other central banks, such as Federal Reserve and Bank of England, the European Central Bank (ECB) established different unconventional monetary measures during the period 2009-2018. Even though the main concern of these policies is the real economy, they have also unintended effects on financial markets that are largely studied by recent literature. Among these effects, it is crucial the positive influence that quantitative easing should have on market uncertainty. Thus, while most authors analyse the effect of unconventional monetary policies on bond market (Boeckx, Dossche and Peersman, 2017; Joyce, Lasaosa, Stevens and Tong, 2010; Krishnamurthy, Nagel and Vissing-Jorgensen, 2014), some others focus on stock market (Ciarlone and Colabella, 2016, 2018; Georgiadis and Grab, 2015) emphasizing the role played by the portfolio-rebalancing channel in transmitting monetary policy decisions (Breedon, Chadha and Waters,2012). Clearly, unconventional policies affect market returns and volatility since, by purchasing assets available in the market, the central bank reduces the amount of those assets incentivizing private investors to rebalance their portfolio opting for a new preferred risk return configuration. In addition, notice how most of the unconventional policies by ECB were established to reduce market uncertainty, which is measured through the expected variance (Rompolis, 2017). Surprisingly, there exists a narrow literature concerning the impact of quantitative easing on volatility as key research objective (Apostolou and Beirne, 2017; Balatti, Brooks, Clements and Kappon, 2016; Converse,2015; Kenourgios, Papadamou and Dimitriou,2015; Shogbuyi and Steeley, 2017), modelling volatility mainly through the GARCH family models (Engle, 1982; Bollerslev, 1986). Despite the effectiveness of GARCH models, the new frontier in analysing volatility is represented by the Multiplicative Error Model, MEM (Engle, 2002; Engle and Gallo, 2006), in which volatility is the product of a time-varying factor (following a GARCH

process) and a positive random variable ensuring positiveness without resorting to logs. Basing on MEM, Otranto (2015) proposes a new model to capture spillovers effects in financial markets, by decomposing the mean equation as the sum of two components, both evolving according to GARCH models. This model could be considered a general framework where inserting the effect of quantitative easing as an unobservable factor, providing its estimate and its weight on the level of volatility. In other terms, we further modify this model to allow volatility to depend on unconventional monetary policy. In particular, in our specification, the first equation composing the mean equation evolves as a GARCH model (capturing the pure volatility mechanism), while the second one follows an autoregressive process with exogenous variables, in order to capture both the announcement effect and the implementation effect of unconventional measures on volatility.

More precisely, our research aims to analyse the impact of unconventional monetary policy by ECB on stock market volatility in four Eurozone countries (France, Germany, Italy and Spain). We proxy for unconventional policies by using three different variables, relating with existing literature in using two of those, i.e. the balance sheet size growth (see for example Apostolou and Beirne, 2017; Voutsinas and Werner, 2011) and the ratio between the securities purchased and total assets (D'amico, English, Lopez-Salido and Nelson, 2012; Voutsinas and Werner, 2011).

In carrying out our analysis we employ realized volatility measure based on high frequency data, which should remove endogeneity arising when monetary policy decisions coincide with a stock price reduction, as argued by Ghysels, Idier, Manganelli and Vergote (2014).

In addition, the volatility dynamics is characterized by several and frequent changes in regimes and frequent jumps, generally with a lower persistence with respect to the quiet periods; this fact could imply changes in the model parameters in unknown (a priori) time. We propose to extend the analysis implementing a Markov Switching model to test the ECB ability to keep volatility in low and high regimes. According to our results, what matters for the effectiveness of these policies is the balance sheet composition rather than the balance sheet size (Curdia e Woodford, 2011). Indeed, it follows that an increase in securities held by ECB for monetary policy purposes relative to total asset reduces volatility in both core and peripheral countries, with disrupted countries, generally, beneficing more. A further proof derives from a different proxy, which tells us that an increase in securities purchased for QE programmes relative to securities held for conventional policies also reduces market volatility. Ultimately, within a Markov Switching framework, it emerges how these programmes contribute in keeping volatility in low regime: the average duration of the QE effects on volatility is about 15 days for France, Italy and Spain. The effect lasts more in Germany, probably because of more favourable economic conditions characterizing this country, during the sample period. In conclusion, there is evidence for the crucial role played by unconventional monetary policies in restoring the proper functioning of the economy when interest rate is closed to the zero lower bound.

**Keywords:** Unconventional monetary policy · Realized Volatility · Multilplicative Error Model · Markov Switching process.

## References

1. Apostolou, A., & Beirne, J., (2017). " Volatility spillovers of Federal Reserve and ECB balance sheet expansions to emerging market economies". Working Paper Series 2044, European Central Bank.

2. Balatti, M., Brooks, C., Clements, M. & Kappou, K., (2016). " Did Quantitative Easing only in ate stock prices? Macroeconomic evidence from the US and UK". Available at SSRN: http://ssrn.com/abstract=2838128.

3. Boeckx J., Dossche M. & Peersman, G. (2017). "Effectiveness and Transmission of the ECB's Balance Sheet Policies". International Journal of Central Banking, International Journal of Central Banking, vol. 13(1), pages 297-333, February.

4. Bollerslev T. (1986). " Generalized autoregressive conditional heteroskedasticity". Journal of Econometrics 31: 307-327.

5. Breedon, F., Chadha, J. S. & Waters, A. (2012). "The financial market impact of UK quantitative easing". BIS Papers chapters in: Bank for International Settlements (ed.), Threat of fiscal dominance?, volume 65, pages 277-304 Bank for International Settlements.

6. Ciarlone, A. & Colabella, A. (2016). " Spillovers of the ECB's non-standard monetary policy into CESEE economies". Ensayos Sobre PolĂtica EconĂłmica, Vol. 34, pp.175-190.

7. Ciarlone A. & Colabella, A. (2018). "Asset price volatility in EU-6 economies: how large is the role played by the ECB?" Temi di discussione (Economic working papers) 1175, Bank of Italy, Economic Research and International Relations Area.

8. Converse, N. (2015). " The impact of unconventional monetary policy on financial uncertainty in emerging markets". Mimeo.

9. Curdia, V. & Woodford, M. (2011). " The central bank balance sheet as an instrument of monetary policy." Journal of Monetary Economics 58 (1): 54-79.

10. D'Amico, S., English, W. B., LĂłpez-Salido, D. & Nelson, E. (2012). " The Federal Reserve's large-scale asset purchase programs: rationale and effects". Finance and Economics Discussion Series 2012-85, Board of Governors of the Federal Reserve System (US).

11. Engle R. F. (1982). " Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom Inflation". Econometrica 50(4): 987-1007.

12. Engle, R. F. (2002). " New frontiers for ARCH models". Journal of Applied Econometrics, 17, 425-446.

13. Engle, R. F., & Gallo, G. M. (2006). " A multiple indicatorsmodel for volatility using intradaily data". Journal of Econometrics, 131, 3-27.

14. Georgiadis, G. & Grab, J. (2015). "Global financial market impact of the announcement of the ECB's extended asset purchase programme". Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute, Working Paper No. 232, March.

15. Ghysels, E., Idier, J., Manganelli, S. & Vergote, O. (2014). " A high frequency assessment of the ECB securities markets programme". Working Paper Series 1642, European Central Bank.

16. Joyce, M., Lasaosa, A., Stevens, I., & Tong, M. (2010). "The financial market impact of quantitative easing". Bank of England working papers 393, Bank of England.

17. Kenourgios, D., Papadamou, S. & Dimitriou, D. (2015). "Intraday exchange rate volatility transmissions across QE announcements". Finance Research Letters, Elsevier, vol. 14(C), pages 128-134.

18. Krishnamurthy, A., Nagel, S. & Vissing-Jorgensen, A. (2014). "ECB policies involving government bond purchases: impact and channels". Working paper, Stanford University.

19. Otranto, E. (2015). " Capturing the spillover effect with Multiplicative Error Models". Communications in Statistics - Theory and Methods, 44:15, 3173-3191.

20. Rompolis, L. S. (2017). "The effectiveness of unconventional monetary policy on risk aversion and uncertainty," Working Papers 231, Bank of Greece.

21. Shogbuyi, A. & Steeley, J. M. (2017). "The effect of quantitative easing on the variance and covariance of the UK and US equity markets". International Review of Financial Analysis, Elsevier, vol. 52(C), pages 281-291.

22. Voutsinas, K., & Werner, R. A. (2011). " New evidence on the effectiveness of "Quantitative Easing" in Japan". CFS Working Paper, No. 2011/30.

# Comparative Investigation of Tests in Modeling Process in Univariate Time Series

Reşat Kasap and Sibel Sancak
Gazi University (Turkey)

## Abstract

Forecasting is one of the most important concepts in time series. To forecast truly the model should be determined that identifies the data set best. However one or more outliers in the model affect the parameters of the model and forecasting. In the scope of this study, firstly, time series and the outliers in time series concepts are identified. The effect of outliers is investigated on the ARMA model parameters and forecasting. For this reason, data of TUIK are used to research outliers and forecasting.

# Modelling the Nigerian Market Capitalization Using Vector Error Correction Model.

Nura Isah, Dr. Sani Ibrahim Doguwa and Basiru Yusuf
Jigawa State Polytechnic Dutse (Nigeria)

## Abstract

This research work intends to empirically develop a model for Nigerian stock Market Capitalization using Vector Error Correction Model. Forecast performance of the estimated Model is analyzed using Quarterly data on Real Gross Domestic Product (RGDP), Inflation rate (INF), Exchange Rate (EXR), Money Supply (MS) and Nigerian Stock Exchange market Capitalization (NSEC) from 1985Q1 to 2016Q3. The result for Augmented Dickey Fuller test indicates that all the variables are stationary after taking the first difference I(1). The Johansen co-integration test revealed that the variables are co-integrated, that is VEC Model is more appropriate to represent the time series data. The long run equation indicates that all the variables have a significant long run relationship with NSEC. However, for the short run equation (VEC Model) only RGDP and MS are significant in the Model while INF and EXR are insignificant to NSEC. The result for forecast performance for estimated VEC Model indicates that, the model has a Root Mean Square Forecast Error (RMSFE) of 22.05, Mean Absolute Forecast Error (MAFE) 17.65 and Mean Absolute Forecast Percentage Error (MAFPE) of 55.72%. The result for Likelihood Method (LM) test indicates that the null hypothesis of no serial correlation at lag 1 to 8 at 1% significant level, has been accepted.

# Modelling and Predicting Air Quality in Visakhapatnam using Amplified Recurrent Neural Networks

Dr. G. Lavanya Devi[1] and K. Srinivasa Rao[2]

[1] Department of CS&SE, Andhra University, Visakhapatnam, India
lavanyadevig@yahoo.co.in
[2] Department of CS&SE, Andhra University, Visakhapatnam, India
sri.kurapati@gmail.com

**Abstract.** Air quality refers to the condition of the air within our surroundings. Good air quality relates to the degree which the air is clean, clear and free from pollutants such as smoke, dust and fog among other vaporous impurities in the air. Emission of pollutants by nature and human made developments have drastically increased all over the globe in the present circumstances. Study of the pollutant concentrations, developing models to predict the future air quality has become areas of interest for the research and industry community. The air pollutants data is a temporal sequence type data and hence proper initiatives are to be taken to handle it. The system which is able to predict the concentration of air pollutants with sufficient anticipation can provide public authorities the time required to manage the emergency. Great progress has been made in the prediction of the concentration of air pollutants over the past decades by using conventional techniques. However, it is still challenging to accurately predict the concentration of air pollutants due to the complex influential factors such as meteorological parameters. On the other hand, the air quality is unique for distinct geographical locations. Hence, it is important to study, analyze and predict the air quality parameters for specific geographical area of interest. This paper aims to study and predict the quality of air at city, Visakhapatnam, India through persistent deep learning technique, amplified recurrent neural networks (ARRN) model to predict air quality. The results obtained were evaluated with the state-of-the-art models. It has been observed that the that proposed frame work significantly improves the prediction compared to widely used benchmarks models.

**Keywords:** Air Quality, Air Pollutants', Temporal Sequence Data, Prediction, , Amplified Recurrent Neural Networks (ARNN).

## 1 Introduction

Ambient Air quality is the state of the air around us. Good air quality refers to clean, clear, pure air. Clean air is essential to maintain the delicate balance of life on this planet. Ambient air quality refers to the quality of outdoor air in our surrounding environment. It is typically measured near ground level, away from direct sources of pol-

lution. Poor air quality endangers humans, animals, plants and the whole environment. It imbalances the ecological equilibrium of the planet which may further lead to a great disaster. The social and economic implications include impacts from human activities such as transport, industrialization has a direct or indirect bearing on the environment[1].

Urban air pollution poses a significant threat to human health and the quality of life of millions of people around the globe. It also places a substantial financial burden on society at large. Being able to comprehensively estimate overall urban air pollution aids air quality organizations in their decision-making and assists in the implementation of preventive actions to reduce emissions. Accurate forecasting of the air quality has become a challenging task in today's scenario. There is an increasing concern on air quality ambiance studies to identify and extract patterns for estimating and predicting pollutants' concentrations for a specific geographical area. Traditional approaches for air quality prediction use mathematical and statistical techniques. These conventional forecasting models demand a great amount of computing resources[2]. In addition, the model's accuracy depends on the model structure itself and cannot improve regardless of the amount of training data.

On the other hand, the deep learning approach has achieved outstanding results in information processing, such as speech recognition, natural language processing and computer vision. In those tasks, deep learning methods have outperformed conventional methods. Inspired by this, people are trying to use deep learning models such as Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) to perform air quality forecasting [3].

This research aims to build air quality prediction models using deep learning techniques such as Amplified Recurrent Neural Networks (ARNN) by considering pollutants' concentration levels over a period of time. ARNN models are accurate in predicting concentration levels of pollutants by extracting temporal patterns in the past data of the pollutants.

The structure of the paper is described as follows: Section II presents the literature review. Section III summarizes the theoretical background of the proposed novel architecture for predicting air quality forecasting. Implementation and performance analysis of the proposed system is presents in section IV. The last section deals conclusion and future directions.

## 2 Literature Review

Literature related to various types of mathematical, statistical, machine learning and deep learning approaches for predicting air quality is reviewed. Deep learning techniques and past applications are examined to show why these methods are likely to perform well in air quality forecasting. Figure 1 labels various types of air quality prediction models

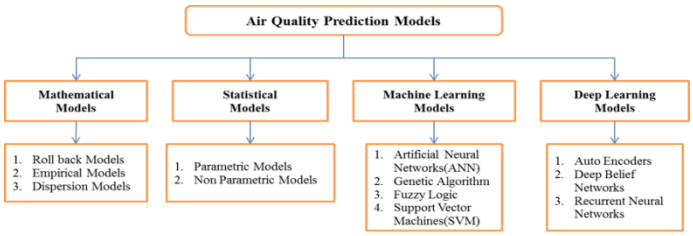



**Fig.1.** Various types of Air Quality Prediction Models

### 2.1 Air Quality Prediction Models

Air quality modelling can be viewed as the attempt to predict, by physical or numerical means, the ambient concentration of criteria pollutants found within the atmosphere of a domain. The principle application of air quality modelling is to investigate air quality scenarios so that the associated environmental impact on a selected area can be predicted and quantified [4].

There are the different type of mathematical as well as statistical models for the prediction and analysis of air quality, but machine learning models are considered to be an excellent predictive and data analysis tool for air quality forecasting. Moreover, these methods cannot draw insights from the abundant data available. To address this issue, deep learning models used to predict ambient air quality puts forward.

**Mathematical models.** Mathematical modelling attempts to predict air quality scenarios, by the intimation of mathematical and physical relationships. When these relationships become too tedious or complex to be used analytically, they are often expressed in algorithmic form and solved using computers. There are different types of mathematical models such as rollback model, empirical model and dispersion model [5].These studies suggested the mathematical models might not perform well in densely populated areas and differences in topography.

**Statistical models.** The main role of statistical models is to analyze past monitored air quality data. They are divided into parametric or linear and non-parametric or non-linear models. Linear Models as Multiple Linear Regression (MLR) can be used to make a linear empirical relationship between air pollutants and meteorological variables. The methodologies of linear and nonlinear models are explained briefly in [6]. Statistical models apply simple parameter based methods surpassing the complicated structures. Hence, these models may not reveal valuable insights into the data.

**Machine learning models.** In the machine learning models, we can note that there are two basic machine learning model techniques. One is supervised learning models, and the other one is unsupervised models. A review of the available on machine learning models literature reveals that Artificial Neural Networks (ANN) have been applied successfully to predict the air quality. According to Kostandina et al. (2018) and

Dixian et al. (2018), the neural networks are promising tools for air quality prediction when compared with other statistical models, such as regression-based models[7][8]. Moreover, the ANN, in particular, the Multi-Layer Perceptron (MLP), has a reliable performance in dealing with highly nonlinear systems such as the phenomenon of interaction between air pollutant concentrations and metrological parameters.

These studies point out that conventional methods need prior knowledge about the model structure and are based on a theoretical hypothesis. Also, they work with various data constraints. Hence, these models may not reveal valuable insights into the data.

**Deep learning models.** Deep learning approaches have emerged as powerful solutions to mitigate these limitations over conventional methods [8]. Most popular deep learning techniques are Multi-Layer Perceptron (MLP), Deep Belief Nets (DBN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Autoencoders (AE) [9], [10], [11]. Particularly, RNN based models for predicting air quality have drawn much attention in recent times. A considerable amount of literature has been published on Deep Learning techniques to predict air quality.

Studies from the fields of deep learning and air quality models show that much effort has been put into air quality forecasting, including the use of various deep learning methods. Deep learning methods have been widely used in environmental science problems and the applications of the Amplified Recurrent Neural Networks (ARNN) tend to provide some advantages over linear methods based on the results of the previous studies. In air quality forecasting, deep learning methods are promising when compared with machine learning models and other baseline models.

## 3 Overview of the Proposed System

This section provides an overview of the proposed amplified recurrent neural network model for predicting air quality in city Visakhapatnam. It presents a step-by-step process followed for the development of the proposed system.

### 3.1 Design of the Proposed Framework

Real world data is often highly *incomplete*, *noisy* and *inconsistent* (dirty) as they originate from multiple, heterogeneous sources [12]. Quality decisions can be drawn only from clean data. Hence, data has to be preprocessed to set it free from dirt for decision making. Uncertainty arises due to lack of knowledge or insufficient information. Decision making under uncertainty is a challenging task in any domain. Application of deep learning techniques for handling uncertainty in air quality data may reveal new insights in prediction.

To address this problem, an amplified recurrent neural network model is proposed, that imputes the data and builds a model that can predict the air quality of a given dataset. The framework of the entire process is depicted in Figure 2. Ten-Fold cross-

validation is used to avoid overfitting of the model. 70% of the data is for training and 30% is used for testing.

### 3.2 Description of the Proposed Framework

This section gives a detailed description of the architecture depicted in Figure 2. The proposed framework divided into four modules such as air quality data preprocessing, feature selection, proposed amplified recurrent neural network models and evaluation system.

**Module1: Air Quality Data Preprocessing.** The proposed model is trained with the real-time data collected from the Central Pollution Control Board, of the city Visakhapatnam. Data obtained from the source as mentioned earlier has to be preprocessed as noisy data affects the performance of the forecasting model. Missing values add more noise to data[12]. For the proposed model, missing values were imputed by the mean values of the respective parameter [13].

**Module 2: Feature Selection.** Pollutant concentration values and meteorological parameters are the features of the air quality dataset. The intuition is to derive inferences from sequential features to better represent data. For this, one pollutant concentration and meteorological parameters consider the feature selection process.

**Module3: Proposed Amplified Recurrent Neural Network models.** The postulate of this model is, given a temporal sequence data of pollutant concentration values and meteorological parameters of city Visakhapatnam. the model will capture the dependencies in the data and predicts the next hour pollutant concentrations. Given, meteorological parameters $m = \{m_1, m_2, m_3, m_4, m_5, m_6\}$ and pollutant concentrations $x_t$, t = 1...T as a paired input $X= \{(m, x_t)\}$, the objective of the proposed model is to recognize patterns and predict $x_{t+1}$.

This study leverages Amplified Recurrent Neural Networks (ARNNs) and its variants for modelling and forecasting concentration values of the pollutants.

**Module4: Evaluation System.** The dataset randomly divided into two groups: a training set that contains 70% of the original dataset, and the remaining 30% used as a test set for the models. The division dataset into training and test sets might be sensitive to the randomly selected pollutants/ meteorological parameters. Therefore, to ensure that evaluation is not vulnerable to the randomness of the division step, we ran the models 10 times, each time with a different partitioning. Cross-validation is adopted to minimize the bias at the training phase.

The system flow diagram of the proposed system is shown in Figure 2. The inputs of the proposed system are the records of the pollutant concentration and meteorological parameters over the last 24 hours from Visakhapatnam air quality dataset. The output is the pollutant concentration of the next hour.

This section has given an overview of the proposed system and the research design adopted. It briefly explained all the components of the system developed. Finally, it

has given the rationale of choice of the dataset used for evaluating the system. If the air quality dataset does not have ambiguous, this model may produce accurate and efficient results.

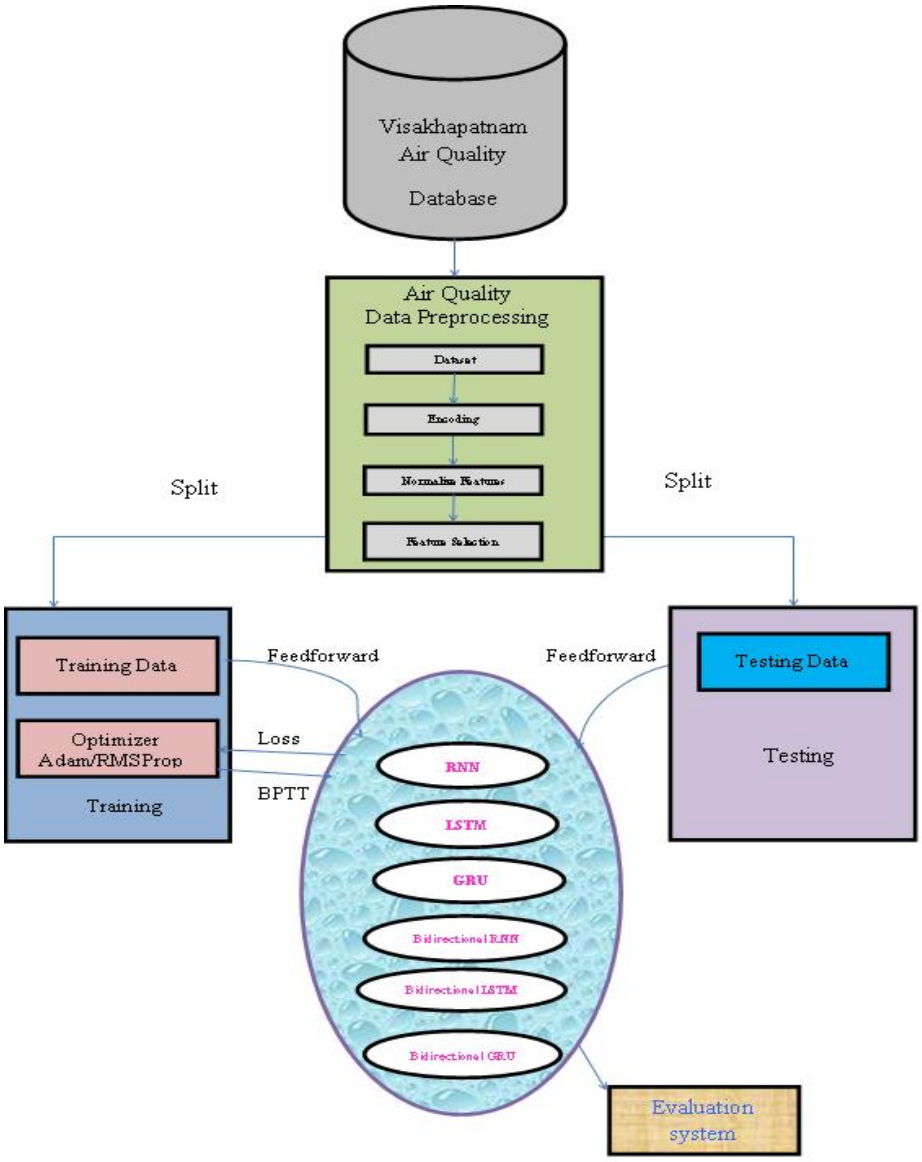



**Fig.2.** Architecture of the Proposed System

# 4 Experimental Evaluation of ARNN Models

Experimental procedures adopted for the evaluation of the proposed models and analysis of the results is shown in this section. The proposed approaches are validated with empirical evaluations against several state-of-the-art methods on real-world data.

The objective of this research work is to develop a framework to predict the air pollutant concentration of the next hour from the past air quality data. Performance obtained by RNN, LSTM, GRU, and Bidirectional (BI-RNN, BI-LSTM, BI-GRU) models have to be evaluated empirically. The main goal of this experimental analysis is to compare the performance of the proposed methods with the state-of-art existing methods using Visakhapatnam air quality dataset[14].

## 4.1 Model Evaluation Measures

Several statistical scores were used to evaluate the performance of proposed models, including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE) and Coefficient of Determination ($R^2$) [15].

## 4.2 Models Training and Implementation

The models are trained with the real-time data collected from the Central Pollution Control Board (CPCB), of the city Visakhapatnam for the period July 1, 2016, to May 17, 2018 [14]. 70% of the dataset was considered for training and the remaining 30% is utilized for testing the models. Cross-validation is adopted to minimize the bias at the training phase.

All the algorithms were implemented in Python platform, using TensorFlow backend. The experimentations were run under Linux OS on a machine with 3.30 GHz Intel Core i5-4590 processor, 8 GB RAM and Intel HD Graphics 4600 card. We implemented all code (ARNN models) using Python, Theano, Keras and Scikit-learn frameworks [16] and executed.

## 4.3 Experimental Results Analysis

Models performance on testing data was evaluated by comparing the six proposed models (RNN, LSTM, GRU, BI-RNN, BI-LSTM, and BI-GRU) with SVR linear, poly and RBF kernels [7], [17][18]. For the ease of comparison of results, each and every pollutant concentration has been taken up for all models. The detailed results for each and every pollutant in the air quality data set of city Visakhapatnam can be found in next subsections.

**Pollutant wise Results Analysis.** Here we analyze the results of each and every pollutant concentration of the air quality dataset of city Visakhapatnam.

*$PM_{2.5}$*. $PM_{2.5}$ pollutant concentration prediction error statistics are shown in table 1. The RMSE for all the models such as RNN, LSTM, GRU, BI-RNN, BI-LSTM, BI-GRU, SVR-RBF, SVR-POLY, and SVR-LINEAR range varies from 17.897 to 30.009. The MSE for all models ranges varies from 320.316 to 900.540. The normalized error MAE varies from 8.367 to 24.554. The Pearson coefficient of determination $R^2$ range varies from 0.082 to 0.749.

According to the correlation and normalized error, models have the best performance with small error and highest correlation attained by the LSTM model. All proposed models better performance than SVR models. Figure 3 demonstrates the forecasting score of various models. SVR-POLY model performed poor than all other proposed models.

**Table 1.** PM $_{2.5}$ prediction error statistics

| $PM_{2.5}$ | | | | | |
|---|---|---|---|---|---|
| Sno | MODEL | RMSE | MSE | MAE | $R^2$ |
| 1 | RNN | 18.556 | 344.334 | 9.729 | 0.730 |
| 2 | **LSTM** | **17.897** | **320.316** | **8.367** | **0.749** |
| 3 | GRU | 18.328 | 335.907 | 9.088 | 0.736 |
| 4 | BI-RNN | 22.722 | 516.301 | 16.054 | 0.595 |
| 5 | BI-LSTM | 18.002 | 324.062 | 8.555 | 0.746 |
| 6 | BI-GRU | 18.766 | 352.167 | 10.011 | 0.724 |
| 7 | SVR-RBF | 25.722 | 661.604 | 19.179 | 0.326 |
| 8 | SVR-POLY | 30.009 | 900.540 | 24.554 | 0.082 |
| 9 | SVR-LINEAR | 19.922 | 396.868 | 13.913 | 0.596 |

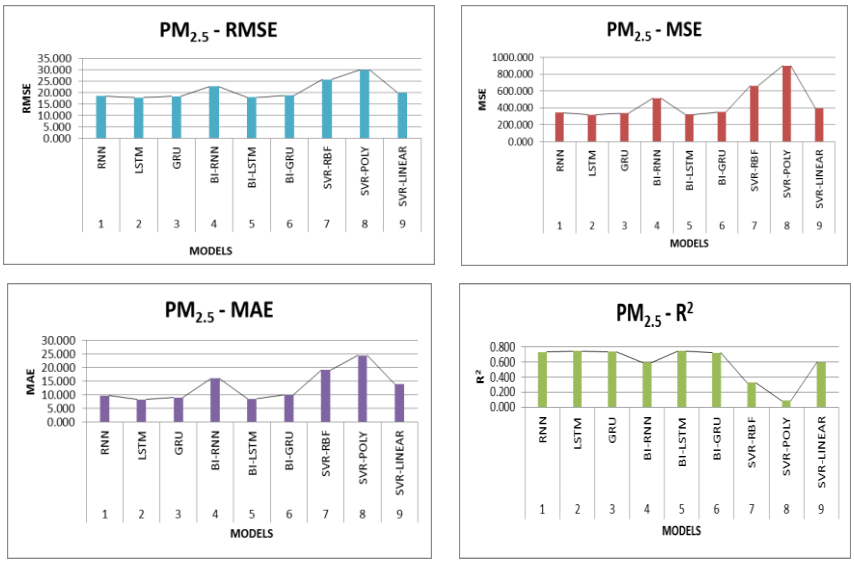



**Fig.3.** PM $_{2.5}$ forecast scores of various models

*NO*. NO pollutant concentration prediction error statistics are shown in table 2. The RMSE for all the models such as RNN, LSTM, GRU, BI-RNN, BI-LSTM, BI-GRU, SVR-RBF, SVR-POLY, and SVR-LINEAR range varies from 19.853 to 39.966. The MSE for all models ranges varies from 394.144 to 1597.269. The normalized error MAE varies from 6.593 to 37.912. The coefficient of determination $R^2$ range varies from -2.029 to 0.276.

According to the correlation and normalized error, models have the best performance with small error and highest correlation attained by the BI-LSTM model. All proposed models better performance than SVR models. Figure 4 demonstrates the forecasting score of various models. SVR-LINEAR model performed poor than all other proposed models.

**Table 2.** NO prediction error statistics

| NO | | | | | |
|---|---|---|---|---|---|
| Sno | MODEL | RMSE | MSE | MAE | $R^2$ |
| 1 | RNN | 20.574 | 423.301 | 9.247 | 0.223 |
| 2 | LSTM | 19.916 | 396.652 | 6.700 | 0.272 |
| 3 | GRU | 20.248 | 409.992 | 6.674 | 0.247 |
| 4 | BI-RNN | 20.914 | 437.411 | 8.144 | 0.197 |
| 5 | **BI-LSTM** | **19.853** | **394.144** | **6.593** | **0.276** |
| 6 | BI-GRU | 20.526 | 421.305 | 8.006 | 0.227 |
| 7 | SVR-RBF | 36.019 | 1297.380 | 34.157 | -1.460 |
| 8 | SVR-POLY | 35.257 | 1243.038 | 33.322 | -1.357 |
| 9 | SVR-LINEAR | 39.966 | 1597.269 | 37.912 | -2.029 |

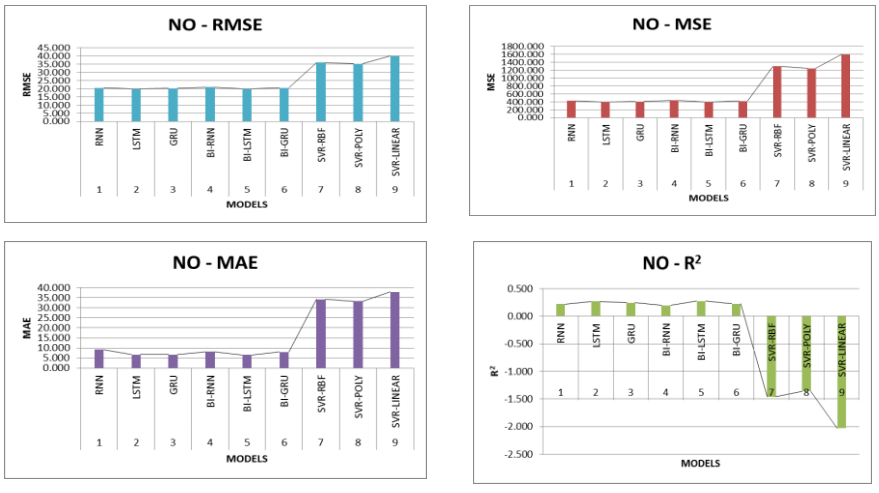



**Fig.4.** NO forecast scores of various models

The study has been carried out by considering 12 air pollutant concentrations. However, due to space constraints, the results of only two pollutants have been displayed.

### 4.4 Summary of Results

As far as results are concerned, on the bottom line, ARNN modelling approaches have provided decent and comparable results. The results of the proposed ARNN modelling approaches justify further development and applications of these methods to air quality data of city Visakhapatnam.

## 5 Conclusion

In this paper, various Amplified RNN models specific to air quality prediction in Visakhapatnam have been studied and their methodology and significance were investigated. The purpose of the current study is to determine an efficient forecasting model for hourly concentration levels of air pollutants. This work proposed RNN, LSTM, GRU and Bidirectional RNN, LSTM, GRU models that perform modelling to predict pollutant concentrations, given temporal sequence data as input. Real-time data of the city Visakhapatnam having 12 air pollutant values is considered for experimental analysis. The outcomes of the proposed framework agree to the idea that deep learning-based techniques for forecasting air quality achieve promising performance over conventional strategies. Also, ARNNs can better work with time-series data. The present work was carried out by considering single air pollutant values and meteorological parameters at a time for all the 12 pollutants chosen. This work can be extended by performing multivariant modelling of the 12 pollutants simultaneously. Also, the proposed models can be extended by convolutional neural networks. However, in the experiments of the proposed approaches; the sample only takes the data from the GVMC, Visakhapatnam monitoring station.

**Acknowledgements**

This Publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

## References


1. Franco DiGiovanni and Philip Fellin.:Transboundary Air Pollution: Environmental Monitoring, ©Encyclopedia of Life Support Systems (EOLSS).
2. Niharika, Venkatadri, M. and Padma, S. Rao.:A survey on Air Quality forecasting Techniques: International Journal of Computer Science and Information Technologies, vol. 5 (1), pp. 103-107(2014).
3. Athira, V., Geetha, P., Vinayakumar, R., Soman, K. P.: DeepAirNet- Applying Recurrent Networks for Air Quality Prediction: Elsevier, Procedia computer science, vol 132, pp 1394-1403(2018).
4. Ganganjot Kaur Kang, Jerry Zeyu, Sen Chiao, Shengqiang Lu, and Gang Xie.:Air Quality Prediction: Big Data and Machine Learning Approaches: International Journal of Environmental Science and Development, vol. 9(2018).
5. Richard, S. Collett and Kehinde Oduyemi.:Air quality modelling: a technical review of mathematical approaches: Meteorol. Appl. 4, pp. 235–246(1997).
6. M. K. Reddy, K. G. Rama Rao and I. Rammohan Rao.:Air Quality Status of Visakhapatnam (India) –Indices Basis: Environmental Monitoring and Assessment 95: 1–12(2004).
7. Kostandina Veljanovska and Angel Dimoski.:Air Quality Index Prediction Using Simple Machine Learning Algorithms: International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), ISSN 2278-6856, vol. 7, Issue 1(2018).
8. Zhongang Qi et al..:Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality: In CoRR abs/1711.00939,arXiv: 1711.00939(2017).
9. Ian Goodfellow, YoshuaBengio and Aaron Courville.:Deep Learning: MIT Press(2016).
10. J. Schmidhuber.:Deep Learning in Neural Networks An Overview(2014).
11. Y. LeCun, Y. Bengio, and G. Hinton.:Deep Learning: Nature, vol. 521, pp. 436-444(2015).
12. Jiawei Han, Micheline Kamber, and Jian Pei.:Data Mining: Concepts and Techniques: 3rd Edition, The Morgan Kaufmann Series in Data Management Systems(2011).
13. M. Sujatha, G. Lavanya Devi, K. Srinivasa Rao, and N. Ramesh.:Rough Set Theory Based Missing Value Imputation: Cognitive Science and Health Bioinformatics. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore, Online ISBN 978-981-10-6653-5(2018).
14. https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing
15. Xiang Li et al..:Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation: Elsevier Ltd., Environmental Pollution 231, pp. 997-1004(2017).
16. Rami Al-Rfou et al..:Theano: A Python framework for fast computation of mathematical expressions: arXiv preprint arXiv: 1605.02688(2016).
17. Nieto P.G, Combarro E.F, Del Coz Diaz J.J.:A SVM- based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study: Appl.Math.Comput, pp 8923-8937(2013).
18. K Srinivasa Rao, G. Lavanya Devi, N. Ramesh, "Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.2, pp.18-24(2019).

# Environmental policies analysis for CO2 emission reduction: evidence across countries 1980-2014

Yi Zheng* and Dess Pearson**

*Department of Political and Economic Studies, University of Helsinki
**Torrens University Australia

**Abstract.** We employ a difference-in-difference regression model to analyse the efficiency of carbon taxation policy and its interaction with an Emission Trading System (ETS). We provide evidence for 26 of the most developed countries in the world in the period 1980-2014. Following the work of Dynarski [2004], Cameron et al. [2008, 2012] and Hoechle [2007], we compare the results under three methods of regression analysis. We confirm that carbon taxation has preformed well in the 35 years under observation in that they efficiently control CO2 emissions. We find that the longer the duration a country uses carbon taxation, the greater the reduction in CO2 emissions. For the countries that use both an ETS and carbon taxation, we find an even more efficient CO2 emission reduction. The results are robust to heteroskedasticity, autocorrelation and cross-sectional dependence.

**Keywords:** Carbon Pricing Instruments, CO2 Emission Reduction, Difference-in-Difference Model, Cross-sectional Dependence, Few Clusters

**JEL classification:** C3, H23, Q5

## 1 Introduction

Carbon taxation and Emissions Trading Systems (ETS) are generally considered to be efficient environmental policies to reduce CO2 emissions. However, in 2017, only 42 national and 25 subnational jurisdictions had some form of carbon pricing — either through an emissions trading scheme or a carbon tax. Instead, more countries have been using nuclear or renewable energies such as wind, solar, geothermal and biomass in the past decades to reduce CO2 emissions. Finland and Sweden are two of the very few countries that have both implemented a carbon tax and an ETS scheme (as both are in the EU-ETS Scheme). Moreover both countries also use nuclear and non-renewable energy sources. Data from these two countries demonstrate a clear and dramatic reduction in CO2 emissions. However, it is difficult to conclude whether the emission reduction results from carbon pricing instruments, the lower consumption of energy, intensity improvement or GDP reduction. In this article, we focus on carbon pricing instruments. We test the treatment effects of carbon taxation policy and a joint use of both taxation and ETS. We then determine whether they efficiently control CO2 emissions and which one has performed better. Our aim is to provide empirical evidence to encourage and support the implementation of carbon pricing instruments.

Many researchers have provided evidence of the effectiveness of carbon pricing instruments, both theoretically and empirically. For example, Newell et al. [2013] provide a clear summary of the carbon market performance in several countries. The existing empirical literature usually chooses a short time horizon. In this article, we try to include most developed countries (that satisfy our countries selection criteria for the difference-in-difference model) as possible and at the same time use a long time horizon to support our analysis. We employ a difference-in-difference regression model to analyse the efficiency of carbon taxation policy and its interaction with an ETS, as evidence for the most developed countries in the world over the period 1980-2014. As of December 2014, 17 countries had introduced carbon taxation as part of their emission reduction strategies. We include 11 of them – 64.1% in our analysis. The six carbon taxation users that are excluded are Latvia, Estonia, Slovenia, Iceland, Poland and Mexico. We drop the first four countries due to insufficient data.[1] We drop Poland and Mexico because of their lower levels of economic development. It finally leaves us with a 35-year panel data series across 26 countries. We include the shares of consumption of the different forms of energy usage as well as commonly used measures such as Gross Domestic Product based on Purchasing Power Parity per capita (GDP on PPP) , energy intensity (EI) as control covariates following Zakarya et al. [2015], Bruvoll and Larsen [2004], Lin and Li [2011], Scrimgeour et al. [2005], Song et al. [2015], Meng et al. [2013], Doda et al. [2012] and Liu et al. [2015], just list a few. To deal with few clusters, we use a country specific bootstrap following the work of Cameron et al. [2008, 2012] and make adequate use of bootstrap replications.[2]. We find evidence of cross-sectionally strong mixing, although every country's environmental policy setting and energy consumption is, in principle, independent.

* Corresponding author.

[1] Hoechle's (2007) approach can handle missing values problem. However, the authors prefer the balanced data.

[2] 3000 bootstrap replications are performed in our main results. In robustness check, we report the results by using 500, 1000 and 2000 replications.

We explain the correlation between the countries as the neighbourhood effects. (see Tanguay et al. [2004])[3] We correct for the cross-sectional dependence following Hoechle [2007]. To test the sensitivity of the choice of treatment and control groups, we follow Dynarski [2004]'s approach. We then compare the results under three classic methods. We find the treatment effects of carbon pricing instruments are statistically significant. The coefficients of interest in the regressions exhibit the 95% confidence intervals away from zero with small p-values. Although the results differ depending on the choice of covariates, we suggest that carbon taxation has preformed well in the past 35 years. It efficiently controls CO2 emissions. We also find the evidence that the longer the duration a country uses carbon taxation, the greater the reduction in CO2 emissions. We provide evidence to confirm the significant effectiveness of carbon pricing instruments for both a carbon tax and an ETS. For the countries that use both instruments, we find an even more efficient CO2 emission reduction process.

With the signing of the Paris Agreement on CO2 reductions by most countries in December 2015, the results of this study will have considerable implications for policymakers in the years ahead as they find ways to implement the commitments made to emission reductions. Indeed, the findings of this research have clear implications for countries that made commitments to cut their CO2 emissions; that is, the use of market-based instruments such as an ETS and especially carbon taxation are effective mitigation methods. The important terms in this articles are listed in the box.

**Fig. 1:** Definitions and measures

1. Eenergy Intensity (EI) = $\frac{\text{Total primary energy consumption}}{\text{Population}*\text{Total GDP}}$
2. Total primary energy consumption = Total renewable energy consumptionTotal non-renewable energy consumption
3. Share of each energy consumption = $\frac{\text{Consumption of each energy}}{\text{Total primary energy consumption}}$
4. $CO2_{it}$: carbon dioxide emissions per capita per year from fossil fuel use and cement production excluding short-cycle biomass burning (for example, agricultural waste burning) and excluding large-scale biomass burning (for example, forest fires) (Million ton CO2 per year)
5. GDP on PPP: gross domestic product based on purchasing-power-parity valuation of country GDP (Current international dollar, Billions)
6. GDP: gross domestic product, current prices (U.S. dollars , Billions)
7. Coal consumption: commercial solid fuels only, i.e. bituminous coal and anthracite (hard coal), and lignite and brown (sub-bituminous) coal, and other commercial solid fuels. Excludes coal converted to liquid or gaseous fuels, but includes coal consumed in transformation processes. (Million tonnes oil equivalent)
8. Natural gas consumption: Excludes natural gas converted to liquid fuels but includes derivatives of coal as well as natural gas consumed in Gas-to-Liquids transformation. (Million tonnes oil equivalent)
9. Hydroelectricity consumption: Based on gross primary hydroelectric generation and not accounting for cross-border electricity supply. Converted on the basis of thermal equivalence assuming 38% conversion efficiency in a modern thermal power station. (Million tonnes oil equivalent)
10. Consumption of nuclear, solar, wind, and geothermal, biomass and other waste: Based on gross generation and not accounting for cross-border electricity supply. Converted on the basis of thermal equivalence assuming 38% conversion efficiency in a modern thermal power station. (Million tonnes oil equivalent)

The rest of the article is organised as follows. In Section, 2 we discuss the setting and underlying assumptions of the model. The Data description is provided in Section 3. We report the regression outputs in Section 4 and provide the robustness checks in Section 5. Section 6 concludes the article.

## 2 Model

We test the efficiency of CO2 reduction through the use of carbon pricing instruments — carbon taxation in particular and its interaction with an ETS. Let $N$ be the total number of countries in our data, $N_1$ be the number of treatment groups that have and $N_0$ be the number of control groups that have not implemented the environmental policy during the years 1980-2014. We employ a linear panel data model with time and entity fixed effects. This is expressed as follows:

$$CO2_{it} = \alpha d_{it} + \mathbf{x'}_{it}\beta + \theta_i + \delta_t + \varepsilon_{it} \tag{1}$$

where $CO2_{it}$ represents carbon dioxide emissions per capita per year from fossil fuel use and cement production excluding short-cycle biomass burning (for example, agricultural waste burning) and excluding large-scale biomass burning (for example, forest fires). $CO2_{it}$ is the dependent variable, where $i$ represents country and $t$ represents year; $d_{it}$ is the environmental policy dummy variable whose coefficient $\alpha$ is the object of interest in this study. $d_{it}$ equals one if the environment policy of interest is in effect and zero otherwise; $\mathbf{x'}_{it}$ represents a vector of independent variables with parameter vector $\beta$ including gross domestic product at purchasing power parity per capita per year, energy intensity[4] and the shares of energy consumption of renewables and non-renewables to the primary energy; $\theta_i$ and $\delta_t$ are country specific and time specific effects respectively; $\varepsilon_{it}$ is the error term under different assumptions.

**Assumption 1.** *The standard error is assumed to be heteroskedastic and autocorrelated. The panel data are assumed to be cross sectionally (spatically) uncorrelated. There is no temporal variation in the environmental policy dummy variable* $\widetilde{d}_{it}$.

[3] Martén (2014) suggests that there would be benefits to neighbouring countries to harmonise their energy policies. See the Wall Street Journal `https://blogs.wsj.com/experts/2014/10/02/neighboring-countries-should-harmonize-energy-policies/`

[4] Energy intensity (EI) is calculated as the amount of energy a country needs to generate a unit of gross domestic product (GDP), while energy consumption per capita represents total primary energy consumption divided by the population of the country.

It is an assumption widely used for the case when time $T$ is fixed and the numbers of both the treatment and control group are large. However, the number of countries that implement carbon pricing instruments is small. To deal with the few clusters issue, we follow Cameron et al. [2008, 2012] and Cameron and Miller [2010] and use a sufficient number of bootstrap replications. Next, we take account of general forms of cross-sectional dependence and analyse complex patterns of mutual dependence in the panels.

**Assumption 2.** *The standard error is assumed to be groupwise heteroskedastic, autocorrelated up to some lag length,*[5] *and cross-sectional (spatial) and temporal dependent of general forms, i.e., Driscoll and Kraay (1997) standard error.*

We can thus rewrite the model as follows:

$$\widetilde{CO2}_{it} = \alpha \widetilde{d}_{it} + \widetilde{\mathbf{x}}'_{it}\beta + \widetilde{\varepsilon}_{it} \tag{2}$$

where the country-year random effects are not averaged away. By regressing $\widetilde{CO2}_{it}$ on $\widetilde{d}_{it}$ and $\widetilde{\mathbf{x}}'_{it}$, we can obtain an estimation of $\alpha$.

We suggest the existence of the neighbourhood effects for policy implementation and energy use: countries that are geographically located close by or in the same region seem to have similar environmental policies. Not unexpectedly, a country's policy potentially has an impact on its neighbouring countries. Some groups of countries that are part of trade blocs (for example, EU countries) lend themselves considerably to the neighbourhood effects of government policies, including carbon pricing and sometimes even the selection of nuclear and renewable energy. It stands to reason that neighbouring countries that trade with each other to a greater degree are more likely to harmonise their environmental policies. Besides the EU-ETS which is applied to all EU countries, similar ties apply to APEC and OECD members. Some environmental regulations are applied to all the members. Another cause of such effects is the geographic nature: neighbouring countries might share similar natural resources and therefore the consumption of energy. The claim of the neighbourhood effects is supported by our data in Table (1). We find the implementation of carbon pricing instruments and the consumption of renewables and non-renewables are more alike regionally. An obvious example are the four Nordic countries. Furthermore, for close neighbouring countries, we often find great similarity pairwise., e.g., Australia and New Zealand, USA and Canada, all of which have similar environmental policies. We therefore correct for cross-sectional dependence because of the neighbourhood effects in environmental policy and in energy use.

## 3 Data description

The variable we are interested is the country specific yearly CO2 emission reduction through the use of carbon pricing instruments. We use the shares of consumption of renewable and non-renewable energy sources as the control covariates[6] as well as commonly used measures such as GDP on PPP and energy intensity. The time series data of CO2 emission totals cross countries is obtained from the Emission Database for Global Atmospheric Research (EDGAR), European Commission's Joint Research Centre (JRC).[7] The country-specific CO2 emission totals exclude short-cycle biomass burning (such as agricultural waste burning) and large-scale biomass burning (such as forest fires). The data of national and subnational carbon pricing instruments is compiled from OECD Economic Surveys,[8] International Carbon Action Partnership (ICAP) and World Bank Group.[9] The annual energy consumption data is from BP Statistical Review of World Energy.[10] The data of annual GDP on PPP and population is extracted from the International Monetary Fund (IMF), OECD[11] and World Bank Group databases. To satisfy the common trend assumption in our difference-in-difference model, we keep only the most developed countries following the IMF's criteria for advanced economies[12], World Bank high-income economies[13] and High-income OECD members[14].

The original dataset includes 207 countries. As of December 2014 the number of countries that implemented either carbon taxation or an ETS was 17 and 37 respectively. Eleven countries used both carbon pricing instruments. They are Canada, Denmark, Finland, France, Republic of Ireland, Japan, Norway, Poland, Sweden, Switzerland and United

[5] The selection of the lag length of $floor[4(T/100)^{2/9}]$ follows Newey and West (1994).

[6] We include energy consumption share to the primary energy instead of energy consumption because of 'bad control' problem. For example, one can argue that carbon tax reduces the consumption of coal, while the reduction of coal consumption also reduces CO2 emission. We test the effect of carbon taxation on the energy consumption by regressing CO2 emission on the yearly consumption of each energy source as well as its share on the primary energy. We find that the direct effect of carbon tax on energy consumption share is much less. The regression outputs are given in the appendix.

[7] See: Trends in global CO2 emissions: 2015 Report by PBL Netherlands Environmental Assessment Agency and the European Commission's Joint Research Centre.

[8] See: OECD Economic Surveys: Poland 2012, Issue 7, Volume 2012.

[9] See: State and Trends of Carbon Pricing 2014 and 2015 by World Bank Group.

[10] See: Statistical Review of World Energy 2015 and 2016 by BP.

[11] See: World Economic and Financial Surveys by IMF and Economic Surveys by OECD.

[12] See: IMF Advanced Economies List. World Economic Outlook, April 2016, p. 148

[13] See: Country and Lending Groups by World Bank Group. Accessed on August 1, 2016

[14] See: Members and partners by OECD. Retrieved 1 August 2016

Kingdom. The carbon pricing instrument users are mostly developed countries. To use a difference-in-difference (DID) model to analyse the effect of carbon pricing policies on CO2 emission, we need to choose the closest matched countries. For this reason, we check each country's yearly CO2 emission, GDP preformance and energy intensity growth and drop the developing and the least developed countries.

We then test the consumption of non-renewable energies (including coal, hydroelectricity, natural gas, nuclear and oil) and renewable energies (including solar, wind, geothermal, biomass and other) in each individual country. We drop the consumption of oil due to multicollinearity. It finally leaves us a panel data across 26 representative countries over the time period 1980-2014. In terms of carbon taxation, the number of the treatment group is 11 representing 17 countries who are the real carbon taxation users as of December 2014. Ten out of these eleven countries have implemented both a carbon taxation and an ETS.

A brief summary of the 35-year panel data across 26 countries in 4 regions (grouped by geographic location) is shown in Table (1): Asia Pacific (7), Europe (16), Middle East (1) and North America (2). It summaries the starting date of the policies that were implemented and their length. Note that an interruption in the continuous use of some energy exists. That is, during the 35-year period, some countries may have stopped using some types of energy and switched to others, for example, for the sake of seeking a more efficient solution of CO2 reduction. Such interruption does not apply to the continuous implementation of carbon pricing instruments. Table (1) includes carbon pricing instruments, renewable energies and nuclear energy consumption in use from 1980 to 2014 which are all widely considered as 'environmental friendly' approaches to reduce CO2 efficiently. As briefly mentioned earlier, we find that countries that are located closely are more likely to design similar environmental policies (and energy use). By the same token, the environmental policies (and energy use) in countries further apart are less alike.

**Table 1:** Starting year and length of the use of carbon pricing instruments and energies of interest as of December 2014[a]

| Region (6) | Country | ETS | Carbon Tax | Renewables[b] | | | Nuclear |
|---|---|---|---|---|---|---|---|
| | | | | Geothermal, Biomass and Other | Solar | Wind | |
| Asia Pacific (7) | Australia | | 2012,3[c] | 1980,35 | 1991,24 | 1993,22 | |
| | Hong Kong SAR, China | | | 2010,5 | 2010,5 | 2006,9 | |
| | Japan | 2010,5 | 2012,3 | 1980,35 | 1990,25 | 1993,22 | 1980,34 |
| | New Zealand | 2008,7 | | 1980,35 | 2007,8 | 1992,23 | |
| | Korea, Republic of | | | 1995,20 | 1991,24 | 1994,21 | 1980,35 |
| | Singapore | | | 1986,29 | 2009,6 | | |
| | Taiwan, Province of China | | | 1982,33 | 2000,15 | 2000,15 | |
| Europe (16) | Austria | 2005,10 | | 1980,35 | 1993,22 | 1995,20 | |
| | Belgium | 2005,10 | | 1980,35 | 2004,11 | 1987,28 | 1980,35 |
| | Denmark | 2005,10 | 1992,23 | 1983,32 | 1996,19 | 1980,35 | |
| | Finland | 2005,10 | 1990,25 | 1990,25 | 1991,24 | 1992,23 | 1980,35 |
| | France | 2005,10 | 2014,1 | 1980,35 | 1992,23 | 1990,25 | 1980,35 |
| | Germany | 2005,10 | | 1980,35 | 1990,25 | 1986,29 | 1980,35 |
| | Greece | 2005,10 | | 1992,19 | 2004,11 | 1987,28 | |
| | Ireland, Republic of | 2005,10 | 2010,5 | 1996,19 | 2009,6 | 1992,23 | |
| | Italy | 2005,10 | | 1980,35 | 1989,26 | 1989,26 | 1980,8 |
| | Netherlands | 2005,10 | | 1980,35 | 1992,23 | 1986,29 | 1980,35 |
| | Norway | 2005,10 | 1991,24 | 1985,30 | 2010,5 | 1999,16 | |
| | Portugal | 2005,10 | | 1980,35 | 1989,20 | 1989,26 | |
| | Spain | 2005,10 | | 1980,35 | 1989,26 | 1990,26 | 1980,35 |
| | Sweden | 2005,10 | 1991,24 | 1980,35 | 1993,22 | 1983,32 | 1980,35 |
| | Switzerland | 2008,7 | 2008,7[d] | 1980,35 | 1990,25 | 1996,19 | 1980,35 |
| | United Kingdom | 2005,10 | 2013,2 | 1990,25 | 1984,20 | 1989,26 | 1980,35 |
| Middle East (1) | Israel | | | 2008,7 | 2009,6 | 2001,14 | |
| North America (2) | Canada | 2007,8 | 2008,7 | 1980,35 | 1992,23 | 1985,30 | 1980,35 |
| | United States of America | 2009,6 | | 1980,35 | 1983,32 | 1983,32 | 1980,35 |

[a] Source: authors own compilation from OECD Economic Surveys, International Carbon Action Partnership (ICAP), World Bank Group and BP Statistical Review of World Energy.
[b] The consumption of renewable energy includes solar, wind, geothermal, biomass and other waste.
[c] The statistics in each cell show as "the starting year of usage" and "the length of use in the period of 1980-2014". Carbon pricing instruments, renewable or nuclear energy shall be used for more than one day in each year. Blank cells means no record for using by December 2014.
[d] The Swiss ETS started with a five-year voluntary phase as an alternative option to the CO2 levy on fossil fuels in 2008. From 2013, companies who participant in the ETS are exempt from carbon tax. Source: Swiss ETS - International Carbon Action Partnership

### 3.1 Subnational jurisdictions

Many subnational jurisdictions have implemented carbon pricing policies such as Québec, California and Tokyo. For our analysis, we use country level data. If subnational carbon pricing instruments (ETS and/or carbon tax) are implemented, for simplicity of model, we consider the instruments to be national. For example, due to the implementation of Alberta SGER (2007 - now), British Columbia carbon tax (2008 - now) and Québec CaT (2013 - now), we consider Canada, as an entity in our model, a country using both carbon taxes and Emissions Trading System. This rule applies to three countries as shown in Table (2): Canada, USA and Japan. The only excluded country is China which started to use a city-level Pilot ETS since 2013.

### 3.2 Trend of CO2 and GDP preformance

From the Stata graphs, countries show a very similar trend of yearly CO2 emission per capita. Moreover, the common trend of neighbouring countries such as Australia and New Zealand; EU countries including UK (almost all of them are X sharped), Hong Kong and Singapore; Canada and USA; Japan, Taiwan, Korea are clearer. The exemption of our

**Table 2:** Subnational carbon pricing instruments in operation

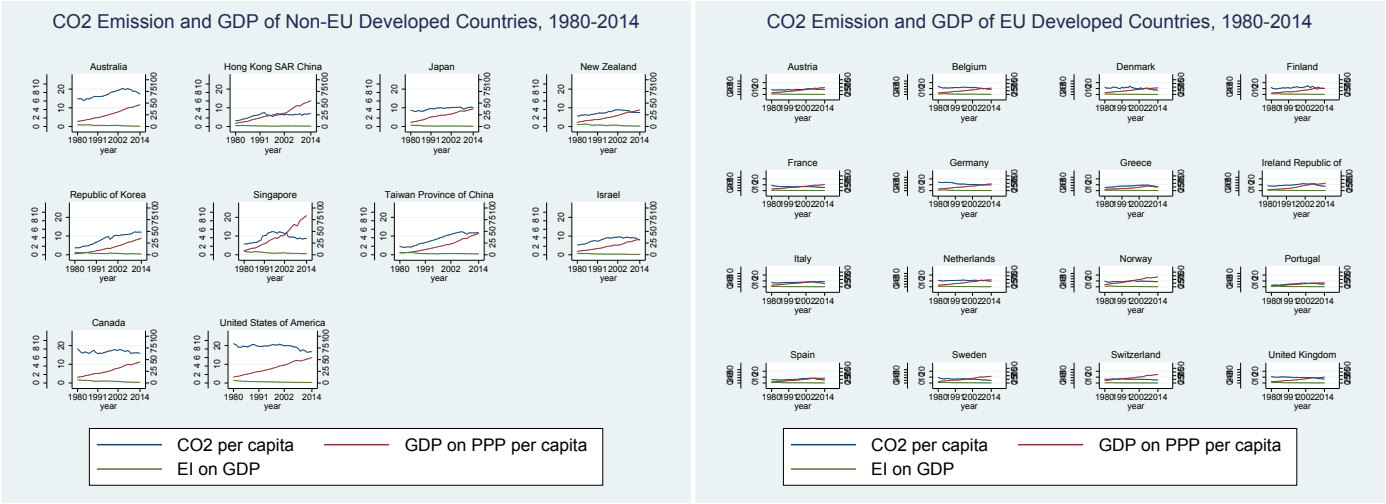
| Country | Carbon Pricing Instruments | |
|---|---|---|
| | ETS | Carbon Tax |
| Canada | Alberta SGER (2007 - now) | British Columbia carbon tax (2008 - now) |
| | Québec CaT (2013 - now) | |
| USA | RGGI (2009 - now) | |
| | California CaT (2012 - now) | |
| Japan | Tokyo CaT (2010 - now) | National (2012 - now) |
| | Saitama ETS (2011 - now) | |
| | Kyoto ETS (2011 - now) | |

Source: World Bank Group, State and Trends of Carbon Pricing, 2015

pairwise/region comparison is Greece — because of the severe economic recession it is facing in the recent years. As CO2 emissions are closely correlated with GDP growth, some can argue that Greece's CO2 reduction is from its sharp GDP reduction rather than environmental policies. However, before its recession, the common trend of CO2 emission can still be found.

We first look at the GDP performance and CO2 emission trend of non-EU countries 1980-2014 from Figure (2). Ten

**Fig. 2**




non-EU countries are all APEC countries except Israel. Australia started to use a carbon tax since 2012 but ended the policy just 3 years later in 2015.[15] New Zealand started to use an ETS from 2008. We can see a clear trend that after the policy implementation in both countries, CO2 emissions showed a gradual reduction. This is due, in no small measure, to the carbon pricing instruments used. Canada is one of the countries that use both carbon pricing instruments — starting tax from 2007 and ETS from 2008. The USA (California) started using an ETS from 2009. Both Canada and the USA have only implemented carbon pricing subnationally. In both countries, CO2 emissions were reduced marginally. The countries with the least impressive records were Korea and Taiwan. Both countries are non-carbon pricing instruments users. Interestingly, Korea started to use nuclear power since 1980 and renewables since 1991. Taiwan has been a renewable energy user since 1982. While Taiwan shows a slight reduction, Korea shows none. Of course, taking account of GDP preformance, these two countries at least have some control on CO2 emissions.

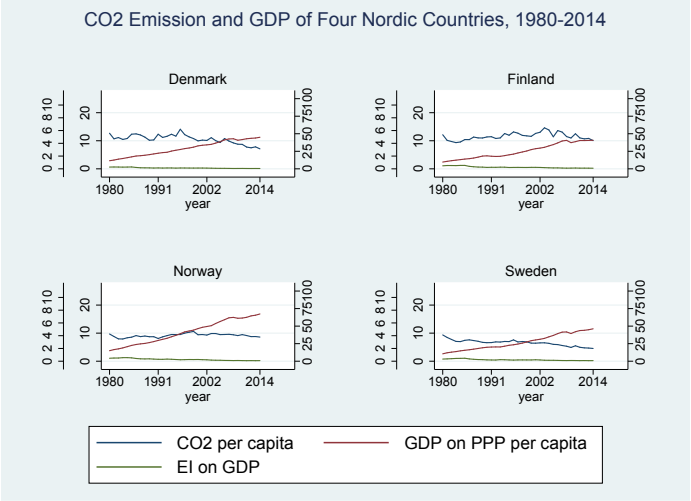
We discuss EU countries separately for a good reason. The EU countries are very similar in many ways; under EU regulation, they have to adapt an ETS. The countries nearby are more alike in terms of renewable energy consumption and advanced 'environmentally friendly' technology development. They also react similarly to the economic shocks. All the 15 EU countries (including UK) in our data have implemented ETS since 2005 without interruption. Switzerland started later in 2008. Besides the four Nordic countries, Ireland (since 2010) and Switzerland (since 2008) have introduced carbon taxation for a considerably long period of time. Carbon Taxation has been implemented in the UK and France for a short period, since 2013 and 2014, respectively. All the EU countries exhibited similar shaped graphs, except Greece since 2008, probably resulting from its GDP recession. Considering each countries' GDP growth, all of them have efficiently controlled CO2 emissions with the use of a carbon pricing mechanism and/or the consumption of renewables and nuclear power.

Figure (3) reports the CO2 emission reduction and GDP preference of four Nordic countries — Sweden, Norway, Finland and Denmark. The four Nordic countries are the first carbon taxation-users, all start from the early nineteen-nineties,

[15] Our analysis on the tax effect on Australia's CO2 reduction was only based on the data from 1980-2014. We have not estimated the effect of the repeal of the carbon tax on Australia's CO2 reductions.

and then implement the EU-ETS in the year 2005 which makes them the first ones to use an ETS in the world as well. Following the Paris Agreement, they made commitments to implement deeper than usual emission cuts.

**Fig. 3**




From the above figures, we can conclude that CO2 emissions do not necessarily always increase with GDP. Some countries have performed extremely well while maintaining GDP growth, for example, Singapore, Hong Kong, USA and most of the EU countries.

## 4 Empirical results

### 4.1 Carbon taxation effects

We first consider a regression model with year and country fixed effects as in Equation (1). We start from including only the carbon taxation dummy to the regression. Then we slowly add more covariates. The regression outputs are reported in Columns (A)-(E) in Table (3). Three methods are used. The usual one-way cluster by country follows Assumption (1), i.e., the standard errors are assumed to be heteroscedastic and autocorrelated. The results are shown in the row of Standard cluster by country. To deal with few clusters problem, still following Assumption (1), we use the wild bootstrap method following Cameron et al. [2008, 2012]. The number of bootstrap replications is 3000 for each regression.The results are reported in the row of CGM. When the standard errors are robust to very general forms of cross-sectional ('spatial') and temporal dependence, i.e., following Assumption (2), we follow Hoechle [2007] and the results are shown in the row of Hoechle. We find clear evidence that the implementation of carbon taxation efficiently reduces CO2 emissions in the 35-year period.

**Table 3:** Estimates for the effect of carbon taxation on CO2 emission per capita per year

| | REGRESSIONS | | | | |
|---|---|---|---|---|---|
| VARIABLES | (A) | (B) | (C) | (D) | (E) |
| Carbon tax | -1.007 | -1.105 | -0.783 | -0.553 | -0.605 |
| Energy Intensity[a] | | -15.346 | -15.829 | -15.855 | -10.405 |
| GDP on PPP per capita | | -0.015 | -0.072 | -0.091 | -0.044 |
| ETS | | | | -1.678 | -1.251 |
| Renewables[b] | | | | | |
| Share of geothermal, biomass and other | | | -26.696 | -19.942 | -18.909 |
| Share of solar | | | -60.017 | -54.087 | -34.599 |
| Share of wind | | | -23.164 | -17.097 | -9.891 |
| Non-renewables | | | | | |
| Share of nuclear | | | -18.036 | -17.585 | -12.777 |
| Share of hydroelectricity | | | | | -13.828 |
| Share of coal | | | | | 6.803 |
| Share of natural gas | | | | | 2.710 |
| Country fixed effect | yes | yes | yes | yes | yes |
| Year fixed effect | yes | yes | yes | yes | yes |
| 95% confidence intervals for carbon taxation effect | | | | | |
| Standard cluster by country | (-2.166 - 0.153) | (-2.187 - -0.023) | (-1.543 - -0.023) | (-1.207 - 0.101) | (-1.206 - -0.004) |
| CGM (bootstrap reps 3000[c]) | (-2.003 - 0.051) | (-2.033 - -0.135) | (-1.595 - 0.029) | (-1.191 - 0.084) | (-1.145 - -0.065) |
| Hoechle | (-1.518 - -0.495) | (-1.723 - -0.488) | (-1.122 - -0.444) | (-0.991 - -0.218) | (-0.882 - -0.225) |
| Sample size | | | | | |
| Number of countries | 26 | 26 | 26 | 26 | 26 |
| Observations | 910 | 910 | 910 | 910 | 910 |
| R-squared | 0.896 | 0.900 | 0.921 | 0.929 | 0.937 |

[a] energy intensity = Primary energy consumption per capita / GDP, where primary energy includes both renewables and non-renewables.
[b] Share of energy consumption = energy consumption / Primary energy consumption.
[c] We apply a country-specific bootstrap.

To estimate the effect of implementation of carbon taxation on CO2 emission per capita per year, we start from a linear regression including only one carbon tax dummy. The estimation equation is as follows:

$$CO2_{it} = \alpha \cdot Tax_{it} + \theta_i + \delta_t + \varepsilon_{it} \tag{3}$$

where the binary regressor $Tax_{it}$ equals one if carbon taxation is in effect in country $i$ in year $t$ and equals zero otherwise. We follow Dynarski [2004], Cameron et al. [2008, 2012] and Hoechle [2007] and compare the regression output. The results are reported in Colomn (A). Countries using carbon taxation reduce their CO2 emissions by $-1.007$ Million ton per year per capita. The 95% confidence intervals are given in the rows. An interval of -2.166 and 0.153 is obtained with standard cluster by countries. When bootstrapping 3000 replications, a narrower interval of -2.003 and 0.051 is obtained. In the third row, the standard errors are robust to heteroskedasticity, autocorrelation and cross-sectional (spatial) and temporal dependence. The carbon taxation treatment effect becomes highly sufficient — an interval of -1.518 and -0.495 is obtained. Renewable and nuclear energies are widely considered to be efficient in CO2 emission reduction. Therefore we add them in Regression (C) as control covirates. For countries that use these energies and carbon taxation, an average CO2 reduction of $-0.783$ is found.

We now slowly add more control covariates as reported in Columns (B)-(E). Although the results differ depending on the choice of covirates, overall, the interval estimates indicate significant treatment effect. In Column (E), Emission Trading Scheme dummy, shares of renewable and non-renewable energy consumption (less oil consumption), GDP on PPP per capita and Enengy Intensity of GDP are the control covariates. The regression now becomes:

$$\begin{aligned} CO2_{it} = {} & \alpha Tax_{it} + \beta_1 EI_{it} + \beta_2 GDP_{it} + \beta_3 ETS_{it} + \beta_4 ShareBiomass_{it} + \beta_5 ShareSolar_{it} + \beta_6 ShareWind_{it} \\ & + \beta_7 ShareNuclear_{it} + \beta_8 ShareCoal_{it} + \beta_9 ShareHydroelectricity_{it} + \beta_{10} ShareNaturalGas_{it} + \theta_i + \delta_t + \varepsilon_{it} \end{aligned} \quad (4)$$

On average, the implementation of carbon taxation reduces CO2 by $-0.605$ Million ton per capita per year. Using the standard cluster by country, the confidence interval is $(-1.206 - -0.004)$. The interval changes to $(-1.145 - -0.065)$ when we apply the country specific bootstrap. When correcting for cross-sectional dependence, the confidence interval of $(-0.882 - -0.225)$ is narrower.

Next, we examine whether the length of use of carbon taxation has an effect on CO2 emissions. We create new variables for every 5 more years of carbon taxation implementation as shown in Table (4). In Regression (A), we exclude any control covirates. The estimates are not statistically significant. The results improves when we control for ETS dummy, energy intensity and GDP on PPP per capita as shown in Regression (B). We observe a clear decreasing trend — a CO2 reduction of $-0.655$ Million ton per capita in the first five years and $-2.382$ for a over 20-year taxation implementation. Finally, we add the shares of renewable and non-renewable energy consumption (less oil consumption) in Regression (C). The results are highly significant as the robust confidence intervals show. In the first 5-year of using carbon taxation, countries in the treatment group reduce CO2 by $-0.632$ Million ton per capita. The longer the duration of carbon taxation implementation, the more CO2 emissions are reduced. For the four countries that have been using carbon taxation for over 20 years: Denmark, Finland, Norway and Sweden, the CO2 reduction increased to $-1.757$ Million ton per capita as reported in the row of tax21to25.

In the robustness check we test the model with a different selection of control group. Following Dynarski [2004], we drop the non-carbon tax users which leave us date of 11 countries. These countries are in the control group before carbon taxes are implemented. Their identities change to the treatments once carbon taxes are introduced. We find that our main results are not sensitive to the choice of the control group. Therefore, we suggest a clear evidence on the effectiveness of the implementation of carbon taxation on CO2 emissions reduction.

### 4.2 The interaction of carbon taxation and ETS

Next, we test the joint effect of carbon taxation and ETS. We try to answer whether the use of both carbon pricing instruments would reduce CO2. By comparing the regression outputs we get above, we try to determine whether a combination of both instruments are a more efficient way to reduce CO2 emissions. The regression is as follows:

$$CO2_{it} = \alpha (Tax * ETS)_{it} + \mathbf{x}'_{it}\beta + \theta_i + \delta_t + \varepsilon_{it} \quad (5)$$

where the environmental policy of interest becomes the interaction of carbon taxation and an ETS. $(Tax * ETS)_{it}$ equals one if country $i$ used both carbon pricing instruments in year $t$ and equals zero otherwise. Ten countries in the treatment group are Japan, Denmark, Finland, France, Ireland, Norway, Sweden, Switzerland, United Kingdom and Canada.

The results are reported in Table (5). We start from regressing CO2 per year per capita on the $Tax * ETS$ dummy only as in Column (A). A yearly reduction of $-1.248$ Million ton is found for the countries who use both carbon taxation and ETS. From the 95% confidence intervals reported in the parentheses, the results are statistically significant. The coefficient changes to $-1.348$ when we control for the Energy Intensity and GDP on PPP per capita to the regression as Column (B) shows. In Regression (C), we add the renewable and nuclear energies as control covirates. For countries that use these energies and both carbon pricing instruments, an average CO2 reduction of $-1.057$ is found. In Column (D), we report the regression including all shares of renewables and non-renewables consumption. The implementation of both instruments effectively reduces $-0.886$ Million ton CO2 per capita per year. As before, to correct the standard error estimates, three methods are used. At the 95% confidence level, the results are statistically significant. Furthermore, we compare the treatment effects of tax-only (Table 3) and a joint use of both carbon pricing instruments. We find that a joint use performs better: more CO2 emission has been reduced.

Next, we examine whether the length of the use of both instruments has an effect on CO2 reduction. We create new variables for every three additional years of implementation. The results are reported in Table (6). We first exclude

**Table 4:** Estimates for the effect of carbon taxation on CO2 emission per capita per year

| | REGRESSIONS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (A) | | | (B) | | | (C) | | | |
| VARIABLES | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle | |
| tax1to5 | -0.348 | -0.348 | -0.348 | -0.655 | -0.655 | -0.655*** | -0.632** | -0.632** | -0.632*** | |
| | (-0.991 - 0.296) | (-0.992 - 0.297) | (-0.971 - 0.276) | (-1.458 - 0.147) | (-1.476 - 0.165) | (-1.106 - -0.205) | (-1.141 - -0.122) | (-1.148 - -0.115) | (-0.965 - -0.298) | |
| tax6to10 | 0.014 | 0.014 | 0.014 | -0.810 | -0.810 | -0.810** | -0.746** | -0.746** | -0.746*** | |
| | (-0.943 - 0.970) | (-0.732 - 0.760) | (-0.695 - 0.723) | (-1.850 - 0.229) | (-1.994 - 0.373) | (-1.437 - -0.184) | (-1.428 - -0.065) | (-1.468 - -0.025) | (-1.216 - -0.277) | |
| tax11to15 | 0.020 | 0.020 | 0.020 | -1.133 | -1.133 | -1.133*** | -1.146** | -1.146** | -1.146*** | |
| | (-1.526 - 1.566) | (-0.644 - 0.685) | (-0.566 - 0.606) | (-2.771 - 0.506) | (-3.002 - 0.737) | (-1.776 - -0.489) | (-2.166 - -0.126) | (-2.269 - -0.024) | (-1.741 - -0.552) | |
| tax16to20 | -0.862 | -0.862 | -0.862* | -1.368 | -1.368 | -1.368*** | -1.305** | -1.305** | -1.305*** | |
| | (-2.642 - 0.918) | (-2.754 - 1.030) | (-1.728 - 0.004) | (-3.176 - 0.439) | (-3.465 - 0.729) | (-2.181 - -0.556) | (-2.421 - -0.190) | (-2.521 - -0.090) | (-1.941 - -0.670) | |
| tax21to25 | -1.663* | -1.663 | -1.663*** | -2.382** | -2.382* | -2.382*** | -1.757*** | -1.757*** | -1.757*** | |
| | (-3.489 - 0.164) | (-3.831 - 0.505) | (-2.496 - -0.829) | (-4.291 - -0.472) | (-4.921 - 0.158) | (-3.508 - -1.255) | (-2.929 - -0.585) | (-3.067 - -0.446) | (-2.644 - -0.869) | |
| ETS | | | | -1.495*** | -1.495** | -1.495*** | -1.074** | -1.074** | -1.074*** | |
| | | | | (-2.588 - -0.401) | (-2.747 - -0.243) | (-2.204 - -0.785) | (-1.990 - -0.157) | (-2.095 - -0.053) | (-1.718 - -0.429) | |
| energy intensity | | | | -14.082* | -14.082** | -14.082*** | -7.139 | -7.139 | -7.139 | |
| | | | | (-29.605 - 1.441) | (-29.569 - 1.405) | (-23.297 - -4.867) | (-19.790 - 5.511) | (-18.821 - 4.543) | (-16.047 - 1.769) | |
| GDP on PPP p.c. | | | | 0.051 | 0.051 | 0.051*** | 0.042 | 0.042 | 0.042*** | |
| | | | | (-0.024 - 0.126) | (-0.029 - 0.130) | (0.031 - 0.071) | (-0.021 - 0.105) | (-0.031 - 0.115) | (0.019 - 0.064) | |
| Shares of energy[c] | | | | | | | | | | |
| share hydro | | | | | | | -19.794*** | -19.794*** | -19.794*** | |
| | | | | | | | (-29.816 - -9.772) | (-34.409 - -5.179) | (-24.283 - -15.305) | |
| share coal | | | | | | | 7.485*** | 7.485*** | 7.485*** | |
| | | | | | | | (2.064 - 12.906) | (2.394 - 12.576) | (5.583 - 9.388) | |
| share natgas | | | | | | | 7.434** | 7.434** | 7.434*** | |
| | | | | | | | (0.685 - 14.183) | (0.589 - 14.280) | (5.328 - 9.541) | |
| share geoth | | | | | | | -0.079 | -0.079 | -0.079 | |
| | | | | | | | (-20.566 - 20.409) | (-2.560e+19 - 2.560e+19) | (-6.082 - 5.925) | |
| share solar | | | | | | | -41.089*** | -41.089*** | -41.089*** | |
| | | | | | | | (-70.035 - -12.142) | (-64.702 - -17.476) | (-68.132 - -14.046) | |
| share wind | | | | | | | -11.301 | -11.301 | -11.301*** | |
| | | | | | | | (-26.134 - 3.531) | (-27.528 - 4.926) | (-17.270 - -5.332) | |
| share nuclear | | | | | | | -9.023** | -9.023*** | -9.023*** | |
| | | | | | | | (-17.840 - -0.205) | (-14.901 - -3.144) | (-13.275 - -4.770) | |
| Country fixed effect | | yes | | | yes | | | yes | | |
| Sample size | | | | | | | | | | |
| Number of countries | 26 | | | 26 | | | 26 | | | |
| Observations | 910 | | | 910 | | | 910 | | | |
| R-squared | 0.864 | | | 0.898 | | | 0.926 | | | |

[a] Robust confidence interval in parentheses. *** p<0.01, ** p<0.05, * p<0.1
[b] CGM Bootstrap reps 3000.
[c] The covariates 'Shares of energy' represent the shares of all renewable and non-renewable energy consumption less oil.

any control covirates as Regression (A) shows. The treatment effect is immediate: the yearly emission reduction of $-0.572$ Million ton is found in the first three-year period. For a country that has used both instruments for 10 years, the emission reduction of $-2.248$ is dramatic. By including Energy Intensity and GDP on PPP per capita in Regression (B), the results improves sightly. We finally control for the shares of energy consumption as shown in Regression (C). Starting from an immediate reduction of $-1.078$ in the first three-year period, approximately 0.1 Million ton more CO2 emission reduces for every three more years of implementation. We therefore see clear evidence of a stable decreasing trend: the longer the duration of the implementation, the more CO2 emission reduces. From the 95% of confidence interval in parentheses, the results are highly significant.

**Table 5:** Estimates for the effect of carbon pricing instruments on CO2 emission per capita per year

| | REGRESSIONS | | | |
|---|---|---|---|---|
| VARIABLES | (A) | (B) | (C) | (D) |
| Tax X ETS[a] | -1.248 | -1.348 | -1.057 | -0.886 |
| Energy Intensity | | -15.363 | -15.723 | -9.863 |
| GDP on PPP per capita | | -0.015 | -0.072 | -0.025 |
| Renewables | | | | |
| Share of geothermal, biomass and other | | | -25.710 | -22.508 |
| Share of solar | | | -70.268 | -45.387 |
| Share of wind | | | -22.144 | -12.341 |
| Non-renewables | | | | |
| Share of nuclear | | | -18.035 | -11.817 |
| Share of hydroelectricity | | | | -12.920 |
| Share of coal | | | | 8.214 |
| Share of natural gas | | | | 3.824 |
| Country fixed effect | yes | yes | yes | yes |
| Year fixed effect | yes | yes | yes | yes |
| 95% confidence intervals for carbon taxation effect | | | | |
| Standard cluster by country | (-2.449 - -0.047) | (-2.529 - -0.167) | (-1.876 - -0.239) | (-1.616 - -0.156) |
| CGM (bootstrap reps 3000) | (-2.431 - -0.065) | (-2.585 - -0.111) | (-1.933 - -0.182) | (-1.683 - -0.089) |
| Hoechle | (-1.713 - -0.783) | (-1.871 - -0.825) | (-1.517 - -0.597) | (-1.271 - -0.502) |
| Sample size | | | | |
| Number of countries | 26 | 26 | 26 | 26 |
| Observations | 910 | 910 | 910 | 910 |
| R-squared | 0.897 | 0.901 | 0.922 | 0.934 |

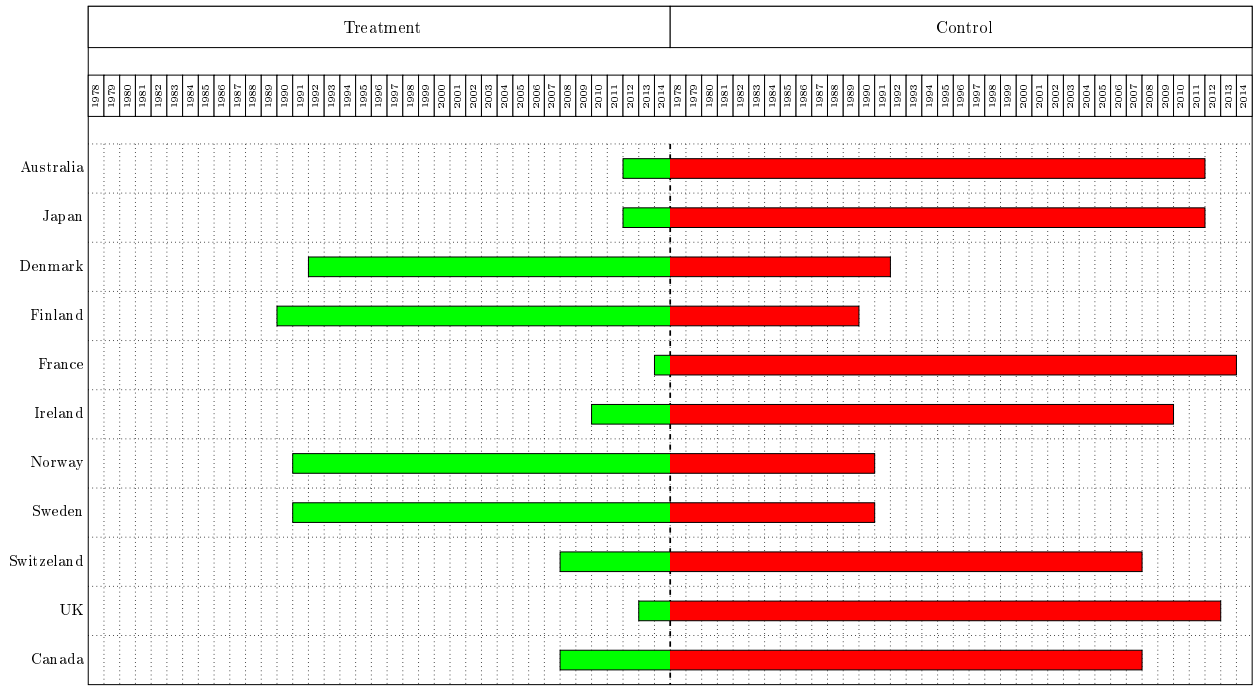
[a] It is to test the effect of the interaction of ETS and carbon tax on CO2 emission. 1 for the countries who use both ETS and carbon tax, 0 otherwise.

# 5 Robustness check

## 5.1 Selection of the control group

We suggest there is little doubt that carbon pricing instruments emitters that are subject to have a higher incentive to reduce CO2 emissions and contribute more to a cleaner environment. Therefore they might have stronger preferences towards the use of nuclear or renewable energy, compared to the countries that have not implemented either carbon taxation or an ETS by 2014. Also taking account of countries' different GDP performance and Energy Intensity improvement as well as other country specific effects which we do not include in our model (such as nature resource, cars emission and fuel economy figures), we suspect that the non-carbon pricing instrument users form a poor control group. We follow Dynarski [2004] and test the sensitivity of our results to the choice of control group. We drop 15 non-carbon taxation users from the sample and test the effect of carbon taxation from the staggered timing of its implementation across countries. The identification of the treatments (in green) and controls (in red) is illustrated in Figure (4). Finland

**Fig. 4:** Timing of introduction of carbon taxation




was the first to introduce carbon tax in 1990. Thus before 1990, all the 11 countries are the controls. In 1990, Finland moves into the treatment group, followed by Norway and Sweden in 1991 and then by Denmark in 1992. France is the

**Table 6:** Estimates for the effect of carbon pricing instruments on CO2 emission per capita per year

| | REGRESSIONS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (A) | | | (B) | | | (C) | | | |
| VARIABLES | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle | |
| taxXets1to3 | -0.572 | -0.572 | -0.572*** | -1.328** | -1.328** | -1.328*** | -1.078*** | -1.078*** | -1.078*** | |
| | (-1.300 - 0.156) | (-1.324 - 0.179) | (-0.975 - -0.170) | (-2.409 - -0.247) | (-2.481 - -0.175) | (-1.796 - -0.860) | (-1.778 - -0.378) | (-1.874 - -0.282) | (-1.451 - -0.705) | |
| taxXets4to6 | -1.110*** | -1.110** | -1.110*** | -2.064*** | -2.064*** | -2.064*** | -1.437*** | -1.437*** | -1.437*** | |
| | (-1.872 - -0.348) | (-2.124 - -0.096) | (-1.447 - -0.773) | (-3.240 - -0.888) | (-3.467 - -0.660) | (-2.713 - -1.414) | (-2.202 - -0.672) | (-2.313 - -0.561) | (-1.890 - -0.983) | |
| taxXets7to9 | -1.696*** | -1.696* | -1.696*** | -2.752*** | -2.752*** | -2.752*** | -1.500*** | -1.500*** | -1.500*** | |
| | (-2.846 - -0.546) | (-3.595 - 0.203) | (-1.994 - -1.398) | (-4.111 - -1.393) | (-4.545 - -0.959) | (-3.405 - -2.099) | (-2.324 - -0.675) | (-2.414 - -0.585) | (-1.984 - -1.015) | |
| taxXets10plus | -2.248*** | -2.248*** | -2.248*** | -3.399*** | -3.399*** | -3.399*** | -1.545** | -1.545*** | -1.545*** | |
| | (-3.590 - -0.906) | (-3.540 - -0.956) | (-2.557 - -1.939) | (-4.891 - -1.906) | (-5.352 - -1.445) | (-4.119 - -2.678) | (-2.688 - -0.402) | (-2.631 - -0.459) | (-2.055 - -1.035) | |
| Energy Intensity | | | | -17.619** | -17.619** | -17.619*** | -9.137 | -9.137 | -9.137** | |
| | | | | (-33.535 - -1.703) | (-34.260 - -0.978) | (-24.733 - -10.505) | (-21.987 - 3.713) | (-22.072 - 3.799) | (-16.812 - -1.462) | |
| GDP on PPP per capita | | | | 0.020 | 0.020* | 0.020* | 0.028 | 0.028 | 0.028** | |
| | | | | (-0.043 - 0.084) | (-0.043 - 0.084) | (-0.003 - 0.044) | (-0.028 - 0.083) | (-0.032 - 0.088) | (0.005 - 0.051) | |
| Shares of energy | | | | | | | | | | |
| share hydro | | | | | | | -18.060*** | -18.060*** | -18.060*** | |
| | | | | | | | (-27.609 - -8.511) | (-31.395 - -4.725) | (-21.997 - -14.123) | |
| share coal | | | | | | | 8.099*** | 8.099*** | 8.099*** | |
| | | | | | | | (2.343 - 13.856) | (2.408 - 13.791) | (5.933 - 10.266) | |
| share natgas | | | | | | | 7.527** | 7.527** | 7.527*** | |
| | | | | | | | (0.520 - 14.533) | (0.649 - 14.405) | (5.315 - 9.738) | |
| share geoth | | | | | | | -12.088 | -12.088 | -12.088*** | |
| | | | | | | | (-35.490 - 11.314) | (-57.983 - 33.807) | (-17.743 - -6.432) | |
| share solar | | | | | | | -57.239*** | -57.239*** | -57.239*** | |
| | | | | | | | (-89.872 - -24.606) | (-90.133 - -24.345) | (-72.659 - -41.819) | |
| share wind | | | | | | | -15.984** | -15.984* | -15.984*** | |
| | | | | | | | (-31.942 - -0.026) | (-32.771 - 0.804) | (-23.775 - -8.193) | |
| share nuclear | | | | | | | -8.749* | -8.749** | -8.749*** | |
| | | | | | | | (-18.500 - 1.002) | (-16.270 - -1.228) | (-13.240 - -4.258) | |
| Country fixed effect | | yes | | | yes | | | yes | | |
| Sample size | | | | | | | | | | |
| Number of countries | | 26 | | | 26 | | | 26 | | |
| Observations | | 910 | | | 910 | | | 26 | | |
| R-squared | | 0.866 | | | 0.888 | | | 0.923 | | |

[a] Robust confidence interval in parentheses.
[b] The covariates 'Shares of energy' represent the shares of all renewable and non-renewable energy consumption less oil.

last country to join the treatments in our sample. It started to use carbon tax in 2014. Thus by 2014, all the eleven countries are in the treatment group.

The regression output is reported in Table (7). Overall, the estimations are not sensitive to the choice of treatment and control group, although the confidence interval becomes less significant. We compare the results in Column (D) in Table (3) and Column (A) in Table (7), the estimation of treatment effect decreases slightly from $-0.605$ to $-0.689$. Similarly, we find the estimations drop by comparing results in Column (C) in Table (4) and Column (B) in Table (7). A 5-year use of carbon taxation reduces CO2 by $-0.845$ Million ton per capita. For every five more year's use of carbon taxation, we find approximately 0.5 Million per capita less CO2 emission. As these results are similar to the ones we obtained earlier, it becomes clear that the longer duration of use of carbon taxation, the more CO2 reduces. For countries that have been using carbon taxation for more than 20 years, the CO2 reduction of $-2.927$ demonstrates its effectiveness.

**Table 7:** Estimates for the effect of carbon taxation on CO2 emission per capita per year, carbon taxation users only

| | REGRESSIONS | | | | | |
|---|---|---|---|---|---|---|
| | (A) | | | (B) | | |
| VARIABLES | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle |
| carbon taxation | -0.689* | -0.689 | -0.689** | | | |
| | (-1.474 - 0.096) | (-1.622 - 0.244) | (-1.369 - -0.009) | | | |
| tax1to5 | | | | -0.845** | -0.845** | -0.845*** |
| | | | | (-1.483 - -0.208) | (-1.624 - -0.067) | (-1.293 - -0.398) |
| tax6to10 | | | | -1.086** | -1.086* | -1.086*** |
| | | | | (-2.020 - -0.152) | (-2.332 - 0.160) | (-1.557 - -0.615) |
| tax11to15 | | | | -1.676** | -1.676 | -1.676*** |
| | | | | (-3.300 - -0.051) | (-3.686 - 0.334) | (-2.272 - -1.079) |
| tax16to20 | | | | -2.174** | -2.174** | -2.174*** |
| | | | | (-3.913 - -0.435) | (-4.144 - -0.204) | (-3.087 - -1.261) |
| tax21to25 | | | | -2.927*** | -2.927** | -2.927*** |
| | | | | (-4.947 - -0.907) | (-5.403 - -0.450) | (-4.195 - -1.658) |
| ETS | | yes | | | yes | |
| Energy Intensity | | yes | | | yes | |
| GDP on PPP per capita | | yes | | | yes | |
| Shares of energy[c] | | yes | | | yes | |
| Country fixed effect | | yes | | | yes | |
| Year fixed effect | | yes | | | | |
| Sample size | | | | | | |
| Number of countries | | 11 | | | 11 | |
| Observations | | 385 | | | 385 | |
| R-squared | | 0.967 | | | 0.966 | |

[a] Robust confidence interval in parentheses.
[b] CGM Bootstrap reps 3000.
[c] The covariates 'Shares of energy' represent the shares of all renewable and non-renewable energy consumption less oil.

Next, we test the sensitivity of our results in the joint effect of both carbon pricing instruments to the choice of control group. We include ten countries that have used both carbon taxation and ETS in the 35-year period. From 1980-2004, all countries are controls. In 2005, the four Nordic countries first move into the treatment group. Switzerland and Canada follow in 2008. France is the last to join the treatments in 2014.

**Table 8:** Estimates for the effect of carbon pricing instruments per capita per year, carbon pricing instruments users only

| | REGRESSIONS | | | | | |
|---|---|---|---|---|---|---|
| | (A) | | | (B) | | |
| VARIABLES | Std cluster by cty | CGM | Hoechle | Std cluster by cty | CGM | Hoechle |
| Tax X ETS | -0.645 | -0.645 | -0.645*** | | | |
| | (-1.442 - 0.152) | (-1.610 - 0.320) | (-1.077 - -0.214) | | | |
| taxXets1to3 | | | | -0.932** | -0.932** | -0.932*** |
| | | | | (-1.590 - -0.274) | (-1.821 - -0.043) | (-1.247 - -0.617) |
| taxXets4to6 | | | | -1.443*** | -1.443** | -1.443*** |
| | | | | (-2.372 - -0.514) | (-2.718 - -0.168) | (-1.910 - -0.976) |
| taxXets7to9 | | | | -1.737*** | -1.737** | -1.737*** |
| | | | | (-2.821 - -0.652) | (-3.174 - -0.300) | (-2.290 - -1.183) |
| taxXets10plus | | | | -1.923*** | -1.923** | -1.923*** |
| | | | | (-3.193 - -0.652) | (-3.513 - -0.332) | (-2.619 - -1.226) |
| Energy Intensity | | yes | | | yes | |
| GDP on PPP per capita | | yes | | | yes | |
| Shares of energy[c] | | yes | | | yes | |
| Country fixed effect | | yes | | | yes | |
| Year fixed effect | | yes | | | | |
| Sample size | | | | | | |
| Number of countries | | 10 | | | 10 | |
| Observations | | 350 | | | 350 | |
| R-squared | | 0.967 | | | 0.960 | |

[a] Robust confidence interval in parentheses. *** p<0.01, ** p<0.05, * p<0.1
[b] CGM Bootstrap reps 3000.
[c] The covariates 'Shares of energy' represent the shares of all renewable and non-renewable energy consumption less oil.

We first compare Regression (C) in Table (5) with Regression (A) in Table (8). The estimate of reduction increases slightly from $-0.886$ to $-0.645$ million tons per year and the results are less significant given by the wider confidence intervals. From Regression (A) in Table (8), we see a similar decreasing trend as shown in Regression (C) in Table (6). An immediate emission reduction of $-0.932$ is found in the first three-year period. This figure doubles for the countries that have used both carbon pricing instruments for ten years as reported in the row of taxXets10plus. Although the results change, overall, we find the choice of the treatment and control groups is not very sensitive. We could still suggest a clear evidence that the implementation of both carbon taxation and ETS has efficiently reduced CO2 emission over the past 35 years. And it performs slightly better than a carbon tax-only implementation.

The selection of bootstrap replications is available upon request. We estimate the effect of carbon taxation (and the joint effect of carbon taxation and ETS) on CO2 emission with 500, 1000 and 2000 bootstrap replications. We show that our main findings hold even with different selections of bootstrap replications. With more replications, better results are obtained. We also test the choice of control covariates. We regress on share of each energy consumption on carbon taxation dummy (and on the interaction of carbon taxation and ETS), Energy Intensity and GDP on PPP per capita. We show that the choice of covariates in our model performs well. The results are available upon request.

## 6 Conclusion

In this article, we have tested the efficiency of carbon taxation by using evidence across 26 of the most developed countries for the period 1980-2014. We employ a simple difference-in-difference model and correct the standard error following Dynarski [2004], Cameron et al. [2008, 2012] and Hoechle [2007]. The error terms are robust to heteroskedasticity,

auto-correlation and cross-sectional dependence. We confirm that in the past 35 years, carbon taxation has effectively reduced CO2 emissions per capita in the developed world. The longer the duration of taxation implementation, the more efficient the reduction of CO2 is found. For countries that have been using both an ETS and carbon taxation, we find the evidence of an even more efficient reduction in CO2 emissions.

The findings of this research have clear implementations for countries that made commitments to cut their CO2 emissions. The use of market-based instruments such as a carbon taxation or a combination with an ETS is an effective mitigation method. Countries around the world will need to give serious consideration to adopting these measures.

An area for further research could be that countries' performance after 2014, especially after the Paris Agreement where most countries, developed and developing, become more open to voluntarily cutting their emissions. Other areas that could be examined include countries that have been the worst CO2 emitters. We especially wish to provide evidence to these countries that have not yet adopted market-based mechanisms like carbon taxes to mitigate emissions.

# Bibliography


A. Bruvoll and B. M. Larsen. Greenhouse gas emissions in Norway: do carbon taxes work? *Energy policy*, 32(4):493–505, 2004.

A. C. Cameron and D. L. Miller. Robust inference with clustered data. *Handbook of empirical economics and finance*, pages 1–28, 2010.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 2012.

B. Doda et al. Evidence on co2 emissions and business cycles. *Work. Pap., Grantham Research Institute on Climate Change and the Environment*, 2012.

S. Dynarski. The new merit aid. In *College choices: The economics of where to go, when to go, and how to pay for it*, pages 63–100. University of Chicago Press, 2004.

D. Hoechle. Robust standard errors for panel regressions with cross-sectional dependence. *The Stata Journal*, 7(3): 281–312, 2007.

B. Lin and X. Li. The effect of carbon tax on per capita co2 emissions. *Energy policy*, 39(9):5137–5146, 2011.

Q. Liu, Y. Chen, C. Tian, X. Zheng, F. Teng, A. Gu, X. Yang, X. Wang, E. Matthews, and R. Song. Peaking China's co2 emissions: Trends and mitigation potential. 2015.

S. Meng, M. Siriwardana, and J. McNeill. The environmental and economic impact of the carbon tax in Australia. *Environmental and Resource Economics*, pages 1–20, 2013.

R. G. Newell, W. A. Pizer, and D. Raimi. Carbon markets 15 years after Kyoto: Lessons learned, new challenges. *The Journal of Economic Perspectives*, 27(1):123–146, 2013.

F. Scrimgeour, L. Oxley, and K. Fatai. Reducing carbon emissions? the relative effectiveness of different types of environmental tax: the case of New Zealand. *Environmental Modelling & Software*, 20(11):1439–1448, 2005.

R. Song, W. Dong, J. Zhu, X. Zhao, and Y. Wang. Assessing implementation of China's climate policies in the 12th 5-year period. *World Resources Institute, September*, 2015.

G. A. Tanguay, P. Lanoie, and J. Moreau. Environmental policy, public interest and political market. *Public Choice*, 120(1):1–27, 2004.

G. Y. Zakarya, B. Mostefa, S. M. Abbes, and G. M. Seghir. Factors affecting co2 emissions in the BRICS countries: a panel data analysis. *Procedia Economics and Finance*, 26:114–125, 2015.

# HYDROGEN ENERGY POTENTIAL DETERMINATION WITH COMPUTATIONAL MATHEMATICS

**Levent Yılmaz**[1]

[1]Department of Civil Engineering, Nisantasi University,Istanbul,Turkey

levent.yilmaz@nisantasi.edu.tr

**Abstract**

Hydrogen is the simplest element. An atom of hydrogen consists of only one proton and one electron. It's also the most plentiful element in the universe. Despite its simplicity and abundance, hydrogen doesn't occur naturally as a gas on the Earth - it's always combined with other elements. Water, for example, is a combination of hydrogen and oxygen (H2O).

Hydrogen is also found in many organic compounds, notably the hydrocarbons that make up many of our fuels, such as gasoline, natural gas, methanol, and propane. Hydrogen can be separated from hydrocarbons through the application of heat - a process known as reforming. Currently, most hydrogen is made this way from natural gas. An electrical current can also be used to separate water into its components of oxygen and hydrogen. This process is known as electrolysis. Some algae and bacteria, using sunlight as their energy source, even give off hydrogen under certain conditions.

NASA uses hydrogen fuel to launch the space shuttles. Credit: NASA

Hydrogen is high in energy, yet an engine that burns pure hydrogen produces almost no pollution. NASA has used liquid hydrogen since the 1970s to propel the space shuttle and other rockets into orbit. Hydrogen fuel cells power the shuttle's electrical systems, producing a clean byproduct - pure water, which the crew drinks.

A fuel cell combines hydrogen and oxygen to produce electricity, heat, and water. Fuel cells are often compared to batteries. Both convert the energy produced by a chemical reaction into usable electric power. However, the fuel cell will produce electricity as long as fuel (hydrogen) is supplied, never losing its charge.

Fuel cells are a promising technology for use as a source of heat and electricity for buildings, and as an electrical power source for electric motors propelling vehicles. Fuel cells operate best on pure hydrogen. But fuels like natural gas, methanol, or even gasoline can be reformed to produce the hydrogen required for fuel cells. Some fuel cells even can be fueled directly with methanol, without using a reformer.

In the future, hydrogen could also join electricity as an important energy carrier. An energy carrier moves and delivers energy in a usable form to consumers. Renewable energy sources,

like the sun and wind, can't produce energy all the time. But they could, for example, produce electric energy and hydrogen, which can be stored until it's needed. Hydrogen can also be transported (like electricity) to locations where it is needed.

***Keywords:*** hydrogen energy potential, computational mathematics

**REFERENCES**

1. Jaehyun Bae,ab Dongwook Kim,b Jong Hyun Jung b and Jisoon Ihm*cb, A computational study on hydrogen storage in potential wells using K-intercalated graphite oxide, RSC Advances, Vol:4, No: 2, 184-201, 2011

# Hybrid Orbit Propagator based on Time Series Forecasting: Predictive Interval

Montserrat San-Martín[2], Iván Pérez[1], Rosario López[1] and Juan Félix San-Juan[1]

[1] Scientific Computing Group (GRUCACI), University of La Rioja, 26006 Logroño, Spain
[2] Scientific Computing Group (GRUCACI), University of Granada, 52005 Melilla, Spain

**Abstract.** The orbital motion of an artificial satellite or space debris object is perturbed by a variety, and sometimes not well-modeled, external forces [1, 2]. The hybrid methodology can be used to predict these unmodeled effects or the uncertainty associated with this process. In this work, a Hybrid Orbit Propagator based on SGP4 [3–7] and a state space formulation of the exponential smoothing method as the forecasting technique is developed. The error terms of the forecasting technique are considered Gaussian noise what allows us to use the maximum likelihood method to estimate the parameters of the exponential smoothing model, as well as computing the point forecast and the reliable predictive intervals. Finally, this Hybrid Orbit Propagator is applied to data from a satellite of the Galileo constellation. This Propagator improves the accuracy of the classical SGP4 and it is particularly good for short forecast horizons.

**Keywords:** Statistical time Series models, artificial satellite problem and Additive Holt-Winters method.

## Acknowledgments

This work has been funded by the Spanish State Research Agency and the European Regional Development Fund under Project ESP2016-76585-R (AEI/ERDF, EU).

## References

1. R. H. Battin, An Introduction to the Mathematics and Methods of Astrodynamics, revised Edition, AIAA Education Series, American Institute of Aeronautics and Astronautics, Inc., Reston, VA, USA, 1999.
2. D. A. Vallado, Fundamentals of Astrodynamics and Applications, 4th Edition, Space Technology Library, Microcosm Press, Hawthorne, CA, USA, 2013.
3. F. R. Hoots, R. L. Roehrich, Models for propagation of the NORAD element sets, Spacetrack Report #3, U.S. Air Force Aerospace Defense Command, Colorado Springs, CO, USA (1980).

4. D. A. Vallado, P. Crawford, R. Hujsak, T. S. Kelso, Revisiting spacetrack report #3, in: Proceedings 2006 AIAA/AAS Astrodynamics Specialist Conference and Exhibit, Vol. 3, American Institute of Aeronautics and Astronautics, Keystone, CO, USA, 2006, pp. 1984–2071, paper AIAA 2006-6753. doi:10.2514/6.2006-6753.
5. J. F. San-Juan, M. San-Martín, I. Pérez, Application of the hybrid methodology to SGP4, Advances in the Astronautical Sciences 158 (2016) 685–696, paper AAS 16-311.
6. J. F. San-Juan, M. San-Martín, I. Pérez, R. López, Hybrid SGP4: tools and methods, in: Proceedings 6th International Conference on Astrodynamics Tools and Techniques, ICATT 2016, European Space Agency (ESA), Darmstadt, Germany, 2016.
7. J. F. San-Juan, I. Pérez, M. San-Martín, E. P. Vergara, Hybrid SGP4 orbit propagator, Acta Astronautica 137 (2017) 254–260. doi:10.1016/j.actaastro.2017.04.015.

# Hybrid Orbit Propagators based on Neural Network

Iván Pérez[1], Rosario López[1], Montserrat San-Martín[2] and Juan Félix San-Juan[1]

[1] Scientific Computing Group (GRUCACI), University of La Rioja, 26006 Logroño, Spain
[2] Scientific Computing Group (GRUCACI), University of Granada, 52005 Melilla, Spain

**Abstract.** Space Situational Awareness current needs demand innovative solutions to the orbit propagation problem, so as to find new algorithms which are simultaneously accurate and fast. The hybrid methodology for orbit propagation constitutes a recent approach based on modeling the error of any orbit propagator with the aim of complementing its calculations and hence enhancing its precision. Diverse sources of inaccuracy can exist in propagators, such as incomplete perturbation models, forces not considered, low-order of the series expansions, etc. The creation of a time series with the differences be- tween ephemerides computed with low-accuracy propagators and their corresponding real observations (or precisely computed ephemerides) allows applying time-series forecasting techniques so as to create a model that includes any dynamics not contained in the original propagator. Then, the adjusted model can be used in order to correct other future predictions. We present an application of the hybrid methodology, in which the time-series forecasting process is performed by means of machine-learning techniques, to the well-known SGP4 propagator [1, 2]. We have adjusted the resulting Hybrid SGP4 propagator [3–5], HSGP4, to the case of Galileo-type orbits. We will show how the use of HSGP4 can reduce the position error of SGP4, hence extending the validity of Two-Line Elements (TLE) from Galileo satellites.

**Keywords:** Forecasting time series, artificial satellite problem and neural network

## Acknowledgments

This work has been funded by the Spanish State Research Agency and the European Regional Development Fund under Project ESP2016-76585-R (AEI/ERDF, EU).

## References

1. F. R. Hoots, R. L. Roehrich, Models for propagation of the NORAD element sets, Spacetrack Report #3, U.S. Air Force Aerospace Defense Command, Colorado Springs, CO, USA (1980).
2. D. A. Vallado, P. Crawford, R. Hujsak, T. S. Kelso, Revisiting spacetrack report #3, in: Proceedings 2006 AIAA/AAS Astrodynamics Specialist Conference and Exhibit, Vol. 3, American Institute of Aeronautics and Astronautics, Keystone, CO, USA, 2006, pp. 1984–2071, paper AIAA 2006-6753. doi:10.2514/6.2006-6753.
3. J. F. San-Juan, M. San-Martín, I. Pérez, Application of the hybrid methodology to SGP4, Advances in the Astronautical Sciences 158 (2016) 685–696, paper AAS 16-311.
4. J. F. San-Juan, M. San-Martín, I. Pérez, R. López, Hybrid SGP4: tools and methods, in: Proceedings 6th International Conference on Astrodynamics Tools and Techniques, ICATT 2016, European Space Agency (ESA), Darmstadt, Germany, 2016.
5. J. F. San-Juan, I. Pérez, M. San-Martín, E. P. Vergara, Hybrid SGP4 orbit propagator, Acta Astronautica 137 (2017) 254–260. doi:10.1016/j.actaastro.2017.04.015.

# A Stochastic Drift Model for Electrical Parameters of Semiconductor Devices

Lukas Sommeregger[1,2] and Horst Lewitschnig[2]

[1] Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt, Austria
[2] Infineon Technologies Austria AG, Siemensstr. 2, 9500 Villach, Austria

**Keywords:** Guardbanding · Longitudinal Data · Parameter Drift · Piecewise Linear Model · Random Slopes

## Introduction

During their development, semiconductor devices are put to accelerated stress tests in order to simulate their life in a short time. Devices are initially tested, stressed for a certain time, then tested again, stressed again, and so on. Such tests before, during and after the stress are called readouts. In the course of such readouts, electrical parameters are tested and their behavior over time, i.e. their drift, can be observed. Potentially they could also drift outside of their specified limits. In order to avoid this, tighter test limits are introduced at the final production test of semiconductors. These tighter limits are called guardbands.

Guardbands should be placed to limit the likelihood that electrical parameters drift outside the specification limits during their lifetime while maximizing the production yield at the same time.

The following abstract shows a way to describe such drift behavior, which serves as a basic framework from which guardbands can be derived. The method is easy to implement and reflects the knowledge that is gained from stress tests and parametric drift investigations.

## Drift Modelling

We are dealing with a special structure of data, that is multiple series of longitudinal data with typically 3 to 4 readouts. Therefore, common time-series approaches such as ARIMA or GARCH models are not sensibly applicable. Here, the approach is to make use of the given data structure and take the time series of each device as a piecewise linear function with a random slope from one readout to the next:

$$X_s = X_{t_n} + a_n \cdot (s - t_n), \qquad \forall\ t_n < s < t_{n+1}. \tag{1}$$

In other words, we see our data as a weighted sum of random slopes between measured points in time $t_n$. The whole series can then be expressed as

$$X_{t_n} = x_1 + \sum_{k=1}^{n-1} a_k \cdot (t_{k+1} - t_k). \tag{2}$$

The slopes $a_k$ are seen as realizations of random variables $A_k$, which are assumed to be normally distributed with parameters $\mu_{a_k}$ and $\sigma_{a_k}$. The advantage is that these parameters can be easily estimated from the time series of all devices using a classical statistical approach:

$$\hat{\mu}_{a_{t_k}} = \frac{1}{t_k - t_{k-1}}\big(\bar{x}_{t_k} - \bar{x}_{t_{k-1}}\big), \tag{3}$$

$$\hat{\sigma}_{a_{t_k}} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(a_k^i - \hat{\mu}_{a_{t_k}})^2}. \tag{4}$$

$\bar{x}_{t_k}$ denotes the arithmetic mean of all readouts $x_{t_k}^i$ of the time series of all devices at time $t_k$.

At each readout, different electrical test equipment with different offsets can be used. This leads to distorted drift data. In order to avoid this effect, unstressed parts - so called reference parts - are tested together with the stressed parts. We correct the time series using these reference parts prior to the calculation. Let $m_t^i$ be the readout data of the $i$-th series at time $t$. The $u$ unstressed reference parts are denoted by $r_t^j$. Then the corrected data $x_t^i$ can be calculated:

$$x_t^i = m_t^i - \frac{1}{u}\sum_{j=1}^{u}(r_t^j - r_1^j). \tag{5}$$

The basic idea is to "center" the reference data and correct the measured data by the average drift of the reference devices.

All in all, we can model the position of the parameter at time $t_n$ as a weighted sum of normally distributed random variables resulting in, again, a normally distributed random variable with parameters $\mu_{t_n}$ and $\sigma_{t_n}$. These distribution parameters can be estimated using the electrical parameters tested at the readouts. Suppose $m$ is the number of devices, then

$$X_{t_n} \sim \mathcal{N}(\mu_{t_n}, \sigma_{t_n}), \tag{6}$$

$$\hat{\mu}_{t_n} = \bar{x}_{t_n} = \frac{1}{m}\sum_{i=1}^{m} x_{t_n}^i, \tag{7}$$

$$\widehat{\sigma}_{t_n} = \sqrt{\sum_{k=1}^{n-1}\sum_{l=1}^{n-1}(t_{k+1} - t_k)(t_{l+1} - t_l)Cov(A_k, A_l)}. \tag{8}$$

With (6) we can also accurately estimate the expected value and variance of the time series at arbitrary points in time.

The likelihood that devices drift outside their specified limits $U$ and $L$ during their lifetime is a quality target. The probability at any given point in time $t$ to

drift outside a given upper or lower Limit $U$ or $L$, respectively can be expressed as:

$$P(X_t > U) = 1 - \Phi\left(\frac{U - \mu_t}{\sigma_t}\right), \tag{9}$$

$$P(X_t < L) = \Phi\left(\frac{L - \mu_t}{\sigma_t}\right). \tag{10}$$

$\Phi$ in (9) and (10) denotes the cumulative density function (cdf) of the Gaussian normal distribution.

With this model we can accurately calculate guardbands based on given quality targets and maximize production yield at the same time. It serves as a powerful tool to stabilize production and quality.

## Outlook

For further research,we see several additional items to consider:

- The measurement error (GR&R - Gauge Repeatability and Reproducibility) needs to be reflected in the model.
- The model should be extended to more complex drift patterns, like clustering of drift data, drift outliers or correlation between the drift of different electrical parameters.
- Finally, we see the need to develop an overall framework to describe general drift patterns. Time series models, stochastic processes and probabilistic models could be used for this purpose.

*This project has received funding from the ECSEL Joint Undertaking under grant agreement No 737469 (AutoDrive). This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and Germany, Austria, Spain, Italy, Latvia, Belgium, Netherlands, Sweden, Finland, Lithuania, Czech Republic, Romania, Norway.*

# Chaos and Slow Earthquakes Predictability

Adriano Gualandi, Jean-Philippe Avouac, Sylvain Michel and Davide Faranda
Technische Hochschule Ulm

## Abstract

Slow Slip Events (SSEs) are episodic slip events that play a significant role in the moment budget along a subduction megathrust. They share many similarities with regular earthquakes, and have been observed in major subduction regions like, for example, Cascadia, Japan, Mexico, New Zealand. They show striking regularity, suggesting that it might be possible to forecast their size and timing, but the prediction of their extension and exact timing is still yet to come. They certainly are a great natural system to study how friction works at scale of the order of hundreds or thousands of km, and their recurrence time being much shorter than that of regular earthquakes, they give us the possibility to study multiple cycles and test their predictability. Here we focus on the Cascadia region, where SSEs recur every about 1 or 2 years, depending on the latitude. We use a catalog containing dozens of SSEs derived from the study of GPS position time series during the time span ranging from 2007 to 2017. The data show a clear segmentation with a few major patches interacting with one another, a behavior that recalls that of a discrete body system. We use both classical embedding theory and extreme value theory applied to the study of dynamical systems to show that, where the signal to noise ratio is sufficiently high, a low-dimensional (< 5) non-linear chaotic system is more appropriate to describe the dynamics than a stochastic system. We calculate major properties of the strange attractor like its correlation and instantaneous dimension, its instantaneous persistence and a possible range for the metric entropy of the system. For the better resolved segments, the onset of large SSEs can be correctly forecasted by high values of the instantaneous dimension. Longer-term deterministic prediction seems intrinsically impossible. In conclusion, SSEs in Cascadia can be described as a deterministic, albeit chaotic, system rather than as a random process. As SSEs might be regarded as earthquakes in slow motion, regular earthquakes might be similarly chaotic and predictable but with a predictable horizon of the order of their duration.

# Load Forecast by Multi Task Learning Models: designed for a new collaborative world

Leontina Pinto* Jacques Szczupak
ENGENHO
Rio de Janeiro, Brazil
* leontina@engenho.com

Robinson Semolini
ELEKTRO - NEOENERGIA
São Paulo, Brazil
robinson.semolini@elektro.com.br

***Abstract*—This paper proposed a forecasting model designed for lack-of-data problems, based on Multi tasking learning techniques. It is especially useful for evolutionary markets and systems, where new paradigms (like renewable penetration or prosumers) significantly impact behavior and dynamics, creating unforeseen responses, unpredictable from past (possibly obsolete) historical data. A case study targeting the recent Brazilian load changes illustrate the approach performance: it was possible to combine data from four different distribution companies, creating a learning network, yielding reliable results where all other models failed.**

***Keywords-load forecast, lack of data, multi tasking learning, collaborative learning***

## I. INTRODUCTION

Energy demand is perhaps the most important pillar of the market: all institutions, agents and processes - from planning and operation to marketing and management are essentially organized to serve it. Although projecting load future evolution is crucial for an economical and secure supply, it is still one of our major challenges. The behavior of the consumer changes continuously, restyling reactions to various stimuli - from prices to economic indicators, including expectations and perceptions not always based on reality.

The Brazilian load offers an interesting case study. Year 2018 experienced an anomalous increase in consumption throughout Brazil, almost always without connection to any of the classical explaining triggers: GDP experienced a sharp fall, as did income and all economic activities' indicators. We currently face a major challenge: consumer behavior has changed, and old history no longer represents the present and we *must* predict the future, without any past basis. In fact, in this context, the longer the history, the worse the prediction.

It is necessary to develop mathematical models and computational tools as agile as the consumer, able to understand, follow and maybe anticipate its behavior, with the speed of our new times.

## II. OBJECTIVE

This paper proposes a model able to accommodate more than just lack of data: we deal with *extreme* scarcity, where forecast needs to be performed from very few observations - for example, one year (twelve months). In this case, historical records are not even enough to allow a backtracking test (identification/ prediction): it will be necessary to start from scratch.

It is necessary to "populate" the load history with valid information – and it is important to distinguish information from numbers: it would be possible to create synthetic samples from the available data, but they would contain the same poor information – anything else could even lead us to distorted results

Although it is not possible to extract more information from a history beyond the availability limits, it is feasible to *combine* similar experiences, observations from different agents that exhibit similar behaviors. For example, it is possible that distributors in neighboring regions share the same dynamics of consumption. In this case, it might be interesting to combine the experiences of each into a single richer, more complete history.

This is the proposal of collaborative learning (MTL) [1-3]. By joining forces, information is shared without losing individuality. The model should select the common dynamics and point specificities, leading to a more consistent and reliable projection.

## III. MULTI TASK LEARNING APPROACH

Considering space limitations, this article summarizes the applied collaborative learning model. More details, including alternative implementations, may be found in [3].

The proposed approach establishes a set of outputs or tasks $\boldsymbol{t}$ (in our case, the target variables, loads or consumption). Each of these tasks is associated with a set of explanatory variables (inputs) $\boldsymbol{x}$ (in our case, economic, climatic, behavioral activities, etc.). The successful collaborative learning model requires that outputs $\boldsymbol{t}$ react similarly to inputs $\boldsymbol{x}$.

The function that "maps" the input $\boldsymbol{x}$ to the output $\boldsymbol{t}$ is written as

$$\boldsymbol{f_t}(\boldsymbol{x}) = \sum_{i=1}^{d} \boldsymbol{a_{it}} \boldsymbol{u_i}(\boldsymbol{x}) \;:\; \forall \boldsymbol{t} \in \boldsymbol{T};\; \boldsymbol{a_{it}} \in \mathbb{R};\; \boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}} \qquad (1)$$

where

**x** is the vector of input variables

$\boldsymbol{f_t}(\boldsymbol{x})$ is the output associated to task **t**.

function $\boldsymbol{u_i}(\boldsymbol{x})$ expresses the shared responses of all inputs $\boldsymbol{x}$ and different tasks $\boldsymbol{t}$

coefficients $\boldsymbol{a_{it}}$ measure the "coupling" between different tasks.

For the sake of simplicity, this work assumes linear functions (non linear extensions are possible and relatively straight forward). In this case, function ***f(t)*** corresponds to a vetor product which may be written as

$$w_t = \sum_{i=1}^{d} a_{it} u_i \quad (2)$$

and therefore

$$f_t(x) = w_t(x) \; : \forall t \in T; \; x \in \mathbb{R}^d \quad (3)$$

where $w_t(x)$ combines the individual task coefficients $a$ to the shared $u$

Finally, for concision

$$W = UA : W \in \mathbb{R}^{d \times T} \quad (4)$$

These coefficients are obtained from the historical observations among all agents (even if scarce). Among other methods, the most intuitive is some technique of function fitting to the available history

$$min \{ \sum_{i=1}^{m} L(y_{ti}, \; < a_{ti}, \; U^T x_{ti} >) \} : a_t \in \mathbb{R}^d \quad (5)$$

where $L(.,.)$ measures the empirical deviation between the model outputs and the available data.

Figure 1 and 2 illustrate the conceptual difference between the classical and collaborative approach. While the classical approach uses each observation isolatedly, the collaborative approach combines all observations, creating a common pattern without loosing each agent's uniqueness.

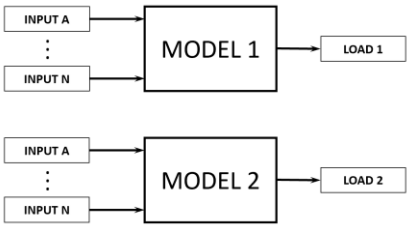



Figure 1. Classic, individual approach

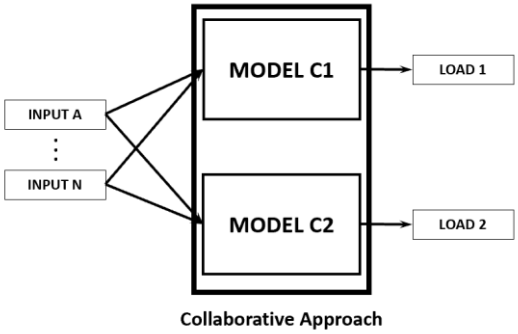



Figure 2. Collaborative approach

## IV. Case Study

### A. *The challenge*

The necessity of a new model, able to deal with lack of data, is shown in Fig, 3. After years of stagnation, the load finally experienced a steep – and unexpected – rise.

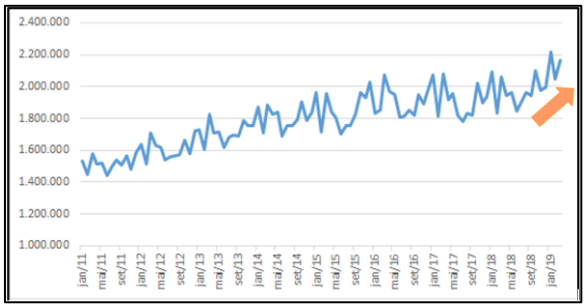



Figure 3. Bahia (COELBA) load growth

The explanation to this phenomenon, however, was unclear. Fig. 4-6 shows the classical forecast phases for a backtracking process (identification and projection) applied to three neighboring distributors (COELBA, CELPE, COSERN), based on usual explaining variables (GDP, Income, Temperature). There is a clear, and unexplained, step associated to 2019 summer in all companies (in fact, all Brazilian distributors exhibited the same behavior.

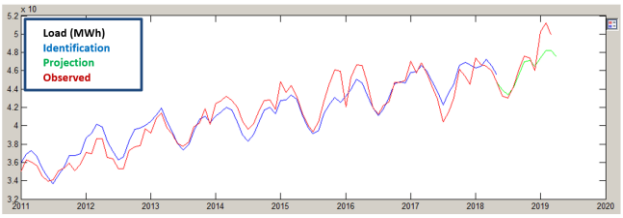



Figure 4. Bahia (COELBA) load dynamics

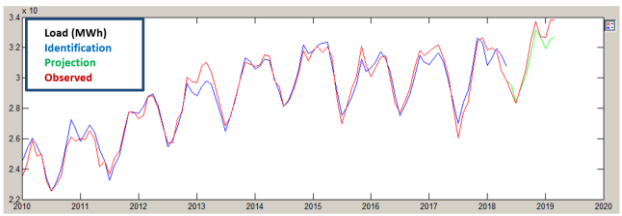



Figure 5. Pernambuco (CELPE) load dynamics

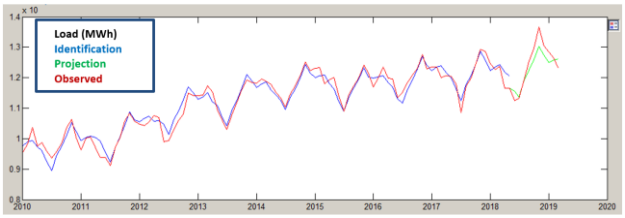



Figure 6. Rio Grande do Norte (COSERN) load dynamics

More than absorbing the deviations, the main question is: should that step be an anomaly, or should it be a change in

consumer's behavior – in other words, is this a new permanent pattern?

This question is, of course, related to the consumer's behavior and the answer requires a deeper – non statistical – understanding.

Extensive field research [4], based on behavioral economics [5-6], showed an interesting fact: a disputed election restored the consumer's belief on a stronger economy and a change for the better. This faith in the future, associated to an unusual warm summer, lead to the highest level of refrigeration equipment purchase observed in a decade.

In other words, consumers possess a new basis of installed demand, and will use it from now on. There is indeed a new standard, which will induce a new response, that must be predicted based on a few observations.

### B. *The proposed solution*

The anomalous behavior was detected from May 2018. It would be very difficult, if not impossible, to use as few as 12 to 18 months for model identification/validation.

We then tested the collaborative learning technique. As our goal was predicting 2019 summer, we based our identification phase on the period from October 2017 to May 2018 – where the behavior was still establishing. Of course, more observations will improve the results and will be used as they become available.

Fig. 7-10 compare the results obtained from our best classical model (individual learning) [7-8] and from the collaborative learning. It is interesting to notice that (as expected) the results show slightly higher errors during springtime (as consumers were still adapting, taking decisions, buying equipment). However, projection for summer months are much better.

In any case, the proposed approach yielded a clear enhancement on the overall forecast quality. All deviations are significantly lower, despite the almost non-existing information. Moreover, the "deviation trend" is broken, offering a more stable and reliable insight of the future.

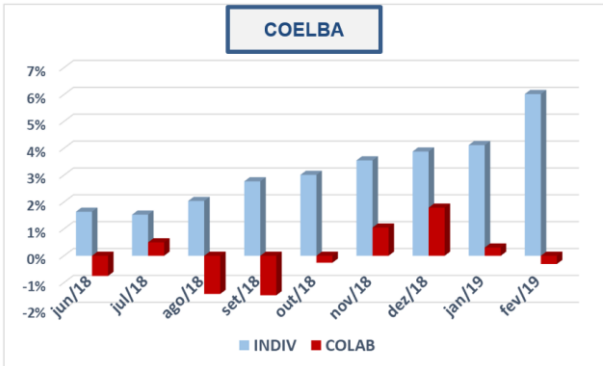



Figure 7. Individual x Collaborative learning, Bahia (COELBA)

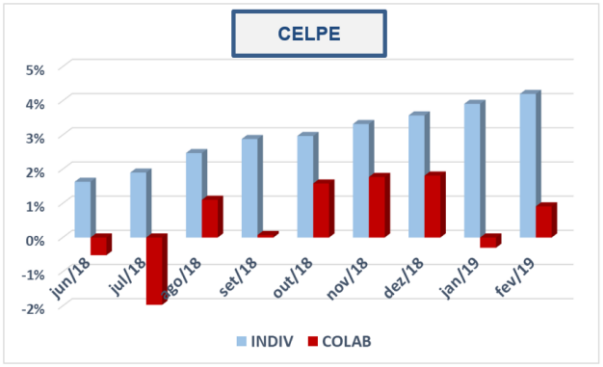



Figure 8. Individual x Collaborative learning, Pernambuco (CELPE)

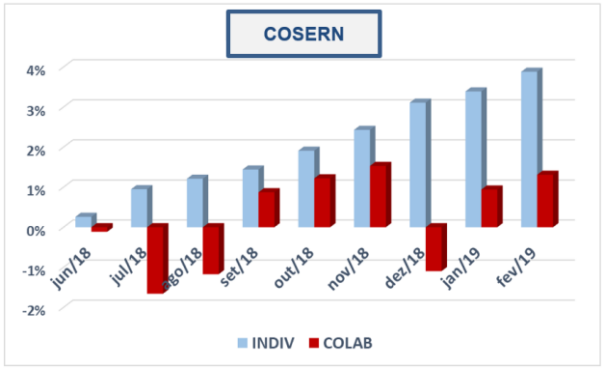



Figure 9. Individual x Collaborative learning, Rio G. do Norte (COSERN)

## V. CONCLUSIONS

We live in a changing world, and consumption dynamic is not an exception. Preparedness for the future requires the forecast of the unknown. It is crucial to build models able to quickly detect modifications – and know the difference from anomalies. It will be necessary to adapt, adjust, absorb novelties.

In the context, classical models, which try to repeat the past, will not be able to foresee the future. The ability to collect and store a huge history may not ensure quality of information. Number quantity will not necessarily yield precision.

We propose a model designed for this new reality: a collaborative learning technique, able to combine information from different agents, identify common and individual characteristics and build a rich history without traveling back to a distant past.

The described approach was applied to a hard challenge: the projection of the summer load for three Brazilian distributors which broke any known record. A mere 8-month observed data was able to provide much better results for all companies, beginning to explain the (previously) unexplainable behavior.

These promising results suggest an interesting way, which will be pursued and reported in the near future.

## REFERENCES


[1] R. Caruana, “Multitask Learning”, Machine Learning, Volume 28, Issue 1, July 1997, pp 41-75

[2] A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-task feature learning,” in Advances in Neural Information Processing Systems 19, pp. 41–48, 2006

[3] Y. Zhang and Q. Yang, “A Survey on Multi-Task Learning”, arxiv pre-print, 2017.

[4] ENGENHO “Brazilian Load Growth Diagnostics”, report, available from www.engenho.com

[5] Eia, US Energy Information Administration, “Behavioral Economics Applied to Energy Demand Analysis: A Foundation”, October 2014

[6] R. H. Thaler, “Misbehaving: The Making of Behavioral Economics”, W. W. Norton & Company, 2016

[7] J. Szczupak, L. Pinto, L.H. Macedo, J. Pascon. R. Semolini, M. Inoue, C. Almeida ; F. R. Almeida, “Load Modeling and Forecast based on a Hilbert Space Decomposition”, 2007 IEEE Power Engineering Society General Meeting, disponível na base de dados do repositório IEEEXPLORE https://ieeexplore.ieee.org/document/4275991

[8] L. Pinto, J. Szczupak, C. Almeida, L. Macedo, M. Inoue, R. Massaro, R. Semolini, J. Pascon, E. Albarelli, D. Tortelli, "Load Forecast under uncertainty: Accounting for the economic crisis impact", PowerTech 2009 IEEE Bucharest, pp. 1-5, 2009.

# Forecasting inflation in the euro area: countries matter!

Angela Capolongo[⋆,†] and Claudia Pacella[⋆,‡]

[⋆]ECARES, Université Libre de Bruxelles
[†]F.R.S.- FNRS
[‡]Bank of Italy

**Abstract.** We construct a Bayesian vector autoregressive model including the key drivers of inflation, cross-country dynamic interactions, and country-specific variables. The model provides good forecasting accuracy with respect to the popular benchmarks used in the literature. We perform a step-by-step analysis to shed light on which information is more crucial for forecasting euro area inflation. The complete model performs better in forecasting inflation excluding energy and unprocessed food. Our empirical analysis reveals the importance of including the key drivers of inflation and taking into account the multi-country dimension of the euro area.

**Keywords:** inflation · forecasting · euro area · Bayesian estimation

## 1 Introduction[1]

The primary objective of the European Central Bank is to maintain price stability in the euro area as a whole. This general goal has been further specified in terms of keeping the year-on-year increase in the euro area Harmonised Index of Consumer Prices (HICP) below, but close to 2% over the medium term.

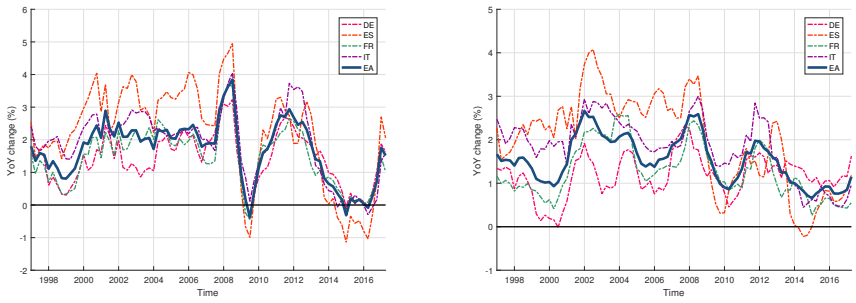
Given this objective, a timely assessment of the economic drivers and the most likely outlook for inflation are a fundamental input for monetary policy. However, as reviewed by Faust and Wright [1], while a large number of models have been proposed to forecast inflation, interpreting the inflation dynamics and providing an informed view on the inflation outlook has always been a challenging exercise.

For the US, Atkenson and Ohanian [2] show that it is difficult to outperform very simple models as the random walk, while Stock and Watson [3] find that the inflation process is well represented by a univariate unobserved component time-varying trend-cycle model. Similarly, for the euro area, Fisher et al [4] highlight the good inflation forecasting performance of the random walk model, and Diron and Mojon [5] provide evidence that the central bank's objective targets yield more accurate forecasts than most inflation forecast models.

The aim of this paper is to contribute to the literature on forecasting inflation in the euro area at the short- and medium-term horizon. We consider forecasts for both the headline HICP and the HICP excluding energy and unprocessed food, which is often referred to as a measure of core

[1] We are very grateful to Michele Lenza and Philippe Weil for their valuable guidance and support. In addition, we would like to thank Fabio Busetti, Stefano Siviero, two anonymous referees and the editor for their helpful comments and suggestions. We also wish to thank Carlo Altavilla and Domenico Giannone for discussions about an early version of this paper. Angela Capolongo gratefully acknowledges financial support from the Fonds National de la Recherche Scientifique (FNRS). Angela Capolongo angela.capolongo@ulb.ac.be; Claudia Pacella claudia.pacella@bancaditalia.it. *The views expressed in the article are those of the authors and do not involve the responsibility of the Bank of Italy.* All remaining errors are ours.

inflation, meant to capture the most persistent component of consumer prices.
Specifically, we address the question of which information is crucial to forecast aggregate inflation in the euro area. While this question is quite traditional in the literature on inflation forecasting, the unique nature of the euro area, as a monetary union among highly interconnected but heterogeneous countries (see Figure 1), adds another possible layer of complexity to it. Moreover, it also raises the issue whether taking into account country information may improve the accuracy of euro area inflation forecasts. In our approach we consider three (overlapping) levels of information for




(a) HICP

(b) HICP excluding energy and unprocessed food

Fig. 1: Country-specific inflation rates. Note: Figure (a) shows the year-on-year percentage change of HICP index. Figure (b) shows the year-on-year percentage change of HICP index excluding energy and unprocessed food. The countries are Germany (DE), Spain (ES), France (FR), Italy (IT), and the euro area (EA).

the forecast exercise: *inflation key drivers*, *cross-country dynamic interactions* and *country-specific variables*. This is done through two steps.
*First*, we employ a large Bayesian Vector Autoregressive model (BVAR) for the biggest four euro area countries (Germany, Spain, France, and Italy) to produce individual country inflation forecasts, which are then aggregated to obtain a forecast for the euro area inflation. Concerning the variables chosen as inflation determinants, the model is broadly inspired to the "triangle model" introduced by Gordon [6][2] with a cross-country twist.
The modelling strategy is drawn from Altavilla et al [7], who, differently from our aim, use it to evaluate the effects of standard and nonstandard monetary policy shocks on the biggest four euro area countries. As shown by Banbura et al [8], a BVAR takes advantage of Bayesian shrinkage to tackle the high-dimensionality problem and allows to capture the dynamic inter-relationships between HICP components and their determinants in a fully unrestricted framework, as opposed to alternative models used in the literature, such as factor models (e.g.[9]), panel VARs (e.g. [10]), and global VARs (e.g. [11]).
We first validate our model, by comparing it with the random walk, which is often used in the literature as a benchmark and it corresponds to the prior in our model specification. We also compare the forecasting accuracy of our model to the one obtained with the Unobserved Component Stochastic Volatility (henceforth UC-SV) of Stock and Watson [3]. It is a simple but tough-to-beat

[2] It consists in an extension of the Phillips curve, realized using inflation data together with its three determinants: inflation persistence, demand-pull inflation and cost-push inflation

benchmark, especially in US. We find that our model with the three layers of information, i.e. the multi-country model with aggregate and country-specific variables (our *baseline model*), produces accurate pseudo out-of-sample inflation forecasts, comparable with the alternative benchmarks considered. This performance seems remarkable given that our sample covers not only the financial crisis, but also the period of the unexpected low inflation both in the US and the euro area (see [12,13]), in which our model remains able to generate accurate pseudo out-of sample forecast.

*Second*, we perform a step-by-step analysis of the model to shed light on the elements that are crucial for a more accurate forecast of euro area inflation. On the one hand, including as many variables as possible seems a good hedging strategy against omitting relevant information. On the other hand, this strategy risks to increase the complexity of the model without gains in terms of forecasting accuracy. Specifically, we carry out three additional exercises. First, we assess if the *inflation key drivers* play a determinant role in forecasting inflation, as the economic theory based on the Phillips curve would suggest. Therefore, we compare our baseline model to a BVAR including only the inflation rates in the four largest euro area countries, i.e. a pure multi-country autoregressive model of inflation. Second, in order to evaluate the importance of *cross-country dynamic interactions*, we compare the accuracy of the euro area inflation forecasts obtained in our baseline model with the one produced by aggregating the forecasts, one for each country, produced by country-specific models. Finally, our goal is to assess if *country-specific variables* matter for the euro area inflation forecasting accuracy. Hence, we build a model including only the euro area aggregates and we compare it with our baseline model.

The results show that, generally, our baseline multi-country BVAR compares favoura- bly to all the other benchmarks VAR models obtained by excluding the information layers, as discussed above. These results suggest that our modeling strategy, which consists in including the inflation key drivers and explicitly accounting for the multi-country dimension of the euro area, is supported by the data. The results in favor of the baseline model are stronger for what concerns HICP excluding energy and unprocessed food, while, for headline HICP, the aggregate euro area BVAR remains quite competitive. An interpretation of the latter finding is that HICP excluding energy and food has a stronger domestic component, for which it is beneficial to consider country-specific information. Instead, headline inflation dynamics are more strongly affected by global factors, like those driving commodity prices, which rather homogeneously affect the different euro area countries.

Turning to our contribution to the literature, we refer here to the studies strictly relevant for our analysis.

A first strand of literature concerns the euro area inflation forecasting. As reviewed by Banbura and Mirza [14], few studies focus on the out-of-sample inflation forecasting performance in the euro area, among these an exhaustive comparison of model performance is provided by Canova [15]. Our paper shows the importance of accounting for key inflation determinants to produce accurate euro area inflation forecasts, in line with the findings of Giannone et al [16].

Another strand of the literature has focused on the comparison between the inflation forecasting performance of aggregate and disaggregate models. Several forecasting exercises of euro area inflation have been performed by aggregating the forecasts of sub-components: sectors of economic activity [17,18,16], countries [19,20] or both [21]. While the theoretical literature, as proved in Kohn [22], agrees on the improvement obtained by using the disaggregate forecasts, with respect to a direct aggregate approach, the empirical evidence is still mixed. Our contribution to this literature is to build a large-scale model, free from restrictions, which accounts for multi-country dimensions, to forecast inflation in the euro area. In our case, the aggregate approach remains quite competitive for headline inflation, while for what concerns the inflation excluding energy and unprocessed food,

that is more driven by domestic factors, our disaggregate approach is more accurate. The underlying idea is in line with Monteforte and Siviero [23], who highlight the economic relevance of accounting for heterogeneity among the countries in the euro area. This suggests that policy making is more effective when supported by disaggregate (multi-country) rather than aggregate (area-wide) econometric models. A data-rich model, as the one we propose, presents indeed many advantages with respect to the country-models: it can be easily used for scenario analysis and for the assessment of a shock propagation, although these policy-relevant applications are not the focus of this paper.
The paper is structured as follows. Section 2 describes the data, the model specification and the estimation. Section 3 presents the results. Section 4 concludes.

# 2 Model

## 2.1 Data

The choice of the time series used to forecast inflation is entirely based on the economic theory. As anticipated in Section 1, we follow the idea of Gordon's "triangle model" [6] to identify the main inflation key drivers. The first driver is the built-in inflation, i.e. the inertial component that can be identified for example by lagged inflation, inflation expectations, cost of labor and prices of producers. The second driver is the demand-pull inflation factor, which is a measure of economic activity, as the Gross Domestic Product (GDP). The third driver is the cost-push shock, that can be measured by a global driver of inflation as the oil price, which affects inflation also through the exchange rate.
Therefore, our dataset is composed by 26 quarterly variables, which are classified in two groups: country-specific and euro area wide variables. The first group includes HICP overall index (HICP), HICP excluding unprocessed food and energy (HICPex), Unit Labour Cost (ULC), Producer Price Index (PPI), Gross Domestic Product (GDP) and the European Commission Consumer Survey on inflation for each of the four countries considered. The second group includes oil price and the Effective Exchange Rate (EER), which are common to the whole area. Moreover, in order to perform the comparison with an alternative model built to directly obtain euro area forecasts, we also consider the corresponding euro area aggregates of the variables listed above in the country-specific group [3].
Considering the abovementioned variables, we are able to detect the main factors affecting inflation in the euro area (see [24]). On the domestic side, the business cycle represents one of the main drivers of inflation. Movements of GDP, feeding into the labour market, tend to put pressure on wages. The latter can, in turn, push unit labor cost and, hence, affect the cost pressures for firms, which modify the producer prices and thereby inflation. This mechanism could be further amplified or moderated through the inflation expectations, which represent one of the key driving force of inflation. On the global side, several factors may affect the inflation development. We picked oil price because in the recent period it represents the main driver of global disinflationary shocks that has contributed to lead inflation in the euro area being persistently below target since 2011 (see [25,26]). The oil price exerts not only direct effects on inflation, via the energy sub-component, but also second-round effects on wage and price-setting, boosted by inflation expectations, which could affect medium-term price developments. The exchange rate lies in between global and domestic factors as juncture between the two, as the main pass-through channel.
The choice of the inflation data, key variables of the analysis, deserves further explanations. First,

[3] An appendix with further details is available upon request.

the decision of using HICP rather than other inflation measures, as for example the GDP deflator or the Consumer Price Index[4], is mainly driven by cross-country comparability reason. The HICP, differently from the alternative measures, is computed according to a harmonized approach, allowing for full comparability across euro area countries. Second, the choice of including in our model both the overall index (HICP) and the HICP excluding the most volatile components (HICPex), i.e. unprocessed food and energy, is motivated by two main reasons. On the one hand, considering both measures, we are able to perform and compare the forecasting accuracy of our model for both indices. The presence of the volatile components could indeed negatively affect the forecasting accuracy of HICP. On the other hand, the inclusion of both indicators allows us to account for direct, indirect and second-round effects on Euro Area inflation. Over short horizon the HICP overall has a better predictive power than the HICPex (see [27]). However, the latter is more informative for medium-term inflationary trends, because it excludes the more volatile components . Therefore, in our model we prefer considering both together to preserve their relationship and informational content, which is of particular interest since 2014, when HICP has dipped below the HICPex.
The countries considered are the four largest economies of the whole area: Germany, Spain, France, and Italy. They account for about 80% of whole Euro Area GDP growth. The normalized weights considered to aggregate the country-specific HICP forecasts did not change significantly in the period considered. In the forecasting exercise we use every period the latest available weights. The model is estimated over the sample period 1996Q1 - 2017Q2.

## 2.2 Estimation and Forecasting Methodology

In this section, we first describe the model estimation and then the forecasting methodology of our baseline model.
The model for performing short- and medium-term inflation projections is represented by the following vector autoregression (VAR) specification:

$$X_{i,t} = A_0 + A_1 X_{i,t-1} + \ldots + A_p X_{i,t-p} + e_{i,t}, \;\; e_t \sim N(0, \Sigma)$$

where $X_{i,t}$ $(i = 1, \ldots, N)$ is the $N$-dimensional matrix of data, $A_0, \ldots, A_p$ are the $N \times N$ matrices of the parameters, $e_{i,t}$ is a vector of size $N$ of the disturbances. The dataset used in the analysis is composed by $N = 26$ variables. The VAR is specified in log-levels and the variables are not pre-transformed to achieve stationarity. Since our analysis aims at detecting the dynamic properties of our quarterly dataset, we allow for five lags in the VAR model ($p = 5$). Hence, the total amount of parameters to estimate is 3757[5]. Since the sample available for the analysis has a short length (86 quarters), there is a clear over-parametrization. This issue is addressed by applying a Bayesian shrinkage.
Following Doan et al [28], we use a Minnesota prior for the autoregressive coefficients $(A_1, \ldots, A_p)$. Hence, we shrink the model's coefficients towards a naïve random walk model with drift, i.e. $X_{i,t} = \delta_i + X_{i,t-1} + u_{i,t}$. Moreover, we use a normal-inverted Wishart prior for the covariance matrix of the residuals, $\Sigma$. The scale parameter is a diagonal matrix $\Psi$ and it has $n+2$ degrees of freedom, that is

[4] This two measures differ from each other for the composition of the underlying basket. CPI includes only goods bought by consumers, both domestic and imported. The GDP deflator is a measures of prices of all domestic goods and services. They both rely on national definition and hence, they are not easy to aggregate across countries.

[5] It is given by the sum of $(26 \times 26) \times 5$ autoregressive coefficients, 26 parameters and $26 \times 27/2$ parameters of the covariances of the residuals.

$\Sigma \sim IW(\Psi, n+2)$, which implies $E(\Sigma) = \Psi$. Therefore, the prior distribution of the autoregressive coefficients, conditional on the covariance matrix of the residuals, is normal with the following mean and covariance:

$$E[(A_s)_{ij}] = \begin{cases} I_n & \text{if } s = 1 \text{ and } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$cov[(A_s)_{ij})(A_r)_{hm})] = \begin{cases} \lambda^2 \frac{\Sigma_{ij}}{(s^2\,\Psi_{ii})} & \text{if } m = j \text{ and } r = s \\ 0 & \text{otherwise} \end{cases}$$

where $\frac{\Sigma_{ij}}{\Psi_{ii}}$ accounts for the different scale and variability of the data, $1/s^2$ is the rate at which the prior variance decreases with an increasing lag length and $\lambda$ controls for the scale of the prior covariance. The latter hyperparameter determines the overall tightness of the prior. For $\lambda \to \infty$, the prior is defined as "diffuse", since we attribute a small weights to our beliefs, hence, the posterior expectations coincide with the ordinary least square estimations. For $\lambda \to 0$, vice versa, the prior is "dogmatic" centred at the random walk, since the posterior is equal to the prior, hence, the estimates are not influenced by the data. In the literature, this hyperparameter, $\lambda$, has been traditionally set on the basis of ad-hoc procedures. The first method, proposed by Litterman [29], consists in choosing the $\lambda$ that maximizes the out-of-sample forecasting performance of the model, calibrated as $\lambda = 0.2$. Banbura et al [8] show that, in order to get the desired in-sample fit, the tightness $\lambda$ of the prior needs to increase with the size of the model. To reduce the subjective choices in the setting of the prior informativeness, Giannone et al [30] introduce a hierarchical approach. The main idea is to interpret the model as a hierarchical model and treat the hyperparameters as additional unknown parameters, i.e. random variables on which we can conduct inference. Here we follow this approach on $\lambda$, whose posterior distribution is given by:

$$p(\lambda|y) \approx p(y|\lambda)p(\lambda)$$

where $y$ represents the data, hence, $p(y|\lambda)$ is the marginal likelihood and $p(\lambda)$ is the prior on the hyperparameter, also defined as hyperprior. For the latter we choose a proper but almost flat distribution, hence the shape of the posterior of $\lambda$ can be approximated with the marginal likelihood. We employ a recursive scheme to perform out-of-sample forecast of the HICP for the period 2006Q1-2017Q2, using an increasing data window using all the available data from 1996Q1. The highest forecast horizon, $H$, consists of 8 periods (two years). The target variable in our forecasting exercise is expressed in terms of $h$-period annualized average growth change in prices:

$$\hat{y}_{c,t+h|t} = \frac{400}{h} \log\left(\frac{\hat{x}_{c,t+h|t}}{x_{c,t}}\right)$$

where $x_{c,t}$ is the HICP (level) for the $c$-th country ($c = 1, \ldots, 4$).
The modelling strategy used to produce euro area inflation forecasts follows a bottom-up approach in a single model framework. This forecasting procedure consists of two steps. In the first step, we produce country inflation forecasts for $h$ quarters ahead for Germany, Spain, France, and Italy. In the second step, we aggregate the country-specific forecasts to obtain forecasts for euro area inflation using country weights to inflation at time $t$:

$$\hat{y}_{t+h|t} = \sum_{c=1}^{4} w_{c,t}\ \hat{y}_{c,t+h|t}$$

# 3 Forecasting Evaluation

In this section we present the results of the forecasting exercise. The forecasting accuracy of our model is measured both in terms of point forecasts and density forecasts.
First, to evaluate the point forecast we use the mean squared forecasting error (MSFE), which is the average of the squared the difference between the median of the predictive density forecast and the realized observation.

$$\text{MSFE}_h = \frac{1}{T-T_0+1} \sum_{t=T_0-h}^{T-h} (\hat{y}_{t+h|t} - y_{t+h})^2$$

where $\hat{y}_{t+h|t}$ is the median of the density forecast for horizon $h$ ($h = 1, \ldots, 8$), $T_0$ is the first forecast period, and $T$ is the last forecast period.
We compare the MSFE of our model to the one of a benchmark naïve model, as introduced by Theil [31]. The resulting metrics, so-called relative MSFE (RMSFE) or Theil's U-statistics, can be computed as follows:

$$\text{RMSFE}_h = \frac{\text{MSFE}_h}{\text{MSFE}_h^{RW}}$$

where $\text{MSFE}_h^{RW}$ is the MSFE of a random walk in levels with drift. If this ratio is bigger than one, the naïve model performs better than the model in terms of forecasting accuracy, and vice versa. In order to assess if the difference between two RMSFEs is statistically significant, we use the test of Diebold and Mariano [32], which is a $t$ test with HAC standard errors. The outcomes of the test should be considered as suggestive because we compare the forecast performance of nested models. Moreover, our forecasts are produced using a recursive scheme, while the reliability of the test has been proved only for forecasts obtained via a rolling scheme (Giacomini and White [33]).
Second, to assess the density forecast we use the average log predictive score, $LS_h$ that is the arithmetic mean of the log scores, $\text{LS}_{t,h}$, computed in each period as:

$$LS_h = \frac{1}{T-T_0+1} \sum_{t=T_0-h}^{T-h} LS_{t+h|t} = \frac{1}{T-T_0+1} \sum_{t=T_0-h}^{T-h} ln\left(f\left(y_{t+h}|I_t\right)\right)$$

where $f\left(y_{t+h}|I_t\right)$ is the predictive density for $y_{t+h}$ constructed using information up to time $t$ and evaluated at the realized $y_{t+h}$. It follows that the same MSFE can correspond to very different log score depending on the uncertainty around the median, i.e. the second moment of the distribution. A more accurate forecast is characterized by a greater average log score. In our case, we use a Gaussian kernel approximation of the predictive density for all models. The log scores of two models can be compared using the test introduced by Amisano and Giacomini [34]. It is a $t$ test, whose null hypothesis states the absence of difference between the weighted logarithmic scores of two models. In our framework, we use an unweighted version of the test with HAC standard error.
To evaluate the forecasting accuracy of our baseline model we follow three main steps. First, in the subsection 3.1, we assess the performance of the baseline BVAR model with respect to the traditional benchmarks used in the literature, i.e. the random walk and the UC-SV model. Second, in the subsection 3.2, we perform a step-by-step analysis of the model, by relaxing each assumption at a time, to shed light on the elements necessary to improve the euro area inflation forecasts.

## 3.1 Model Validation

We first evaluate the overall performance of the baseline model by making a comparison, in terms of forecasting accuracy, with the random walk model with drift. This naïve model forecasts the future

inflation as the average historically observed inflation without any further variable for predicting the future path of inflation. We chose the random walk as benchmark for two main reasons. First, as shown in the literature, it often exhibits good forecasting accuracy for inflation. Second, it is the prior for our model specification, hence, as argued by Banbura et al [8], if the model outperforms the random walk, this implies that it is able to extract valuable information from the sample.
We then compare the forecasting accuracy of a simple benchmark model often used in the literature, the UC-SV model of Stock and Watson [3]. The UC-SV is an univariate unobserved component model with stochastic volatility and consists in decomposing inflation into a stochastic trend and a cycle, whose shocks have time-varying variances. The model is defined as follows:

$$\pi_t = \tau_t + e^{\frac{1}{2}h_t}\varepsilon_t^{\pi}, \quad h_t = h_{t-1} + \omega_h \varepsilon^h$$
$$\tau_t = \tau_{t-1} + e^{\frac{1}{2}g_t}\varepsilon_t^{\tau}, \quad g_t = g_{t-1} + \omega_g \varepsilon^g$$

where $\varepsilon_t^{\pi}, \varepsilon_t^{\tau}, \varepsilon^h, \varepsilon^g \sim N(0,1)$, and $\omega_h, \omega_g$ are parameters to be estimated. The point forecast for inflation at horizon $t+h$ is obtained as the estimate of the current trend:

$$\hat{\pi}_{t+h|t} = \hat{\tau}_t.$$

The results, summarized in Table 1, show that the baseline model (henceforth *BVAR-Base*) outperforms the random walk and produces forecasts comparable to the UC-SV, both for the HICP and HICPex. In particular, by looking at the Theil's U-statistics, i.e. the relative mean squared error, we can reach two main conclusions. First, the forecasts produced by the BVAR-Base are uniformly more accurate than the random walk model forecasts up to two-year ahead horizon. Second, the UC-SV model produces worse inflation forecasts than the BVAR-Base for horizon longer than one year, for both the inflation measures considered. These results are confirmed by the Diebold and Mariano test [32] (Table 2).

Table 1: RMSFE of benchmark models

| | HICPex | | HICP | |
|---|---|---|---|---|
| horizon | BVAR-Base | UC-SV | BVAR-Base | UC-SV |
| one quarter | 0.76 | 0.54 | 0.92 | 0.78 |
| two quarters | 0.75 | 0.47 | 0.88 | 0.91 |
| one year | 0.74 | 0.88 | 0.89 | 1.02 |
| two years | 0.97 | 1.50 | 0.92 | 1.39 |

Relative MSFE with respect to benchmark models. Note: A value smaller than 1 indicates that the model outperforms the RW.

## 3.2 A step-by-step analysis

In this section, we perform a step-by-step analysis of our baseline large multi-country model, which includes three levels of information: *inflation key drivers*, *cross-country dynamic interactions* and *country-specific variables*. The goal is to validate our model, by analysing if every group of variables included does help improving the projections of the euro area inflation. We proceed in three steps: (i) we remove from the baseline model one piece of information; (ii) we perform the Euro Area inflation forecasts by means of the new model obtained in the first step; (iii) we make a comparison

Table 2: Diebold and Mariano test with respect to benchmark models

| horizon | | HICPex | HICP |
|---|---|---|---|
| | | BVAR-Base | BVAR-Base |
| one quarter | RW | 3.05 | 1.11 |
| | UC-SV | -1.46 | -0.94 |
| two quarters | RW | 2.68 | 1.51 |
| | UC-SV | -1.97 | 0.15 |
| one year | RW | 2.45 | 1.30 |
| | UC-SV | 0.88 | 0.98 |
| two years | RW | 0.19 | 0.66 |
| | UC-SV | 3.56 | 2.78 |

Note: A negative value of the $t$ statistic indicates that the model on the tables row is more accurate than the model on the tables column.

between the forecast performance of this model and our baseline model to assess if any improvement is reached. Therefore, we introduce three alternative models to our baseline multi-country large BVAR, BVAR-Base.

First, we compare our model to a BVAR including only the inflation rates of the four largest countries, henceforth *BVAR-H*. Our goal is to understand whether the information provided by *inflation key drivers* is valuable to predict euro area inflation. The results show that the BVAR-Base has higher predictive power than the BVAR-H, for both HICPex and HICP . When considering the point forecast, in Table 3, the MSFE shows that the model excluding all key determinants, BVAR-H, generates better forecasts than the random walk model only at a very short horizon. As shown in Table 4 this result is very robust: the $t$ statistic is negative at every horizon for both the measures of inflation, thus pointing to higher forecasting accuracy of the BVAR-Base than the BVAR-H. If we consider the entire predictive density this result is confirmed. Table 5 shows that the average difference of log scores is always positive, meaning that the BVAR-Base performs better than the BVAR-H.

Second, we analyze the euro area inflation forecast obtained by our BVAR-Base with respect to those produced by aggregating the four forecasts, one for each country, generated by country-specific BVARs, henceforth *BVAR-C*[6]. Our aim is to assess whether *cross-country dynamic interactions* matter for the euro area inflation forecasting accuracy. The comparison with this alternative methodology allows to understand if taking into account all existing dynamic relationships among variables of different countries really matters for forecasting euro area inflation. We find that the forecasting accuracy, for HICP and HICPex, differs among short- and medium-term forecasts. This result is valid for both point (Table 3) and density forecast (Table 5). In particular, for the first year of the forecasting horizon, the forecasts for euro area inflation obtained through country-specific models, BVAR-C, outperform those of the large multi-country model, BVAR-Base. In the second year, the forecasting exercise shows an opposite evidence. This could lead to conclude that in the very short horizon the informational content of cross-country dynamic interactions is actually disregarded. However, it becomes an important element for the medium-term horizon inflation forecasts.

Third, we compare our disaggregate BVAR-Base, which is built using an indirect two-step approach, to an aggregated model with only euro area variables, which uses a direct approach, henceforth *BVAR-EA*. The final purpose is to understand if a model exploiting *cross-country heterogeneity*, as

[6] The country forecasts are aggregated using normalized country weights to inflation.

the BVAR-Base, is able to produce accurate forecasts for the euro area as a whole. By looking at Tables 3 and 5, we can detect different findings for HICP and HICPex. For HICP, the model including euro area aggregates, BVAR-EA, shows superior forecasting accuracy at all horizons, although the improvement is small in magnitude. For HICPex, the multi-country model, BVAR-Base, performs better in terms of forecasting accuracy for the medium-term, that is for one year onward. The different results between the two indices might potentially be explained by the diverse trend of HICP and HICPex in the out-of-sample period considered (see Figure 1). While the HICP differentials, defined as the difference between country inflation rates and euro area inflation rate, have been decreasing since the beginning of the financial crisis, the HICPex differentials have remained quite large. Therefore, in the last case there is a great amount of country information to exploit. Albeit the short sample size available and the big number of country-level variables, our multi-country model produce results comparable to a smaller model including euro area aggregates.

Table 3: RMSFE of alternative models

| | HICPex | | | | HICP | | | |
|---|---|---|---|---|---|---|---|---|
| horizon | BVAR-Base | BVAR-H | BVAR-C | BVAR-EA | BVAR-Base | BVAR-H | BVAR-C | BVAR-EA |
| one quarter | 0.76 | 0.84 | 0.51 | 0.50 | 0.92 | 0.98 | 0.73 | 0.66 |
| two quarters | 0.75 | 0.83 | 0.50 | 0.56 | 0.88 | 1.01 | 0.78 | 0.72 |
| one year | 0.74 | 0.89 | 0.69 | 0.83 | 0.89 | 1.04 | 0.86 | 0.81 |
| two years | 0.97 | 1.21 | 1.26 | 1.12 | 0.92 | 1.12 | 0.91 | 0.57 |

Relative MSFE with respect to the random walk model of the pseudo out-of-sample forecasts computed by using different models. Note: A value smaller than 1 indicates that the model outperforms the RW.

Table 4: Diebold and Mariano test with respect to alternative models

| | HICPex | | | HICP | | |
|---|---|---|---|---|---|---|
| horizon | BVAR-H | BVAR-C | BVAR-EA | BVAR-H | BVAR-C | BVAR-EA |
| one quarter | -1.53 | 2.06 | 1.63 | -1.45 | 1.94 | 2.09 |
| two quarters | -1.57 | 1.82 | 0.97 | -2.64 | 1.13 | 1.19 |
| one year | -2.54 | 0.37 | -0.33 | -2.77 | 0.25 | 0.49 |
| two years | -2.38 | -1.41 | -0.39 | -2.41 | 0.12 | 2.45 |

Note: A negative value of the $t$ statistic indicates that the BVAR-Base is more accurate than the model on the table's column.

# 4 Conclusions

In this paper, we build a large multi-country Bayesian VAR model for the EA. It is able to capture information on *inflation key drivers*, *cross-country dynamic interactions* and *country-specific variables*. We proceed in two steps.

First, we validate the model, by measuring its predictive power in comparison to benchmarks traditionally used in the literature. Covering not only the financial crisis, but also the period of the unexpected low inflation in the euro area, we find the model to produce accurate pseudo out-of-sample inflation forecasts for both short- and medium-term horizon. Therefore, we can conclude that our baseline model contains valuable information to forecast inflation in the euro area and, as such, it can be applied for empirical studies.

Table 5: Average difference between log scores of alternative models

| | HICPex | | | HICP | | |
|---|---|---|---|---|---|---|
| horizon | BVAR-H | BVAR-C | BVAR-EA | BVAR-H | BVAR-C | BVAR-EA |
| one quarter | 0.04 | -0.18 | -0.22 | 0.04 | 3.02 | -0.23 |
| | *(0.03)* | *(0.11)* | *(0.13)* | *(0.03)* | *(2.43)* | *(0.10)* |
| two quarters | 0.08 | -0.12 | -0.17 | 0.46 | 2.86 | 0.37 |
| | *(0.05)* | *(0.21)* | *(0.27)* | *(0.37)* | *(2.26)* | *(0.53)* |
| one year | 0.26 | 0.26 | 0.14 | 0.61 | 5.58 | -0.31 |
| | *(0.15)* | *(0.39)* | *(0.53)* | *(0.40)* | *(5.04)* | *(0.36)* |
| two years | 0.49 | 0.39 | -0.34 | 0.68 | 1.52 | -0.79 |
| | *(0.40)* | *(0.49)* | *(0.22)* | *(0.58)* | *(1.61)* | *(0.59)* |

Note: A negative value indicates that the model outperforms the BVAR-Base. HAC standard errors are in parentheses.

Second, we perform a step-by-step analysis of the model to shed light on which features are more crucial for forecasting euro area inflation. By comparing our baseline BVAR with three different alternatives, we are able to analyse the importance of the informational content of our model to forecast euro area inflation. This is of fundamental importance in our forecasting exercise, because if on the one side, including as much information as possible is indeed desirable for forecasting purpose, on the other side, the risk of overfitting could be increasing the instability, hereby deteriorating the model forecasting accuracy. Our results show that the large multi-country BVAR model presents a good performance in comparison to the alternatives; in some cases, as for the HICP excluding energy and unprocessed food in the medium-term horizon, the model is able to produce an improvement in terms of forecasting accuracy with respect to all the alternatives considered in the analysis.

Therefore, we can conclude that including information concerning *inflation key drivers* and *country-level variables* in a multi-country model may help improving the forecasting accuracy of inflation in the euro area, compensating the instability due to the data-richness and the period of high uncertainty.

## References


1. Faust, J., Wright, J.H.: Forecasting inflation. Handbook of economic forecasting **2**A, 3–56 (2013).
2. Atkeson, A., Ohanian, L.E.: Are Phillips curves useful for forecasting inflation? Federal Reserve Bank of Minneapolis. Quarterly Review-Federal Reserve Bank of Minneapolis **25**(1), 2–11 (2001).
3. Stock, J.H., Watson, M.W.: Why has US inflation become harder to forecast? Journal of Money, Credit and banking **39**(1), 3–33 (2007).
4. Fischer, B., Lenza, M., Pill, H., Reichlin, L.: Monetary analysis and monetary policy in the euro area 19992006. Journal of International Money and Finance **28**(7), 1138–1164 (2009).
5. Diron, M., Mojon, B.: Forecasting the central banks inflation objective is a good rule of thumb. ECB Working Paper **564** (2005).
6. Gordon, R.J.: Inflation, Flexible Exchange Rates, and the Natural Rate of Unemployment. In: Workers, Jobs, and Inflation. Brookings Institution (1982).
7. Altavilla, C., Giannone D., Lenza M.: The Financial and the Macroeconomic Effects of the OMT Announcements. International Journal of Central Banking **12**, 29–57 (2014).

8. Banbura M., Giannone, D., Reichlin, L.: Large Bayesian vector auto regression. Journal of Applied Econometrics **25**(1), 71–92 (2010).
9. Forni, M., Hallin, M., Lippi, M., Reichlin, L.: Do financial variables help forecasting inflation and real activity in the euro area? Journal of Monetary Economics **50**(6), 1243–1255 (2003).
10. Dees, S., Gunther, J.: Forecasting Inflation Across Euro Area Countries and Sectors: A Panel VAR Approach. Journal of Forecasting **36**(4), 431–453 (2017).
11. Pesaran, M.H., Schuermann, T., Smith, V.L.: Forecasting economic and financial variables with global VARs. International journal of forecasting **25**(4), 642–675.
12. Coibion, O., Gorodnichenko, Y.: Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation. American Economic Journal: Macroeconomics **7**(1), 197–232 (2015).
13. Bobeica, E., Jarociski, M.: Missing Disinflation and Missing Inflation: A VAR Perspective. International Journal of Central Banking, International Journal of Central Banking **15**(1), 199–232 (2019).
14. Banbura, M., Mirza, H.: Forecasting euro area inflation with the Phillips curve. mimeo (2013).
15. Canova, F.: G-7 inflation forecasts: Random walk, Phillips curve or what else? Macroeconomic Dynamics **11**(1), 1–30 (2007).
16. Giannone, D., Lenza, M., Momferatou, D., Onorante, L.: Short-term inflation projections: a Bayesian vector autoregressive approach. International journal of forecasting **30**(3), 635–644 (2014).
17. Hubrich, K.: Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? International Journal of Forecasting **21**(1), 119–136 (2005).
18. Hendry, D.F., Hubrich, K.: Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. Journal of business & economic statistics **29**(2), 216–227 (2011).
19. Marcellino, M., Stock, J.H., Watson, M.W.: Macroeconomic forecasting in the euro area: Country specific versus area-wide information. European Economic Review **47**(1), 1–18 (2003).
20. Cristadoro, R., Saporito, G., Venditti, F.: Forecasting inflation and tracking monetary policy in the euro area: does national information help? Empirical Economics **44**(3), 1065–1086 (2013).
21. Benalal, N., Diaz del Hojo, J.L., Landau, B., Roma, M., Skudelny, F.: To aggregate or not to aggregate? Euro area inflation forecasting. ECB Working Paper **374**
22. Kohn, R.: When is an aggregate of a time series efficiently forecast by its past? Journal of Econometrics **18**(3), 337–349 (1982)
23. Monteforte, L., Siviero, S.: The economic consequences of euroarea macro-modelling shortcuts. Applied Economics **42**(19), 2399–2415.
24. European Central Bank: Domestic and global drivers of inflation in the euro area. Economic Bulletin, 72–96 (2017).
25. Draghi, M.: Introductory statement to the press conference. European Central Bank, Frankfurt am Main, 22 January 2015
26. Ciccarelli, M., Osbat, C.: Low inflation in the euro area: causes and consequences. ECB Occasional Paper Series **181** (2017).
27. European Central Bank: The relationship between HICP inflation and HICP inflation excluding energy and food. ECB Economic Bulletin **2** (2016).
28. Doan, T., Litterman, R., Sims, C.: Forecasting and conditional projection using realistic prior distributions. Econometric reviews **3**(1), 1–100 (1984).
29. Litterman, R.: A Bayesian Procedure for Forecasting with Vector Autoregression. MIT Working paper (1980).
30. Giannone, D., Lenza, M., Primiceri, G.E.: Prior selection for vector autoregressions. Review of Economics and Statistics **97**(2), 436–451.
31. Theil, H.: Applied econometric forecasting. Rand McNally. Chicago (1996).
32. Diebold, F., Mariano, R.S.: Comparing Predictive Accuracy. Journal of Business & Economic Statistics **13**(3), 253–263 (1995).
33. Giacomini, R., White, H.: Tests of conditional predictive ability. Econometrica **74**(6), 1545–1578 (2006).
34. Amisano, G., Giacomini, R.:. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business & Economic Statistics **25**(2), 177–190 (2007).

35. Constancio, V.: Understanding inflation dynamics and monetary policy in a low inflation environment (2015)

# On the automatic identification of Unobserved Components Models

Diego J. Pedregal* and Juan R. Trapero**

*Industrial Engineering School
**Faculty of Chemical Science and Technology
Universidad de Castilla-La Mancha, Ciudad Real, Spain

## 1 Abstract

Automatic identification of time series models is a necessity once the big data era has come and is staying among us. This has become obvious for many companies and public entities that has passed from a crafted analysis of each individual problem to handle a tsunami of information that have to be processed efficiently, online and in record time. Automatic identification tools are the usual way to go in the Machine Learning area and also in some other statistical approaches, but it never has been tried out in Unobserved Components models (UC). There are many reasons for this. Firstly, UC have been developed mainly in academic environments with the purpose of research, with little dissemination among practitioners for their everyday use in business and industry. Secondly, the widely-held feeling that UC models do not really have anything relevant to add to exponential smoothing methods has deterred its use in practice. Third, UC models are usually identified by hand, with automatic identification being very rare. Finally, software is rather scarce compared with other methods.

After much experimentation, the automatic forecasting algorithm proposed below performs remarkably well in practice. The algorithm avoids estimating all the possible models, by selecting models in a hierarchical and logical way. Hence, the computation burden is lighter and speed is enhanced considerably. This is the first time that an algorithm of this nature is proposed in the literature on UCs. This algorithm may be used for forecasting, but also for automatic and reliable modelling of components models, useful for operations like detrending data, seasonal adjustment, signal extraction in general, etc.

The algorithm proceeds along the following steps:

- Step 1: Select either to use the Box-Cox transformation or not [3]. This step is left to the user discretion, because of a number of different habits that have developed over time. Some users would ignore it completely, others would prefer ever to use the log transform, while some others would select it according to some criteria. The approach by [5] is preferred here. It consists on finding the transformation parameter that minimizes the variation coefficient. The main advantage of this approach is that it is not model dependent and may be run independently to the rest of the algorithm.

- Step 2: Trend test. Whether a trend is present in the data is decided on the basis of the Augmented Dickey-Fuller unit root tests [4] with a number of delays automatically chosen by information criteria such as Akaike's (AIC) or Schwarz's (BIC), i.e.,

$$AIC = -2ln(L^*) + 2k$$

$$BIC = -2ln(L^*) + ln(T)k$$

  where $L^*$ is the likelihood value at the optimum, $T$ is the length of the time series and $k$ the number of parameters in the model.
- Step 3: Seasonality test. A time series is considered to have seasonality based on a conservative and quick test. Such test assumes seasonal time series those with a one-year lag autocorrelation coefficient with a p-value smaller than 10% (or a t-test greater than 1.645 in absolute value [2]). A second sub-step here is selecting the number of harmonics on a regression of the de-trended data on sines and cosines on the fundamental frequency and its harmonics. The preliminary trend (if Step 2 indicates its presence) is calculated by a standard UC model.
- Step 4: UC model selection. Several models are tried out and the best is selected according to the minimization of any information criterion, either the AIC or BIC. The set of models to search for are a subset of all the possible combinations of trends (none, Random Walk, Local Linear Trend, Damped Trend) and seasonal components (none, all harmonics with equal variance, all harmonics with different variances) and irregular (none, white noise). the search is restricted to a subset of the whole bunch of previous possible combinations because the final range of models depends on the results obtained in the previous steps. For example, if there is no trend, there is no need to search among models with trends and the computation time is considerably reduced.
- Step 5: ARMA model selection. A low order non-seasonal ARMA model is then selected for the innovations of the UC identified so far. ARMA models are selected according to AIC or BIC minimization, following the automatic procedure developed by [7], but with all ARMA models estimated in regression form using Hannan-Rissanen approximation [6].
- Step 6: If an ARMA model is detected, then the full UC with the ARMA model embeded is jointly re-estimated by Maximum Likelihood and its AIC or BIC values computed. If the information criterion is better it is retained as the best model. Otherwise, the best option is the UC without ARMA perturbations. One important point about ARMA identification is that AR and MA orders should be smaller than the seasonal period to avoid confusion with the seasonal component and they should be estimated with stationarity and invertibility constraints in order to avoid overlapping with the trend component [1]. A non-stationary ARMA model would tend to model a trend, while a non-invertible one would show unit roots that would cancel out with unit roots included in the trend altering the nature of the components.

A piece of software has been developed in C++ that is integrated into the R environment via the RcppArmadillo package. The preliminary results are encouraging and the forecasting results suggest that UC models are powerful potential forecasting competitors to other well-known methods. Though there are several pieces of software available for UC modeling, this is the first implementation of an automatic algorithm for this class of models, to the authors knowledge.

## References

1. Ansley, CF and Kohn, R: A Note on Reparameterizing a Vector Autoregressive Moving Average Model to Enforce Stationarity. Journal of Statistical Computation and Simulation, 24, 99-106 (1986)
2. Assimakopoulos, V. and Nikolopoulos, K.: The theta model: a decomposition approach to forecasting. International Journal of Forecasting, 16, 521 - 530 (2000)
3. Box, GEP, and Cox, DR: An analysis of transformations. Journal of the Royal Statistical Society, Series B, 26, 211-252 (1964)
4. Dickey, D. and W. Fuller: Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. Econometrica, 49, 1057-1072 (1981)
5. Guerrero, Vctor M.: Time-series analysis supported by power transformations. Journal of Forecasting, 12, 37-48 (1993)
6. Hannan, E. J. & Rissanen, J.: Recursive estimation of mixed autoregressive moving average order. Biometrika, 69, 81-94 (1983)
7. Hyndman, RJ and Khandakar, Y: Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27, 1-22 (2008)

# Theory and Simulaion of Procrastination: The Before and After the Releasing of a Cash Credit

Huber Nieto-Chaupis

Universidad Autonoma del Perú
Programa de Ingenierí a de Sistemas
hubernietochaupis@gmail.com

**Abstract.** Commonly the apparition of procrastination in some social layers are not well understood in all. In this paper we present some stochastic simulations by which customers fail to pay the monthly fee when have accessed to a cash credit from a bank. In according to these computational studies, one of the main causes is the expended time either the bank or customer to study and analyze the possible scenarios by which one of them is in risk. Thus, customers might be in a kind of pseudo-procrastination even the cash is not yet released because the rapid decision to access the cash becomes a type of loss.

## 1 Introduction

Alice got a call from a Cars Company agent to inform her that she applies for an offer to buy a car within the next 24 hours after of receiving the call. The first reaction of Alice is the rapid communication with a Bank.

In this manner, she is searching for cash credit. She contacted a Bob, a Bank agent whom takes the case and paid attention to the cash credit although there is not any decision in next 24 hours. Bob is honest and tell her that a sincere answer would be after of 7 working days. During these days, Bob:

(i) Perform an analysis on the historical credit of Alice in the last 5 years,

(ii) Ask to Alice genuine information about real incomes and outcomes per month,

(iii) Perform a crude statistic about on the rate of interest,

(iv) Got numbers about what is of entire convenience to the Bank and what are the profitable scenarios,

(v) Study realistic scenarios of procrastination and legal actions against her.

When all issues are entirely clear and Bob makes sure that the Bank has in front a profitable opportunity, then the preparation to release the cash of the Bank is in the way.

On the other side, Alice is expecting a positive rate from the Bank respect to her case. During the days previous to receive the cash, Alice

(i) Do not take into account the possible risks as to a get a cash credit,

(ii) Paid a minimal time as to evaluate scenarios of a debt that She cannot assume in cases where She is removed from her employment,

(iii) Is entirely submerged in subjective facts that has nothing to do with the fact of making a cash credit with a Bank in a serious manner.

Clearly, the Bank have shown a genuine seriousness as to make the approval of the cash credit, and have invested enough time to evaluate all possible scenarios, in contrast to Alice, that she has preferred to paid less time that the one taken by the Bank. This establishes a kind of asymmetry in timelines with respect to a contract of cash credit between and customer and the bank.

In this paper, we review these asymmetries through mathematical methodologies that might be relevant as to understand the scenarios of credit [1] by which one of both parties is free of risks.

Since credits are decided in the basis of the good history of credits of customer, so that bank applies concrete algorithms to define is a credit is approved or not [2][3].

Normally all these actions are optimized so the bank minimizes resources being the most important: the time that it takes previous to release the cash.

However, often in the side of the customer, the decision to access to a money credit is too short compared to the one of the bank and reflecting the lack of a solid and consistent analysis as seen in the unavoidable apparition of procrastination that will deteriorate the relation of bank and customer [4][5][6][7][8].

The paper is structured as follows: in second section we review all those quantitative concepts that constructs a business relationship between a customer and bank [9].

In third section, we propose algorithms that are applied to the case where N customers are applying for cash credits but each one is managing his own strategy of debt [10]. For this, we use a mathematical model that allows us to quantify the advantages and disadvantages for the different schemes of payment.

In fourth section, we derived the success or fail of the relationships between customers and bank, and underlined the importance of having a scheme that minimizes the possible asymmetries in timelines.

## 2 General Concepts

predominate in a relationship between a customer and a bank institution (state or private) aiming for a cash credit would be the following:

- Sincere amount for cash credit,
- Real purpose for the credit,
- Availability to make a cash credit,
- Real schedule for payments,
- Early identification of risks,
- Compromise to carry out the payments with the most realistic interest rates,
- A minimal level of confidence,
- Continue communication.

## 3 Credit Algorithms

The action for releasing a cash credit might involve the usage of well-designed algorithms. The entire meaning of a robust algorithm would depend on the usage of quantities that are part of the system defined by: Bank and customer. Classically, the dynamics of an algorithm has as object the action of make a loan with the purpose of provide a cash support [11] with the clear assumption that it shall be returned in the short or middle term, even in the long term but depending on the released cash. Therefore, for the building of the algorithm, the crucial inputs are seen to be the cash, and the payment schedule. However, a sincere algorithm has as target to assess all possible scenarios where the relationship between Bank and customer might be broken or enter in conflict as cause of the procrastination.

**Algorithm 1: Fast Decision of One Customer:**

Below is listed the lines that correspond to a cash credit transaction between a Bank and 1 customer. We assume that the customer makes the decision in a time that does not fit with the one of the Bank.

**1** Enter cash approved = Y
**2** Enter months to be paid = J
**3** Enter monthly amount = M
**4** Enter factor of Risk = r
**5** W(0) = 0.0
**6** W(1) = M (first payment)
**7** DO j = 1, J
**8** W(j) = m (monthly payment)
**9** W(j) = W(j) + W(j-1)
**10** Y(j) = W(j)
**11** IF( Y(j) iS NOT EQUAL TO U(J))THEN
**12** Q = Q + 1
**13** IF( j = J )THEN
**14** EF = (J - Q)/J
**15** ENDIF
**16** ENDIF
**17** ENDDO

It is actually an ordinary algorithm where the main inputs are the approved cash and the payment schedule that essentially the number of months to be paid. It should be noted that lines 2 and 3 does not necessarily are the same. It is clear that line 3 that is the monthly amount to be paid is including the interest rate corresponding to that month. In line 4 the risk factor is introduced. Lines 5 and 6 are generalities, whereas in line 7 the loop overall scheduled months is initialized. In line 8 the paid amount as fixed is contracted when the cash is decided. Line 9 performs the summation of the paid amounts, Thus, for each j there is an accumulated amount which should end on the total released cash exact as the declared in line 1. In line 10, this accumulated amount is compared with the expected value calculated by the Bank. In other words, Y(j) must be exactly equal to W(j). In line 11, the inequality between both amounts would

lead to the apparition of procrastination. Therefore, the month that were not paid in time, are summarized in line 12. Between lines 13 and 15, the efficiency of the payment is estimated in a straightforward manner as seen in line 14. When the customer keeps a solid discipline as to payments, this efficiency is 1, that means that all scheduled months were paid without delays in the programmed dates.

**Algorithm 2: Decision of One Customer:**

Below is listed the lines of the algorithm that analyzes the relationship between a customer whom has evaluated the possible scenarios that might be disadvantageous for him. While a cash of amount R is desired by the customer, the choice of a realistic amount Y by which the Bank can release, appears how a healthy point for both parties.

**1** Enter Cash requested = R
**2** Expected Scenarios Customer = Ec
**3** Expected Scenarios Bank = Eb
**4** Enter Cash approved = Y
**5** Enter Months to be paid = J
**6** Enter Monthly Amount = M
**7** Enter Factor of Risk r = Ec/Eb
**8** W(0) = 0.0
**9** W(1) = M (first payment)
**10** DO j = 1, J
**11** W(j) = m (monthly payment)
**12** W(j) = W(j) + W(j-1)
**13** Y(j) = W(j)
**14** IF( Y(j) iS NOT EQUAL TO U(J))THEN
**15** Q = Q + 1
**16** IF( j = J )THEN
**17** EF = (J - Q)/J
**18** IF(r GREATER THAN EF) THEN
**19** J = J + JE
**20** ENDIF
**21** ENDIF
**22** ENDIF
**23** ENDDO
**24** END

Between lines 1 and 7, are defined all these parameters that are necessary for perform calculations. While W(0) is null as written in line 8, the first payment in according to line 9 is given by M. A loop over all schedule months is carry out from line 10 in order to register payments being these saved in vector W(j) when the amount m is already paid. Thus, the amounts are added in line 12, thus in line 13 Y(j) describes the actual amount for the j month. We ask about if W(j) becomes same as the expected by the Bank, i.e., if the customer is carrying out the program of payments correctly and without delays. When this happens, then we count the months that did not were paid in time as seen in line 15. When the

loop is finalized we estimate in line 17 the efficiency EF = (J - Q)/J calculated in basis of integer numbers. In line 18 the risk factor r is compared with EF. Actually the risk factor has a random nature in the sense that the number of expected scenarios reviewed and assessed by the customer is not fixed and might depend on a large number of factors. If this risk is bigger than the efficiency, then it means that the Q is large, fact that implies that the number of months should be reconfigured to one more realistic. This is explicitly written in line 19 where J + JE is the new number of months necessary to accomplish the total payment of the debt. The relationship between EF and r is also manifested in the number of months that the customer has left pay to. While a small Q tells us that that the procrastination is high, an r small would indicates us that the customer have manifested to some extent an interest to become educated with the issue of acquire a debt to be paid in the long or middle term.

**Algorithm 3: Slow Decision by N Customers:**

Below is the full generalization of previous algorithms 1 and 2, by which is understood to be applied to N customers, being this number fully arbitrary. In contrast to Algorithm 2 we introduce the trial parameter that is a vector namely 0.99 - Exp[-AS(n)-CR2]where As(n) defined in line 7 and CR a mean value that measures the departure of the states of recoverability and lost. As(n) can also be seen as the one that measures the capability of the customer to assess the consequences of acquiring a cash credit. Of course, this number might be entirely random and clearly one expects two well-defined cases: (i) As(n) is near to 1 by which the customers might to assume a solid position to evaluate the consequences after of receiving the credit. The scalar quantities defined in Algorithm 2 are now defined as vectors Y(n), M(n), D(n) and r(n). Although the structure is in essence same as Algorithm 2, in line 25 the trial parameter is submitted to be compared with the efficiency (line 23) by which the cases where it turns out to be smaller than the efficiency, constitutes a full fault of customer to face seriously his debt. In these cases the new monthly amount is redefined to be D(n,j) = As(n)D(n,j) by emphasizing the fact that the amount might be smaller than the initial agreement but by making the schedule a bit more largest than the initial one.

**1** Enter number of customers = N
**2** M(1,0)=0.0
**3** DO n = 1 , N
**4** Enter cash requested = R(n)
**5** Expected Scenarios Customer = Ec(n)
**6** Expected Scenarios Bank = Eb(n)
**7** Trial Vector: As(n) = Ec(n)/Eb(n)
**8** Trial Parameter: 0.99 - Exp[-AS(n)-CR2]
**9** Enter Cash approved = Y(n)
**10** Enter Months to be paid = M(n)
**11** Enter Monthly Amount = D(n)
**12** Enter Factor of Risk = r(n)
**13** DO j = 1, M(n)

**14** W(n,0) = 0.0
**15** W(n,1) = F(n,1)(first payment)
**16** W(n,j) = D(n,j)(monthly payment)
**17** W(n,j) = W(n,j) + W(n,j-1)
**18** Y(n,j) = W(n,j)
**19** U(n,j) = M(n,j) + M(n,j-1)
**20** IF( Y(n,j) LESS THAN U(n,j))THEN
**21** Q(n,j) = Q(n,j) + 1
**22** IF( j = M(n) )THEN
**23** IF( n = N )THEN
**24** EF(n,j) = ( M(n,j)  Q(n,j))/M(n,j)
**25** ENDIF
**26**  IF ( As(n) ¡ EF(n,j) ) THEN
**27** D(n,j) = As(n)D(n,j)
**28** ENDIF
**29** ENDIF
**30** ENDIF
**31** ENDDO
**32** ENDDO
**33** END

Thus one can arrive to define the sincere monthly payment as defined by:

$$D(n,j) = \frac{E_c(n)}{E_b(n)} \times \frac{M(n) - Q(n)}{M(n)} \tag{1}$$

that being this translated in terms of a probability distribution function (p.d.f.), we have in a first instance:

$$D(x,N) = \sum_{n=1}^{N} \frac{U(n,x) - V(n,x)}{W(n,x)} \tag{2}$$

where $E_c(x)M(n) \to U(n,x)$ whereas $E_b(x)M(n) \to W(n,x)$ are hard approximations from the discrete space to one continuous. With these mathematical assumptions, it is possible to express the p.d.f. We underlined the fact that all three functions $U(n,x), V(n,x)$ and $W(n,x)$ might have a fully different morphology being all of them fully independent among them. Certainly, the different manners as to be expressed in (2) would depend on the $E_c(n)$ and $E_b(n)$. Despite of the fact that these functions becomes continuous along the number of customers, it is not difficult to see that the zeros of $W(n,x)$ the total recovery of the Banks for a large amount of customers appears to be large.

## 4 SUCCESS AND FAILS ON THE TRANSACTIONS

We can use (2) in order to estimate the realistic scenarios where a Bank can recover the loans as well as the ones of pure lost. In Fig.1 is shown the recoverability as function of the average (black curve) trial parameter that is ranging

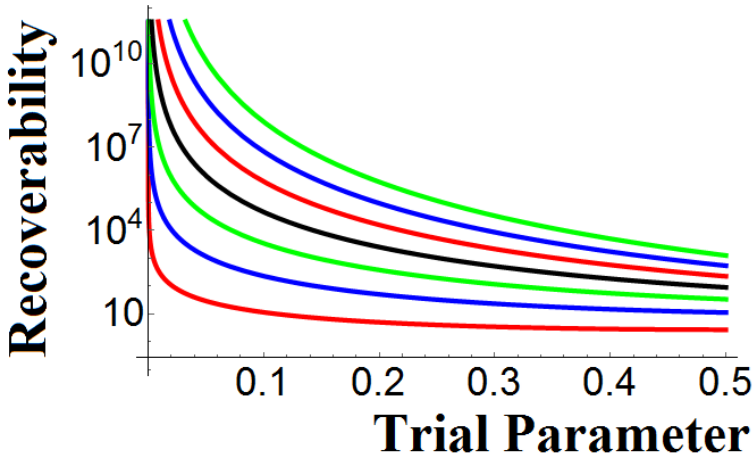



**Fig. 1.** The recoverability as function of trial parameter.

between 0 and 1. The colors green, blue and yellow down the black curve denote the cases where customers are presenting delays in their obligations. The ones above the black curve are the representative cases where the recoverability falls down with a sustainable growth of the recoverability as seen in such curves is still affordable for Banks. However, for the middle value of the trial parameter, this recoverability might be negative with a hardly capability of 1

**On The Concept of Recoverability:**

While Banks can cross the thin line that separates the legal and ordinary methods in that situations where is seen a successive occurrence of events of procrastination, customers while have not carry out a plan of payments, the recoverability would be weak in time and the models of profitability expected by Banks from cash credits becomes volatile. By assuming a scenario where customers might be well educated as to get a full view on the realistic disadvantages after of getting a cash credit, the notion of recoverability is enclosed in an arena that is limited by economic and social methods more than legal issues through the usage of local laws that in most cases turn out to be favorable to Banks, however the costs of the usage of these ways would cause a notable divergence of the mission and vision of Banks.

As mentioned above, the scenario of lost is fully expected in a random cash credit market where customer have not acquired an acceptable list of instructions as to manage efficiently the schedule of payments. In Fig.2 CR has been moved to 0.5 originating a discontinuity of the recovery cash volumes. We can see that for those values less than 0.5 the recovery volumes have turned out to be minimal to reach the value of 10 for the first 10 months (curves of color green, blue, yellow) just those below the black arrow, might tell us that the dynamics between Bank and customers is rather weak by exhibiting. Only in the values near to 0.5 that is translated as the one that makes a hard separation between the scenarios of

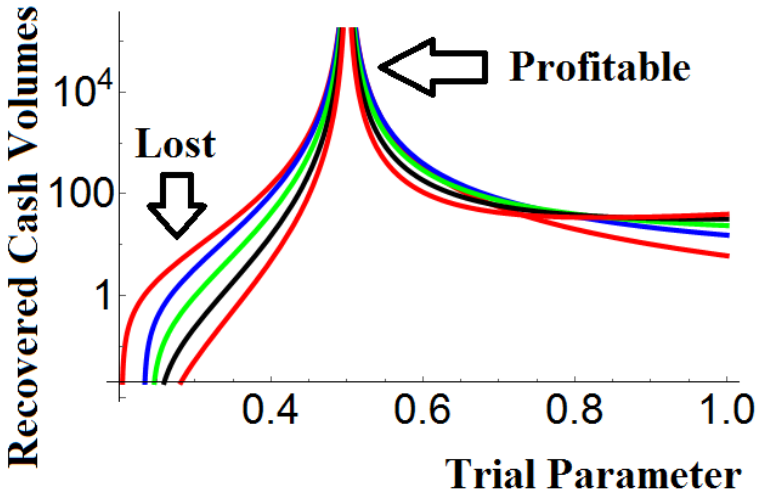



**Fig. 2.** The recovered volumes of cash as function of trial parameter.

being an action profitable or lost respect to the cash credits one can see that for those values beyond that 0.55, all possible scenarios are understood as being dominated by one morphology that is stable allowing us to conclude preliminarily that the profitability can be self-protected as an inherent mechanism to credit dynamics and ethics more than the apparition of pressure by the Bank and the reconfiguration of the schedules of payment.

Finally in Fig.3, is shown the apparent symmetry and asymmetry of the curves of Failed Customers as function of the trial parameter. Concretely, we have paid implemented a function of the number of failed customers given by:

$$N(x,N) = \frac{\mathrm{Sin}^{0.5+n}(nx) - 0.65\mathrm{Exp}(x^2 - 2n)}{(x-0.5)^{2n+1}} \tag{3}$$

In this case from bottom to top, the colors green, black, and yellow are displaying a kind of asymmetry in the right side of the plots. It is due to the fact that the trial parameter is exhibiting an oscillatory character as seen in (3) because the sinusoid Sin function that is interpreted as a oscillating behavior of customers against to the schedule of payments.

## 5 CONCLUSION

In this paper we have presented an analysis based on algorithms that are built in basis of the dynamics of the interaction Bank and customer after a cash credit is released. Our central focus has been the fact that customers do not proceed to make a wide assessment about the advantages and disadvantages of

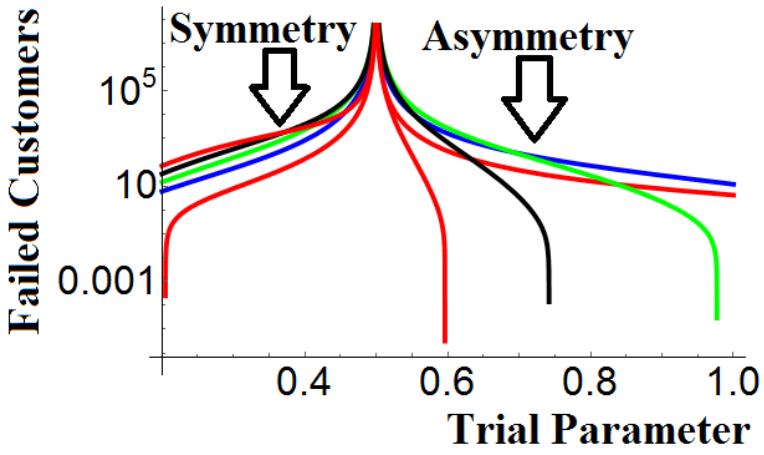



**Fig. 3.** The number of failed customers as function of trial parameter.

the real cost of getting a cash loan or credit by a Bank, that have performed a study concerning the pro and cons the cash credit, fact that would have to take more time than the one taken by the customer [12]. Thus, the results of the simulations have demonstrated that customers without a solid strategy to face a cash loan, might be potentially subject to procrastination thereby deteriorating their relationship with credit institutions

## References

1. A. P. Zaitsev, The State Bank and the Enterprise: Is the Influence of Bank Credit Effective, Problems in Economics, Volume 26, 1983 - Issue 5, Published Online: 08 Dec 2014.
2. Gerhard Kling, A theory of operational cash holding, endogenous financial constraints, and credit rationing, The European Journal of Finance, Volume 24, 2018 - Issue 1 Published Online: 03 Sep 2016.
3. Jos Amrico Pereira Antunes, Claudio Oliveira De Moraes and Adriano Rodrigues, How financial intermediation impacts on financial stability?, Applied Economics Letters, Volume 25, 2018 - Issue 16, Published Online: 15 Nov 2017.
4. Andrea Moro, Daniela Maresch and Annalisa Ferrando, Creditor protection, judicial enforcement and credit access, The European Journal of Finance, Volume 24, 2018 - Issue 3 Published Online: 09 Sep 2016.
5. Naoki Wakamori and Angelika Welte, Why Do Shoppers Use Cash? Evidence from Shopping Diary Data, Journal of Money, Credit and Banking, Volume 49, Issue 1.
6. R. Matthew Darst, Ehraz Refayet, Credit Default Swaps in General Equilibrium: Endogenous Default and Credit Spread Spillovers, , Volume 50, Issue 8.
7. Leonardo Becchetti, Maria Melody Garcia, Giovanni Trovato, Credit Rationing and Credit View: Empirical Evidence from an Ethical Bank in Italy, Journal of Money, Credit and Banking Volume 43, Issue 6.

8. Luigi Guiso, Raoul Minetti, The Structure of Multiple Credit Relationships: Evidence from U.S. Firms, Journal of Money, Credit and BankingVolume 42, Issue 6.
9. Junghwan Hyun, Raoul Minetti, Credit Reallocation, Deleveraging, and Financial Crises, Journal of Money, Credit and Banking.
10. Sumit Agarwal, Souphala Chomsisengphet, Chunlin Liu, The Importance of Adverse Selection in the Credit Card Market: Evidence from Randomized Trials of Credit Card Solicitations, Journal of Money, Credit and BankingVolume 42, Issue 4.
11. Paula Lopes, Credit Card Debt and Default over the Life Cycle, Journal of Money, Credit and BankingVolume 40, Issue 4.
12. Pedro Gomis Porqueras, Daniel Sanches, Optimal Monetary Policy in a Model of Money and Credit, Journal of Money, Credit and BankingVolume 45, Issue 4.

# Theory of Blockchain Based on Quantum Mechanics

Huber Nieto-Chaupis

Universidad Autónoma del Perú
Programa de Ingeniería de Sistemas
hubernietochaupis@gmail.com

**Abstract.** Because the cryptocurrency dynamics inside the framework of Bitcoins would require of advanced schemes to encrypt information, the usage of powerful technologies aimed to guarantee anonymous information becomes a must. In this paper we present a theory of blockchain entirely based on Quantum Mechanics formalism the which might be used as part of advanced algorithms to generate random ensembles that is in fully concordance with the purpose of the markets based on cryptocurrency. Simulations have demonstrated the reliability of a simple model of blockchain of up to a 75% fact that supports the idea for using the Quantum Mechanics formalism in these advanced and highly secured modern markets.

## 1 Introduction

With the apparition of emergent markets based on the so-called Bitcoins [1][2][3], the software associated to these novel technologies aims to be more robust against attacks or any type of sophisticated spying. One of the critic issues related to the Criptocurrency becomes the strength of the technology to guarantee a high level of encryption so that both buyers and sellers are carrying out sure transactions with a null risk [4]. One of the notable encryption procedures is called the BB84 or the Bennet-Brassard algorithm [5] that is aimed to protect transferred information between two parties. Certainly, BB84 requires of strong algorithms to produce random sequences of bits through an orthogonal basis inside the N-dimensional Hilbert space. For example, a random basis:

$$|\Psi >= \sum_{n}^{N} \mathcal{R}_n |\phi_n > \tag{1}$$

where $\mathcal{R}_n$ is a complex number and it can be perceived as a kind of random number from the fact that the scalar product $< \phi_n|\Psi >$ since that the projection of the $|\Psi >$ onto the basis $|\phi_n >$ is done in a random manner, with $|\Psi >$ a quantum state whose time evolution depends on the so-called evolution operator:

$$\hat{\mathrm{U}}(\mathbf{t}, \mathbf{t_0})|\mathbf{\Psi}(\mathbf{t_0}) >= \mathrm{Exp}\left[-\mathbf{i}\frac{\mathcal{H}(\mathbf{t}-\mathbf{t_0})}{\hbar}\right]|\mathbf{\Psi}(\mathbf{t_0}) > \tag{2}$$

and $\hat{\mathcal{H}}$ the hamiltonian of the system. A large list of examples can illustrate the usage and practical applications of the evolution operator. In this paper, we use the formalism and methodologies of the quantum mechanics to propose robust algorithms to be employed in Cryptocurrency dynamics. While Bitcoins transaction needs of untouchtable softwares that provides high fidelity of security in e-commerce and e-payment, the reliability that a random number generator might give to users is considered as a must that in all cases would have to be implemented in servers.

## 2 The Quantum Mechanics Formalism

In Quantum Mechanics [6][7], the estimation of any physical observable demands to derive the amplitude of probability by which it would enable to measure physical quantities. In most cases, the usage of the mathematical machinery would imply to use the basic elements commonly called the formalism of Hilbert or bra and kets state vectors.

### 2.1 The Bra and Ket Formalism

Consider any blockchain process where the encryption of a concrete event requires aleatory numbers by which all of them should be of a large period. In this manner, the process of generation of a "pure" random number must have a beginning (previous the payment) and end (after the bitcoin payment was done). Thus, the "initial state" can be written as an infinite sum of kets

$$|I\rangle = \sum_{\ell=0}^{L} |\ell\rangle \, . \tag{3}$$

where $|0\rangle$ would denote the "ground state" of system. Indeed the existence of the unitary operator as derived of the completeness relation using orthogonal basis,

$$\mathbb{I} = \sum_{j=0}^{L} |j\rangle \langle j| \tag{4}$$

using orthogonal basis that satisfies $\langle i|j\rangle = \delta_{i,j}$. When $\mathbb{I}$ is projected onto the initial state the we have:

$$\mathbb{I}\,|I\rangle = \sum_{j=0}^{J} |j\rangle \langle j| \sum_{\ell=0}^{L} |\ell\rangle \, , \tag{5}$$

so that one gets the initial state "previous" to the bitcoin operation as follows:

$$|I\rangle = \sum_{j=0,\ell=0}^{J,L} |j\rangle \langle j|\ell\rangle \, . \tag{6}$$

Now we turn on the phase "during" the which is understood as the incorporation of a different orthogonal basis:

$$\mathbb{I} = \sum_{q=0}^{Q} |q\rangle \langle q| . \tag{7}$$

In this way, the end-to-end process can be written as

$$\langle F|I\rangle = \sum_{j,q,\ell}^{J,Q,L} \langle F|q\rangle \langle q|j\rangle \langle j|\ell\rangle . \tag{8}$$

It is noteworthy to note that actually we have the initial state $\langle F|$ splitted in an infinite number of sub-states $|\ell\rangle$ initial states with a only one final state $\langle F|$. Clearly there is advantage in extract a random number from the continue spectra than the discrete case. Under this view, we write down the completeness for the continuous case

$$\mathbb{I} = \int |x\rangle \langle x| \, dx \tag{9}$$

that is inserted in (8) between the first braket in the integration:

$$\langle F|I\rangle = \sum_{j,q,\ell}^{J,Q,L} \langle F| \left[ \int |x\rangle \langle x| \, dx \right] |q\rangle \langle q|j\rangle \langle j|\ell\rangle . \tag{10}$$

that finally we can rewrite it as the one given below:

$$\langle F|I\rangle = \sum_{j,q,\ell}^{J,Q,L} \int \langle F|x\rangle \langle x|q\rangle \langle q|j\rangle \langle j|\ell\rangle \, dx \tag{11}$$

with $\langle F|x\rangle = F(x)$ and $\langle x|q\rangle = \phi_q(x)$ then

$$\langle F|I\rangle = \sum_{j,q,\ell}^{J,Q,L} \int F(x)\phi_q(x) \langle q|j\rangle \langle j|\ell\rangle \, dx. \tag{12}$$

It should be noted that the inclusion of the basis $|j\rangle \langle j|$ is inhibited from the integration inserting the summation over "$j$",

$$\langle F|I\rangle = \sum_{j,q,\ell}^{Q,L} \int F(x)\phi_q(x) \langle q| \left[ \sum_{j}^{J} |j\rangle \langle j| \right] |\ell\rangle \, dx. \tag{13}$$

In this manner the sum over $j$ is operated by knowing that $\sum_j^J |j\rangle \langle j| = 1$, so that one gets

$$\langle F|I\rangle = \sum_{q,\ell}^{Q,L} \int F(x)\phi_q(x) \langle q|\ell\rangle \, dx. \tag{14}$$

The last step is justified in terms of the validity of the states transition $\langle F|I\rangle$. In fact, while $J \to \infty$ the completeness applies. It also means that one perceives the "during" as all possibles paths to carry out a **end-to-end bitcoin transaction**. Now we can close the transaction by making an extra operation that again demands to insert the continuous completeness $\mathbb{I} = \int |x'\rangle \langle x'| \, dx'$. It implies that

$$\langle F|I\rangle = \sum_{q,\ell}^{Q,L} \int F(x)\phi_q(x) \langle q| \int |x'\rangle \langle x'|\, \ell\rangle \, dxdx'. \tag{15}$$

Inside the coordinate space, we can again define the following "vectorial" states:

$$\langle q|x'\rangle = \phi_q(x), \tag{16}$$

$$\langle \ell|x'\rangle = \phi_\ell(x'), \tag{17}$$

The full transaction then opts the following form

$$\langle F|I\rangle = \sum_{q,\ell}^{Q,L} \int \int F(x)\phi_\ell(x)\phi_q(x')\phi_\ell(x')dxdx'. \tag{18}$$

Integration of the product of orthogonal polynomials have been studied by Glasser and Montaldi [8]. We can rewrite previous equation as

$$\langle F|I\rangle = \sum_{q,\ell}^{Q,L} \int (x)\phi_\ell(x)dx \int_0^{bx} \phi_\ell(x')\phi_q(x')dx'. \tag{19}$$

To note that for the last integration we opt to include the quantity $bx$ as upper limit. To some extent this action might to restrict the property of integration of the orthogonal polynomials by which is known that in most cases their existence overtakes the whole range of integration. Inspired on the Pierre-Humbert-Bessel integrals [9], in a first instance we write down that,

$$\lim_{Q\to\infty} \sum_{q}^{Q} \int_0^{bx} \phi_\ell(x')\phi_q(x')dx' = J_\ell(bx) \tag{20}$$

with $J_\ell(bx)$ the integer-order Bessel function. Thus, we can test the functionality of the full transition $\langle F|I\rangle$ as

$$\langle F|I\rangle = \sum_{q,\ell}^{Q,L} \int \; F(x)\phi_\ell(x)J_\ell(bx)dx. \tag{21}$$

We interpret Eq.(21) as the one that is a infinite chain of blocks depending the sums over $q$ and $\ell$. Therefore, the full transaction acquires the form of a matrix element

$$S_{q,\ell} = \sum_{q,\ell}^{Q,L} \int \; F(x)\phi_\ell(x)J_\ell(bx)dx. \tag{22}$$

The initial state can be parametrized in the sense that it acquires dependence of one parameter of full control by the client in the quality of any transaction. So that, one gets

$$S_{q,\ell} = \sum_{\ell}^{L} \int F(ax)\phi_\ell(x)J_\ell(bx)dx. \tag{23}$$

It is in according to [10] where a similar expression was derived involving the Bessel functions. It should be noted that Eq.(23) is essentially a kind of convolution integration. In terms of path integrals as commonly used in Quantum Mechanics, Eq.(23) denotes the integration of a single initial state $F(ax)$ through $\ell$ propagators or Green's functions $J_\ell(bx)$ that ends on also $\ell$ final states $\phi_\ell(x)$.

## 3 Eliminating Intermediate Processes

The main purpose of a blockchain is the coherent elimination of intermediate processes in a single transaction. As an illustration, consider the following example:

- Bob decides to send a quantity of money $M$ to Alice
- Bob uses the network to send an amount of bitcoins
- the request to send money arrives to a chain of blocks to approve the sending
- During the approval, Bob request acquires a code that is randomly generated and has a very small probability of being eavesdropped for a third party.
- The system checks if this probability keeps very small otherwise the transaction is cancelled and the blocks are set reset.

We turn now to the usage of Eq.(23). In order to numerically evaluate it, we assign the following functions: $\phi_\ell(x) = x^\ell$ as well as $\mathbf{F}(ax) = \mathbf{Exp}(ax)$. Putting everything together, we arrive to the formulation of a Quantum-Mechanics-based blockchain

$$S_\ell = \sum_{\ell}^{L} \int_0^{D_\ell} \mathcal{N}_\ell \mathbf{Exp}(-ax)J_\ell(bx)x^\ell dx. \tag{24}$$

The adjustment of the upper limit set to $D_\ell$ regulates the contribution per each order. Here we assume that $D_N, D_{N-1}, D_{N-2}.....D_{N-M}, D_{D-M-1}$ by which the employed algorithm uses in according to its numerical convenience. Indeed $\mathcal{N}_\ell$ a normalization constant. Working out in terms of probabilities, Eq.(24) should guarantee that the resulting integration acquires values between 0 and 1. We can also to calculate the efficiency of the blockchain operation with respect to first action $S_0$:

$$\mathcal{E} = \frac{S_0}{S_1 + S_2 + ...S_L}.$$

and in terms of the full mathematical machinery as derived in Eq.(24) one gets for $\ell = 0, 1$ and 2:

$$\frac{\int_0^{D_0} \mathcal{N}_0 \text{Exp}(-ax) J_0(bx) dx}{\int_0^{D_1} \mathcal{N}_1 \frac{J_1(bx)}{\text{Exp}(ax)} x dx + \int_0^{D_2} \mathcal{N}_2 \frac{J_2(bx)}{\text{Exp}(ax)} x^2 dx}. \tag{25}$$

Eq.(25) actually measures the efficiency of the encryption and blockchain sequences. Althouogh it is parametrized, the dependence on the "user" parameters: $a$, $b$, and $D_S$ provides a certain advantage to control of the bitcoin transaction.

### 3.1 Full Blockchain Process

The processes that would involve bitcoin as unit of currency must complete a chain of sequences that is mandatory to accomplish a task either for sell or buy any product that is in the e-Market. The main purpose of a bitcoin process lies in the anonymity of in the sender of bitcoin as well as the amount of cash that has been sent. Therefore, how to hidden ciphers and amount of volumes during the bitcoin transaction constitutes a critic point in the dynamics of blockchain. In order to test the effectiviness of Eq.(25) we must clarify the following items listed below:

- Choice of degree of chain $L$
- Preparation of Bessel function
- Assignment of values to $a$ and $b$
- Solving of normalization constants
- Perform the integration
- Compute sequence of random numbers

Once all these items have been established, the next step is the building of the full algorithm that mainly target to: (i) protect cliens under a bitcoin transaction, (ii) avoid the leak of information namely numbers that take part of the transaction. Below is written the blockchain algorithm that employs the Hilbert basis that entirely based on the Bessel functions as seen in Eq.(25). In essence we have imposed up to restrictions (i) when the efficiency is less than 0.5 then a new parametrization is done and no any completion to the bitcoin payment or transfer is cancelled, and (ii) when $\langle F|I\rangle > 0.90$ the transaction is fully approved so that the customer has access to the desired product. In the language of Bob and Alice, one can finally state that Alice has received the transfer of Bob without any intermediate event that poses in risk in their identities and total amount of transferred bitcoins.

```
1 INITIALIZES PARAMETERS a, b
  2 DO L = 1, MAX
  3 RANDOM NORMALIZATION CONSTANT
  4 PERFORMS INTEGRATION
  5 IF (L.EQ.MAX) THEN
  6 S(L) IS SAVES
```

**7** S(L-1) = $S_{L-1}$
**8** S(L) = $S_L$
**9** S(L-1) = $S_{L-1}$
**10** S(L+1) = $S_{L+1}$
**11** CALCULATES EFFICIENCY
**12** E(L+1) $\frac{S(L-1)}{S(L)+S(L+1)}$
**13 IF** E(L+1)<0.5 **THEN**
**14** RE-INITIALIZES AGAIN PARAMETERS a, b
**15** NEW RANDOM NORMALIZATION CONSTANT
**16 DO** L = 1, NEW-MAX
**17** REPEAT THE BLOCKCHAIN AGAIN
**18** NEW HILBERT BASIS
**19 IF** $\langle F|I\rangle > 0.90$ **THEN**
**20** TRANSACTION IS ACCOMPLISHED
**21** SELF-DELETE OF BLOCKCHAIN HISTORY
**22 ENDIF**
**23 ENDDO**
**24 ENDIF**
**25 ENDIF**
**26 ENDDO**
**27 END**

## 4 Simulations

### 4.1 The Bessel Functions as Propagators of Blockchain

As seen in Eq.(24) the Bessel fucntion from the point of view of Quantum Mechanics would play the rol as the propagator. Inside the context of blockchain and bitcoin dynamics, we relate it with the capacity of the system to manage the nodes or degree of interconnection under a blockchain scenario. In this way the well-known Bessel equation $x^2\frac{d^2J_\ell(x)}{dx^2}+\frac{dJ_\ell(x)}{dx}+(x^2-\ell^2)=0$ thus the Green's function $\left[x^2\frac{d^2J_\ell(x)}{dx^2}+\frac{dJ_\ell(x)}{dx}+(x^2-\ell^2)\right]G(x,x')=\delta(x-x')$ might be seen as the law that governs the chains during the bitcoin transaction. This has implications in our approach: the order of the Bessel function may be perceived as the degree of chain in a bitcoin transaction.
In Fig.1 top. middle and bottom, are plotted the computational simulations corresponding to Eq.(24). With the assumption that the order of the Bessel fucntion denotes the degree of chain, in top panel is displayed the amplitude $\langle F|I\rangle$ veresus the parameter $b$ that is part orf the argument of the Bessel function. The color blue, orange and magenta denotes the degrees 0, 1 and 2. In the first case, the curves were normalized to 10. Como seen along the range of $a$, the amplitudes falls down with $a$. The orange color $\ell=1$ has same behavior, while $\ell=2$ color magenta has stable behavior along the value of $\langle F|I\rangle$. In the middle panel we can see the behavior of the amplitude that is normalized to 1. Only the green color attached to $\ell=0$ has the highst values. We can see that this degree "the ground

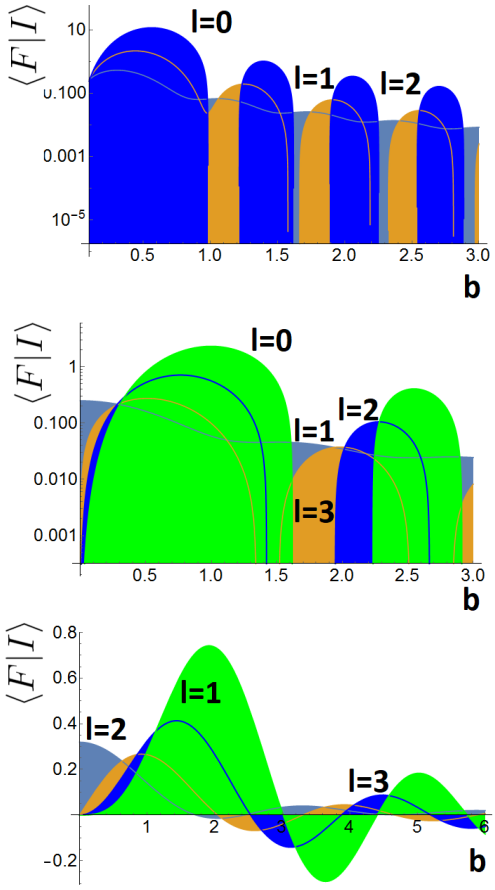



**Fig. 1.** The amplitude as function of the parameter $b$. The color has been assigned to the order of the Bessel function. Top, middle and bottom have used the Eq.24.

level" allows values close to 1, contrarily to the colors blue, orange and magenta with amplitudes below 0.5. Clearly in this case, the transaction might have been cancelled due to lack of any efficient mechanism that protect the transaction. In bottom panel, we used the approximation: $e^{-ax} \approx 1 - ax + (\frac{ax}{2!})^2$. The effect of this in the estimation of the amplitude $\langle F|I\rangle$ is reflected on the resulting curves that display negatives values that turns out to be unphysical. Certainly a negative amplitude is analogue to a negative probability so the blockchain algorithm might collapse in dead periods where no any customer or service is active. As seen in line-19 of algorithm, the impossibility to accomplish gives as result the truncation of the dynamics of a blockchain and subsequently the cancellation of the transaction, therefore bitcoins owners still have the chance of return to a common currency. Finally the morphology of bottom panel in Fig.1 appears

to be remarkable in the sense that the amplitude reach up to a 75% fact that indicates that even in the case of one only chain the guarantee of the secrecy of the transaction is still conserved.

## 5 Conclusion

In this paper we have presented a formalism based entirely in the Quantum Mechanics mathematics that have allowed to simulate a dynamics of blockchain for a e-commerce transaction. Clearly, the full transaction has been described as a convolution. Finally, a blockchain of 75% was identified on the curves of amplitude supporting the fact that the theory of blockchain is one of the most robust to perform e-commerce using the bitcoins. In a future work, we exploit the features of the Quantum Mechanics to model the full dynamics of flux of bitcoins in large economies.

## References

1. Morgen E. Peck, The Bitcoin Arms Race is on! Morgen E. Peck IEEE Spectrum Year: 2013 Volume: 50 , Issue: 6 Pages: 11 - 13.
2. Christine Evans-Pughe ; Alexei Novikov ; Vitali Vitaliev, To bit or not to bit? Engineering and Technology Year: 2014 Volume: 9 , Issue: 4 Pages: 82 - 85.
3. Florian Tschorsch ; Bjrn Scheuermann, Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies IEEE Communications Surveys and Tutorials Year: 2016 Volume: 18 , Issue: 3 Pages: 2084 - 2123.
4. Jong-Hyouk Lee ; Marc Pilkington, How the Blockchain Revolution Will Reshape the Consumer Electronics Industry [Future Directions] IEEE Consumer Electronics Magazine Year: 2017 Volume: 6 , Issue: 3 Pages: 19 - 23.
5. Hui-Fang Li ; Li-Xin Zhu ; Kai Wang ; Kai-Bin Wang, The Improvement of QKD Scheme Based on BB84 Protocol, 2016 International Conference on Information System and Artificial Intelligence (ISAI) Year: 2016, Pages: 314 - 317.
6. J. Sakurai, Advanced Quantum Mechanics, Addison Wesley; Revised edition (September 10, 1993).
7. Claude Cohen-Tannoudji, Quantum Mechanics, Vol. 1 1st Edition Wiley; 1st edition (January 8, 1991).
8. M. Lawrence Glasser, Emilio Montaldi, Some integrals involving Bessel functions, arXiv:math/9307213.
9. Pierre-Humbert, Proceedings of the Edinburgh Mathematical Society, Volume 3, Issue 4 August 1933 , pp. 276-285.
10. H. Nieto-Chaupis, Encrypted Communications Through Quantum Key Distributions Algorithms and Bessel Functions, 2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Year: 2018, Pages: 1 - 4.

# Different frequencies in term structure forecasting

Alexander B. Matthies*

Institute for Quantitative Business and Economics Research
Christian-Albrechts-Universität
Heinrich-Hecht-Platz 9, 24188 Kiel, Germany
a.matthies@qber.uni-kiel.de
http://www.qber.uni-kiel.de/de

**Abstract** Different frequencies of German, US, and UK government term structure data are used in a dynamic factor model forecasting exercise. Employing a purely data-driven approach model selection for the purpose of forecasting is compared for different data frequencies and forecast horizons. In a principal components analysis of all term structures we find evidence of a stable global level factor. With regard to forecasting lower frequencies can produce better forecast statistics. The selection of the number of factors is stable over data frequencies. But it may vary over the yield curve for the US.

**Keywords:** Principle components; dynamic factor method comparison; government bond interest rates; term structure; directional accuracy; directional forecast value; different data frequencies; global yield.
JEL classification: C32; C53; E43

## 1 Introduction

Modeling and forecasting the term structure of interest rates is an important aspect of risk modelling The yield curve is an important indicator of future economic development. An inverted yield curve is often taken as an indicator of a future recession. Government bond interest rates determine the ability of a country to finance itself and are important factors in the fixture of other interest rates. Modelling the term structure helps in understanding this economic base variable and determine financial risk.

Using the concept of an approximate factor model by Chamberlain & Rothschild (1983) Stock & Watson (2002) develop an estimation and forecasting method based on principal components analysis (PCA) that can be applied to dynamic factor models. In a dynamic factor model it is assumed that the unobservable static factors are driven by some dynamic factors that are equally not directly observable. Factor models for term structures usually assume three

* QBER - Institute for Quantitative Business and Economics Research, Christian-Albrechts-Universität zu Kiel

factors (Nelson & Siegel, 1987). These are the level, slope, and curvature factors. Factors can be either be estimated with predetermined loadings (Diebold & Li, 2006, Yu et al., 2009, Yu & Zivot, 2011) or data driven methods (Litterman & Scheinkman, 1991, Blaskowitz & Herwartz, 2009, 2011, Matthies, 2014, 2018). For global term structure data factor analysis suggests the presence of a global level factor (Diebold et al., 2008, Abbritti et al., 2018, Matthies, 2018). Diebold & Li (2006) test Nelson and Siegel model for forecasting. Blaskowitz & Herwartz (2011) use a data driven approach employing autoregressive factors. Matthies (2014) and Matthies (2018) expand on these results and tests for the use of financial data and additional term structures as well as different estimation and forecasting methods.

Term structure data can be used for different economic and business applications. Term structure forecasts are then needed at different frequencies (e.g. daily or monthly). Using data of different frequencies in economic modelling can be challenging (Armesto et al., 2010). If data is available at different frequencies it is is for example possible to produce daily, weekly, and monthly forecasts with daily data. While weekly and monthly data requires larger time windows for modelling. One can use simple one, two, three step forecasts for monthly frequencies with AR methods. While weekly data needs to compute 4-step ahead forecasts and daily data must compute 21-step forecasts to produce monthly forecasts.

Using a data driven factor model for term structures this study forecast government term structures with data of different frequencies. Thereby testing for the best frequencies to forecast the term structure at different forecast horizons and test the stability of forecast model selection. The data consists of daily, weekly, and monthly observations of government bond interest rates for Germany, the UK, and the US for the time period from 1980 to 2016. Unseen dynamic factors of the term structure are estimated in a data driven approach using principal components analysis in rolling time windows to produce yield curve forecasts. Statistical and economic evaluation is provided for different time windows, factor numbers, and lags. Furthermore, the three government term structures are used to test for the presence of a global level factor over time.

This paper continues as follows. Section 2 presents the factor model representation, factor model estimation via PCA, the forecasting method, and forecast evaluation. In Section 3 presents the data, results of the PCA, and forecast strategies. Section 4 discusses some select results. Section 5 concludes.

## 2 Methods

We now briefly discuss dynamic factor models in their static representation and estimation via PCA. Then we sketch the forecasting method using autoregressive factors. Finally, we present our loss functions for forecast evaluation.

### 2.1 Static Representation of Dynamic Factor Model and PCA estimation

We follow Stock & Watson (2002), Breitung & Eickmeier (2006), and Stock & Watson (2011) with regards to presentation and estimation of dynamic factor models. The underlying factors of the term structure are estimated via PCA. This method allows the factor loadings to be estimated via a data driven approach. This is in contrast to the estimation with predetermined loadings that can be then used to estimate factors via a Kalmann Filter or similar methods.

The use of factor models is motivated by representing a large number of $N$ correlated variables can be represented by a smaller number of $R$ factors. The dynamic factor model is represented in its static form as:

$$\widetilde{X}_t = F_t A_t' + E_t, \qquad t = \tau, \ldots, T-h. \tag{1}$$

Where $\widetilde{X}_t = (\tilde{\mathbf{x}}_1', \ldots, \tilde{\mathbf{x}}_\tau')'$ $(\tau \times N)$; $N$ appropriately modified interest rates. In this case $\widetilde{X}_t$ is the deviation from the time window conditional mean. $F_t = (\mathbf{f}_1', \ldots, \mathbf{f}_\tau')'$ $(\tau \times R)$ are the $R$ unobservable factors, with loadings $A_t$ $(N \times R)$ and errors $E_t$ $(\tau \times N)$. For $A_t$ the $n$'th row is $\mathbf{a}_{n\bullet}$ $(1 \times R)$ and the $r$'th column is $\mathbf{a}_{\bullet r}$ $(N \times 1)$.

Following Blaskowitz & Herwartz (2009) the factor model is estimated in a moving time window to account for possible temporal instability. Estimates of Eq. (1) are therefore taken from a rolling time window of size $\tau$ sequentially at each instance $s$. $X_t$ are the individual countries term structures. For the purpose of identifying a global level factor we also perform a rolling time window estimation with all three term structures.

Conditioned on a given time window of sample information the nonlinear objective function is

$$V_{OLS}(\tilde{F}_t, \tilde{A}_t) = \sum_{n=1}^{N} \sum_{s=t-(\tau-1)}^{t} (\tilde{x}_{n,s} - \tilde{\mathbf{f}}_s \tilde{\mathbf{a}}_{n\bullet}')^2. \tag{2}$$

The eigenvectors $\hat{\mathbf{a}}_{\bullet,r}$ corresponding to largest eigenvalues $\lambda_r$, $r = 1, \ldots, R$ of $(\tilde{X}_t' \tilde{X}_t)$ and the corresponding $\hat{F}_t = \tilde{X}_t \hat{A}_t$ are the minimising factors[1].

In Matthies (2014) and Matthies (2018) it is shown that term structure models will usually outperform simple AR forecast models. The selection of the number of factors used to model the term structure can be selected according to forecast performance.

### 2.2 Forecasting with autoregressive Factors

The $M$–dimensional vector $\mathbf{y}_{t+h} = (y_{1,t+h}, \ldots, y_{m,t+h}, \ldots, y_{M,t+h})'$ is to be forecasted. Having the deviation from the mean, $\tilde{y}_{m,s} = y_{m,s} - \bar{y}_{m,t}$, where $\bar{y}_{m,t} =$

[1] Estimation via this method is analogous to the estimation of the static model in Eq. (1).

$\frac{1}{\tau}\sum_{s=t-(\tau-1)}^{t} y_{m,s}$. and the information set $\Xi_{t,\tau} = \{\tilde{\mathbf{x}}_s | s = t-(\tau-1), \ldots, t\}$, so that we have $\hat{\mathbf{y}}_{t+h|t} = E[\tilde{\mathbf{y}}_{t+h}|\Xi_{t,\tau}] + \bar{\mathbf{y}}_t$.

Using factor models in forecasting exercises, it is intuitive to exploit the relationship between factors and variables given by the factor loadings. In the approach from Blaskowitz & Herwartz (2011), the factors $\check{\boldsymbol{f}}_{s+h|s}$ are first predicted, and then the factor model Eq. (1) is used to determine forecasts $\hat{y}_{m,s+h|s}$. The basic assumption is that factors follow a VAR process in first differences,

$$\Delta\hat{\mathbf{f}}_s = \nu + \Phi_1\Delta\hat{\mathbf{f}}_{s-1} + \cdots + \Phi_q\Delta\hat{\mathbf{f}}_{s-q} + \eta_s,$$

where $\Delta\hat{\mathbf{f}}_s := \hat{\mathbf{f}}_s - \hat{\mathbf{f}}_{s-1}$ and $\nu$ $(R \times 1)$; $\Phi_l$ $(R \times R)$ for $l = 1, \ldots, q$. The error vector is $\eta_t$ $(R \times 1)$. As factors are orthogonal we assume AR processes. The $\Phi_l$ are therefore assumed to be diagonal and for each factor we have a univariate autoregressive process:

$$\Delta\check{f}_{r,t+s|t} = \hat{\nu}^{(r)} + \hat{\phi}_1^{(r)}\Delta\check{f}_{r,t+s-1|t} + \ldots + \hat{\phi}_q^{(r)}\Delta\check{f}_{r,t+s-q|t}.$$

Here $\hat{\nu}^{(r)}$, $\hat{\phi}_1^{(r)}, \ldots, \hat{\phi}_q^{(r)}$ are OLS estimates in $\Xi_{t,\tau}$. Obtaining factor forecasts $\check{f}_{r,t+h|t} = \hat{f}_{r,t} + \sum_{s=1}^{h}\Delta\check{f}_{r,t+s|t}$ the $h$-step term structure forecast is then

$$\hat{\mathbf{y}}_{t+h|t} = E[\tilde{\mathbf{y}}_{t+h}|\Xi_{t,\tau}] + \bar{\mathbf{y}}_t = \hat{A}_t\check{\mathbf{f}}_{t+h|t} + \bar{\mathbf{y}}_t.$$

The term structure factors estimated via PCA are usually autocorrelated. This finding motivates the employed forecasting method. Yet, with regard to model selection for the purpose of forecasting the presence of autoregressive factors need not imply that lags must be implemented to improve forecasts. Indeed, in Matthies (2018) I found that models with additional lags are often outperformed by those without lags. Forecasting the factors first and then employing the factor model to produce forecasts of the entire term structure assumes that factor loadings are stable out-of-sample, or that they are at least sufficiently stable until the forecast horizon.

### 2.3 Forecast evaluation criteria

Statistical and economic forecast evaluation are related but distinct assessments of model specific features for the purpose of forecasting. Statistical criteria might not necessarily conform with economic measures (Diebold & Mariano, 2002). Here, evaluation of forecasts are done via one statistical and two economic criteria. Specifically, in addition to the standard mean squared forecast error MSFE criterion directional accuracy DA and big hit ability (BHA) are considered as in Blaskowitz & Herwartz (2011), Matthies (2014), and Matthies (2018). The (MSFE) might not be as correlated with profits as the DA of forecasts are. Overall performance of competing forecasting methods is evaluated via an analysis of variance (ANOVA). This estimates the average impact of any model specification.

The most intuitive criterion to evaluate forecasts is the MSFE:

$$MSFE^{h,m} = \frac{1}{T - (\bar{\tau} + \bar{h} - 1)} \sum_{t=\bar{\tau}+\bar{h}}^{T} \hat{\varepsilon}^2_{m,t+h|t}. \tag{3}$$

where $T_0$ and $\bar{h}$ are the respective time window and forecast horizon applied in this study. The forecast error is $\hat{\varepsilon}_{m,s+h|s} = \hat{y}_{m,s+h} - y_{m,s+h|s}$. Here, MSFE can be interpreted as squared percentage points. The strategy with the minimal MSFE value is the one that performs the best. Furthermore, we will expect MSFE values to increase at larger forecast horizons.

The measure of directional accuracy (DA) "loss" is determined via the function

$$\overline{DA}^{h,m} = \frac{1}{T - (\bar{\tau} + \bar{h} - 1)} \sum_{t=\bar{\tau}+\bar{h}}^{T} I\left[(\hat{y}_{m,t+h|t} - y_{m,t})(y_{m,t+h} - y_{m,t}) > 0\right], \tag{4}$$

where $I[\bullet]$ is an indicator function. Perfect directional accuracy would be indicated by a DA value of one. DA values are in fractions. Larger DA values mean better forecast performance. The DA values represent the ability to accurately predict the direction of a financial time series.

Hartzmark (1991) proposed the BHA measure. Correct prediction of large changes are weighed by their size. The BHA "loss" function is

$$\overline{BHA}^{h,m} = \frac{1}{T - (\bar{\tau} + \bar{h} - 1)} \sum_{t=\bar{\tau}+\bar{h}}^{T} BHA^{h,m}_{t+h|t}, \tag{5}$$

$$BHA^{h,m}_{t+h|t} = \begin{cases} |y_{m,t+h} - y_{m,t}| & \text{if } (\hat{y}_{m,t+h|t} - y_{m,t})(y_{m,t+h} - y_{m,t}) > 0 \\ -|y_{m,t+h} - y_{m,t}| & \text{if } (\hat{y}_{m,t+h|t} - y_{m,t})(y_{m,t+h} - y_{m,t}) < 0. \end{cases}$$

The BHA function is weighted version of the DA criterion. As with DA larger values mean better forecasting performance.

## 3 Data, Factors, and Forecasts

We now shortly sketch the data used in this study and discuss some of its' features. Then we present the results of the rolling PCA used to estimate factors for the forecasting exercises. In addition we discuss the results of a rolling PCA of all term structures. Lastly, we discuss the forecasting strategy design and the ANOVA evaluation concept.

### 3.1 Data

The time series of the yield to redemption of 3 government term structures at daily, weekly, and monthly frequency are employed in this study. We have

five German government bond interest rates (GER), six US treasury benchmark bonds (US), and eight UK government liability nominal spot rates (UK) (See Table 1. The data starts at the 01.01.1980 and goes until 06.09.2016. We therefore have 9571 daily, 1915 weekly, and 441 monthly observations.

| Mat | 1 | 2 | 3 | 5 | 7 | 10 | 15 | 20 | 30 | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|
| GER | | × | × | × | × | × | | | | 5 |
| US | | × | × | × | × | × | | × | | 6 |
| UK | | × | × | × | × | × | × | × | × | 8 |

Table 1: Maturities used for GER, SWI, the UK, and the US.

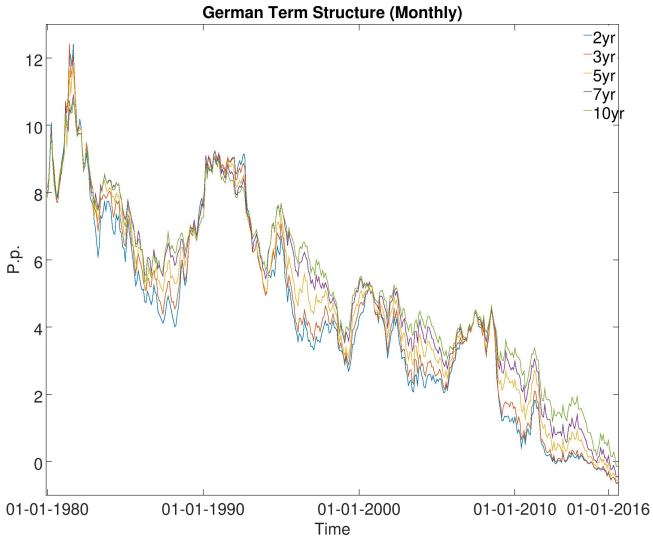



Figure 1: Term Structure of German government bond term structure monthly

In Figure 1 we display the evolution of the German term structure. One prominent feature is the downward trend of interest rates that is also present in the US and UK term structure. For Germany we also have the case that interest rates become negative for some periods towards the end of data set.

Given the maturities in our data set we follow Blaskowitz & Herwartz (2011) and calculate the level, slope, and curvature of the term structure. This is done as: $lev_s = (2yr_s + 5yr_s + 10yr_s)/3$ (level), $slo_s = (10yr_s - 2yr_s)/2$ (slope), and $cur_s = 2yr_s/4 - 5yr_s/2 + 10yr_s/4$ (curvature). Forecast performance for these

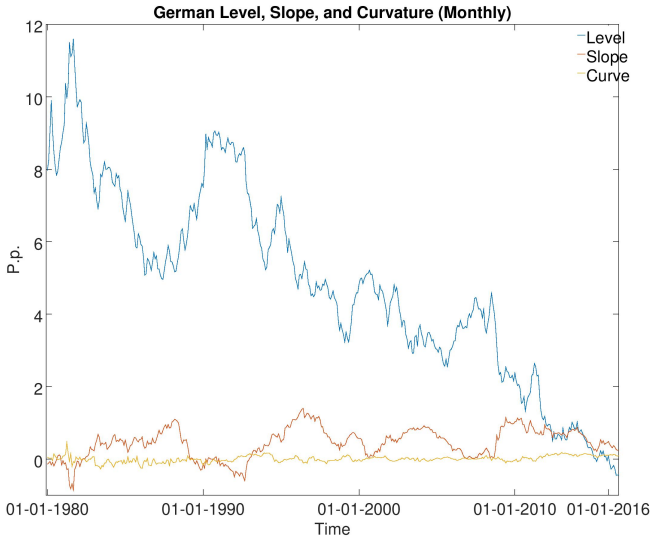


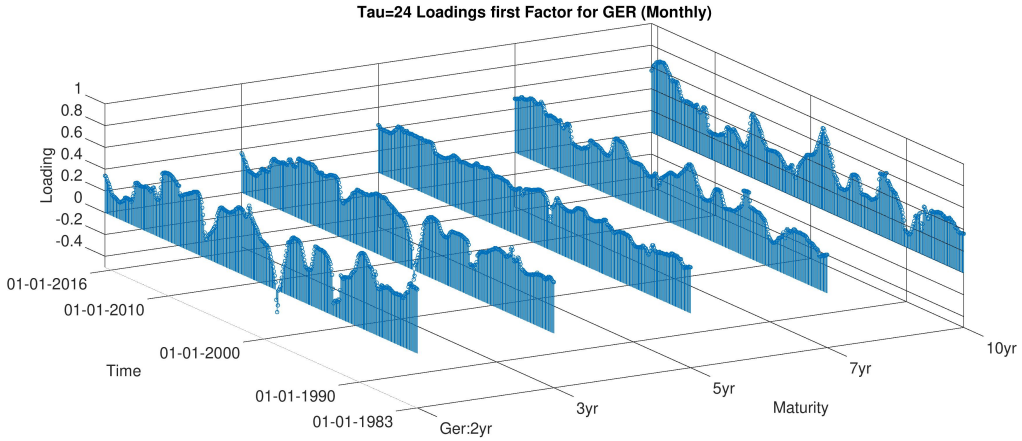

Figure 2: Level, Slope, and Curve German government bond term structure monthly

three linear combinations is also investigated. Figure 2 shows how level, slope, and curvature change in the German monthly data set. It is noteworthy that at certain points the slope value becomes negative. This feature is typically interpreted as an indicator for a looming recession (Estrella & Hardouvelis, 1991).

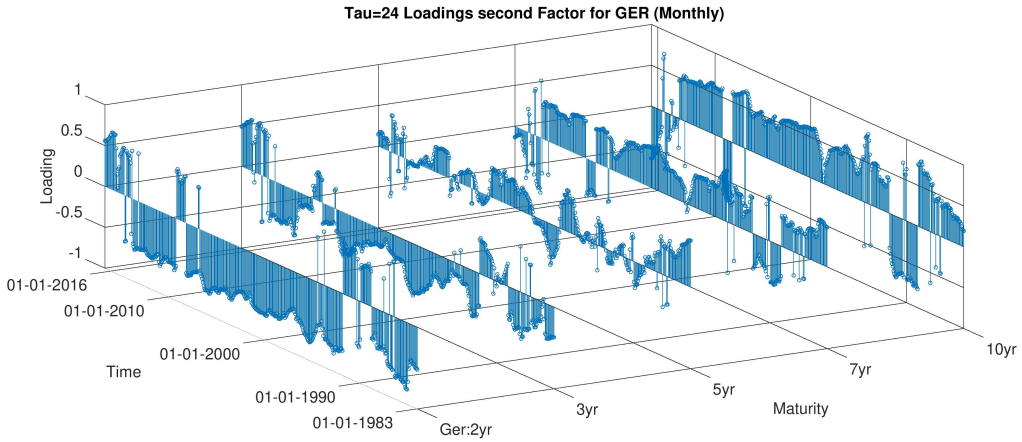
### 3.2 Factors

The first three factors of the term structure are of special interest. They are seen as to represent the essential features of the term structure. The intuitive approach of Blaskowitz & Herwartz (2011) to represent the loadings can be represented visually described as follows. A horizontal line for the level factor features no change in sign, i.e. all loadings should have the same sign. For the second factor we expect one change in sign to represent a slope factor. The third factor is expected to be 'U' or 'V' shaped. It requires two sign changes (e.g. negative – positive – negative) to represent the curvature factor. We now compare select results of the empirical PCA with this construct.

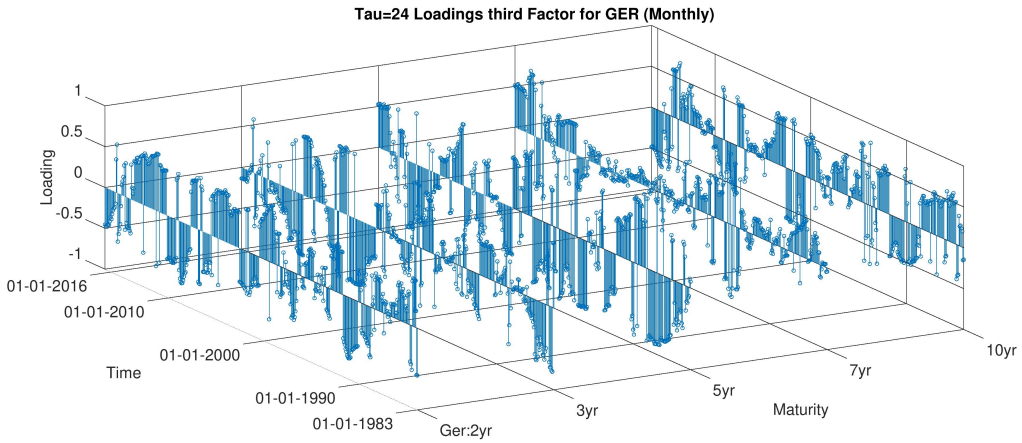
The three panels of Figure 3 depict the PCA from a rolling time window for monthly data for the German term structure. In Panel (a) the first factor is depicted. It represents 393 eigenvectors corresponding to the largest eigenvalues of the rolling PCA. All loadings are consistently positive (with a minor exception). In Panel (b) that negative loadings at the short term end of the yield curve correspond to positive loadings at the long term end, or vice versa. Thus




(a) GER 1




(b) GER 2




(c) GER 3

Figure 3: Loadings of the first, second, and third factor for GER with $\tau = 24$ (monthly).

the loadings have a single sign change along the yield curve. In Panel (c) we find that positive (negative) loadings for $2yr$ and $10yr$ maturities correspond with negative (positive) loadings for the $5yr$ maturity. This corresponds to two sign changes. These observations hold in general for all three term structures under all considered time frequencies and time window estimation sizes. The PCA estimation thereby supports the idea that the underlying factors of term structures correspond to a level, slope, and curvature factor. Furthermore, the factor loadings are locally stable which supports the assumption of our forecasting model.

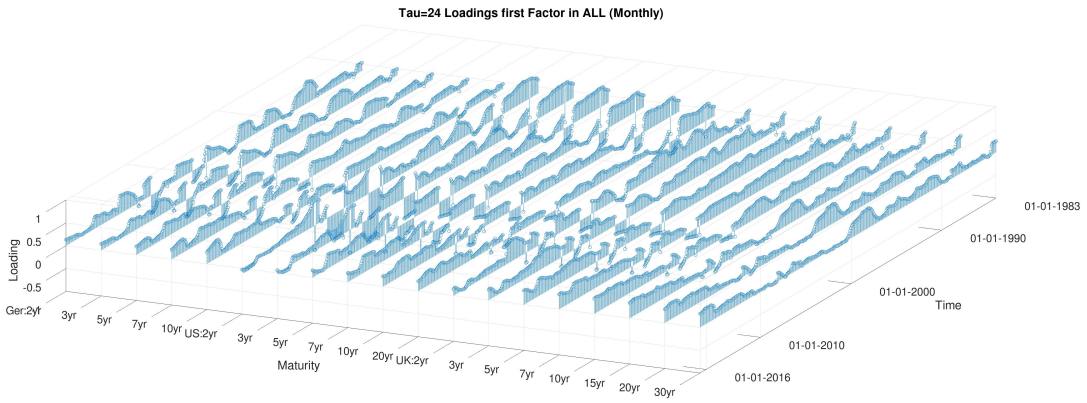
In Matthies (2018) an alternative data set was employed for the purpose of forecasting. This data set contained all term structures used for forecasting (the US, the UK, Germany, and Switzerland). Using daily data from 2000 to 2016 PCA indicated the presence of a global level factor. Later factors were more difficult to interpret as loadings were not stable. Abbritti et al. (2018) and Diebold et al. (2008) have provided interpretations for the second, third, and even the fourth factor extracted from global term structure data. As a side investigation we perform PCA from rolling time windows from a data set with all term structures. These results are important empirical contributions to the discussion of global term structure factors.

Statistically the factors in the global data set describe the correlation between and across term structures. Abbritti et al. (2018) interpret the first factor of their global PCA as a 'global expected inflation' factor based correlation analysis. The loadings can be interpreted as representing a global level factor. Based on further correlation analysis they interpret the second factor as a 'global expected growth' factor.

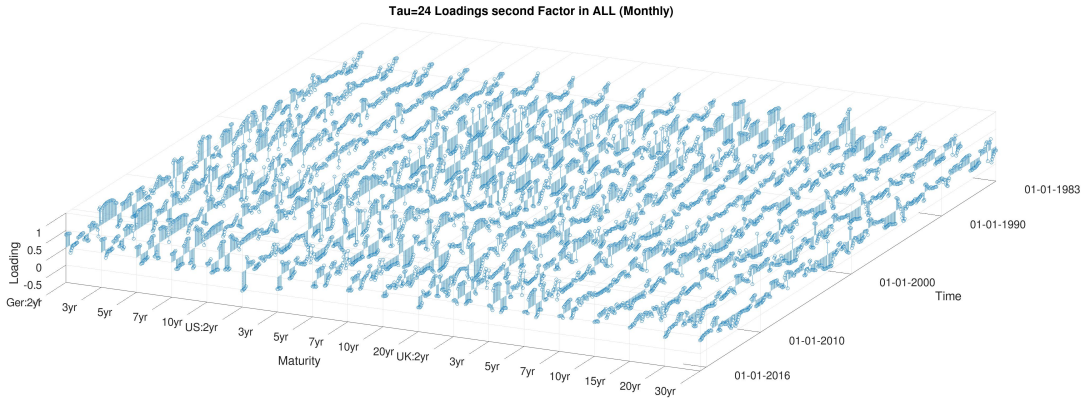
The eigenvectors corresponding to the three largest eigenvalues of our global data set from a rolling time window with monthly data and $\tau = 24$ are depicted in Figure 4. Consistent with the findings of Abbritti et al. (2018) and Matthies (2018) Panel (a) shows that loadings are positive over all three term structures. This holds over most of the sample with few exceptions. This finding supports the notion of a global level factor. In stark contrast loadings on the second and third factor depicted in Panel (b) and (c) respectively are far less stable than those of the first factor. The second factor appears to depict correlation between the levels of the three term structures not captured by the level factor. The third factor at certain instances depicts a global slope factor, i.e. a change in sign of the loadings over each term structure. But this pattern is not stable. Given the instability of the factor loadings the second and third factor arguably defy a constant interpretation when estimated from a rolling time window. Never the less, the loadings are locally stable. This would allow the use of global factors for the purpose of forecasting as in Matthies (2018).

### 3.3 DFM Strategies and ANOVA Analysis

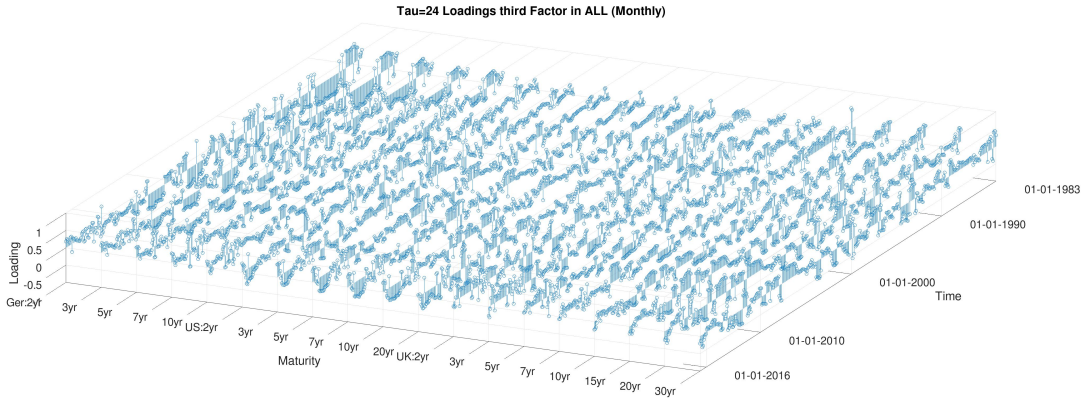
Forecasting strategies are defined by $\{\tau, r, q\}$. $\tau$ determines the size of the rolling time window from which factors are extracted. $r$ determines the number of




(a) ALL 1




(b) ALL 2




(c) ALL 3

Figure 4: Loadings of the first, second, and third factor for All with $\tau = 24$ (monthly).

factors employed. The autoregressive lag number for factors is given by $q$. In contrast to Blaskowitz & Herwartz (2011) and Matthies (2014), and Matthies (2018) this study expands the used frequencies from daily to daily, weekly, and monthly. Accordingly, alternative, forecast horizons, time window sizes and lag numbers differ. Forecast horizons are $h \in \{5, 10, 15, 21, 42, 63\}$, $h \in \{1, 2, 3, 4, 8, 12\}$, $h \in \{1, 2, 3\}$ for daily, weekly, and monthly data respectively. Time window sizes are $\tau \in \{125, 189, 250, 300\}$, $\tau \in \{27, 34, 50, 60\}$, and $\tau \in \{24, 48\}$ for daily, weekly, and monthly data respectively. The number of factors are $r \in \{1, 2, 3, 4\}$ for all frequencies. The alternative number of lags employed are $q \in \{0, 1, 2, 3, 5, 10, 15, 20\}$, $\in \{0, 1, 2, 3, 5\}$, and $\in \{0, 1, 2, 3\}$ for daily, weekly, and monthly data respectively. The number of strategies are for daily data $(4 \times 4 \times 8)$ 128, for weekly $(4 \times 4 \times 5)$ 80, and for monthly $(4 \times 4 \times 5)$ 32.

With an ANOVA we can determine the average impact of each alternative strategy characteristic. To perform ANOVA's we select benchmark strategies for daily, weekly, and monthly data. These are $\{125, 1, 0\}$, $\{27, 1, 0\}$, $\{24, 1, 0\}$ for daily, weekly, and monthly data respectively. Furthermore, we construct dummy variables for each alternative. For daily data we have 3 for the time window sizes, 3 for the number of factors, and 7 for the alternative lags. This adds up to 13 dummy variables. Accordingly, there are $(3 + 3 + 4)$ 10 dummy variables for the weekly data and $(1 + 3 + 3)$ 7 for the monthly data. In addition with a constant a simple least squares regressions then produce parameters that represent the impact of each strategy characteristic, under each evaluation criterion, for each country, at every forecast horizon, for each frequency, and for every maturity and linear combination[2].

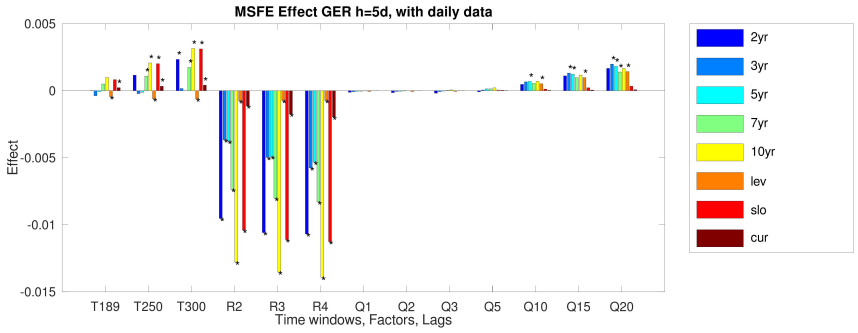
## 4 Results

The setup described above gives us grouped ANOVA results for 3 evaluation criteria for 3 government term structures at 3 or 6 forecast horizons. For daily and weekly data this adds up $(3 \times 3 \times 6)$ 54 and for monthly data to $(3 \times 3 \times 3)$ 27. Together these are 135 ANOVA results for term structure forecast evaluation. We have selected some results here that highlight some important aspects of data driven factor model forecasting of yield curves.

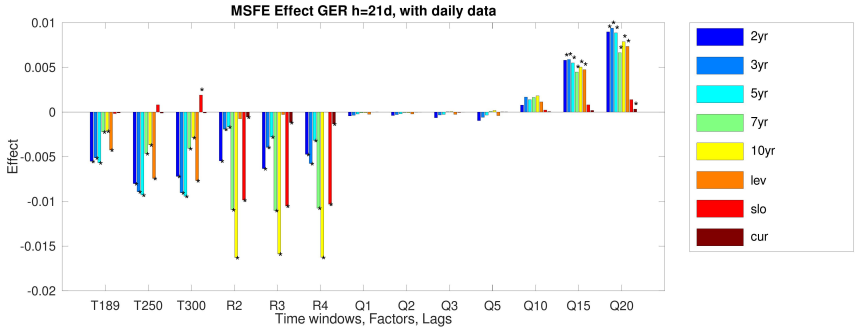
### 4.1 Time Window Sizes and Factor Numbers

The number of factors that govern a term structure are of high interest for the purpose of modelling yield curves. Diebold & Li (2006) assume that three factors underly the term structure of interest rates. In Section 3.2 we showed that these factors can be interpreted as the level, slope, and curvature of the term structure. We now use forecast performance as a criterion to determine the number of factors.

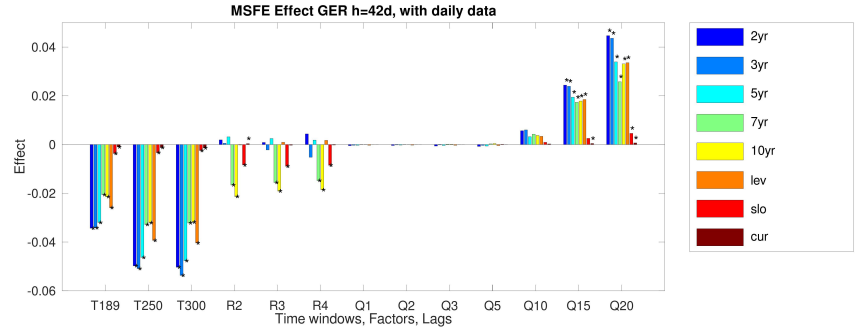
[2] We eliminate unreliable forecasts, i.e. if the forecast at $s + h$ differs more than 10 percentage points from the value at $s$. The forecast is then replaced with random walk forecast $y_{m,s+h|s} = y_{m,s}$. This happens very rarely.




(a) GER MSFE D5




(b) GER MSFE D21




(c) GER MSFE D42

Figure 5: Parameters of the ANOVA regressions for Germany with daily data. Here, MSFE effects at the $h = 5, 21, 42$ forecast horizons in Panels (a), (b), and (c) are depicted. The groups of bars represent the effects time window sizes, factor number, or lags for all maturities and linear combinations (i.e. level, slope, and curvature) A star ($\star$) above the bar indicates that the estimate is significant at the 1% level.

We will first focus on daily data. Specifically, the German term structure forecasts evaluated with MSFE for different forecast horizons can be used to illustrate the varying effects of the number of factors and time window sizes employed for estimation. Figure 5 depicts the parameters of the ANOVA regressions of daily data for 5 maturities and the three linear combinations levle ($lev$), slope ($slo$), and curvature ($cur$).

Panel (a) of Figure 5 depicts the results at the 5 day forecast horizon. Here, larger time window sizes are outperformed and lags cannot improve forecasts significantly. Strategies with more than one factor improve forecasts significantly over those with one factor[3] In Panel (b) of Figure 5 at the $h = 21$ days forecast horizon larger forecast horizons now consistently and significantly improve forecasts. Although the effects of additional factors are larger than those of larger time windows. In Panel (c) for $h = 42$ this feature is reversed and the effects of larger time windows are greater than those of additional factors. Larger time windows now improve forecasts by roughly two times the effect of additional factors. It is noteworthy that the effects of additional factors remain approximately equal over the three forecast horizons while the effects of larger time windows change signs and increase tenfold. Models with ten or more lags ($q \geq 10$) are consistently outperformed.

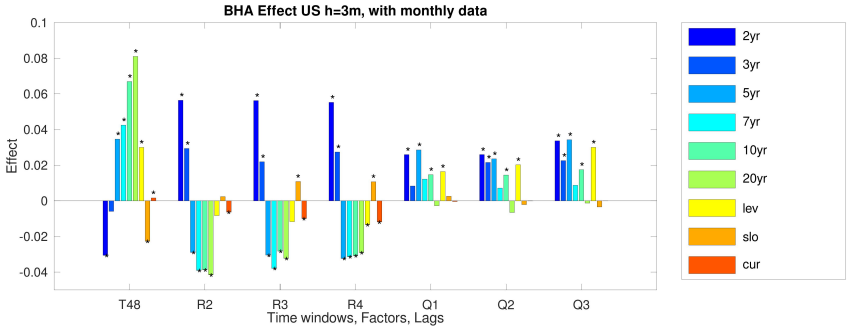



Figure 6: BHA effects for US monthly data with $h = 3$.

Figure 6 depicts the BHA effects for the US term structure using monthly data at the 3 month forecast horizon. Note that under BHA effects that improve forecasts have the opposite sign than under MSFE, i.e. positive. Here, we can observe that the larger alternative time window ($\tau = 48$) over the benchmark time window ($\tau = 24$). Furthermore, models with lags significantly improve forecasts with few exceptions. For the effects of additional factors we can not an odd feature across the term structure. While models with more than one factor

[3] In Matthies (2018) I noted that effects for model with more than 2 factors did not significantly improve forecasts if a model with 2 factors was chosen as the benchmark model.

improve forecasts for short maturities $2yr$ and $3yr$ as well as the slope factor they are outperformed for maturities of $5yr$ and larger as well as level and curvature.

### 4.2 Comparison of Different Frequencies

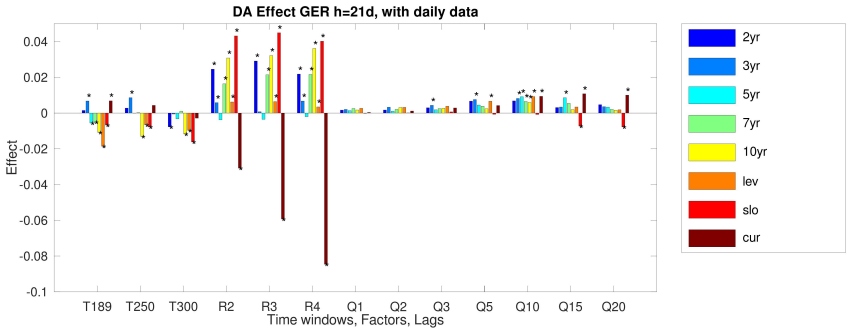
| Mat | 2 | 3 | 5 | 7 | 10 |
|---|---|---|---|---|---|
| | | | MSFE | | |
| D | 0.079957 | 0.076827 | 0.069107 | 0.056333 | 0.051683 |
| W | 0.075459 | 0.074509 | 0.066409 | **0.053496** | **0.049204** |
| M | **0.064123** | **0.064458** | **0.056601** | 0.054198 | 0.051366 |
| | | | DA | | |
| D | 0.57498 | 0.56108 | 0.56217 | 0.56010 | 0.55674 |
| W | 0.56508 | **0.57267** | 0.56453 | **0.56887** | 0.56128 |
| M | **0.59591** | 0.56522 | **0.58568** | 0.56010 | **0.58056** |

Table 2: For GER: Value of the respective minimising and maximising strategy for MSFE and for DA at 1 month forecasts with daily, weekly, and monthly data.

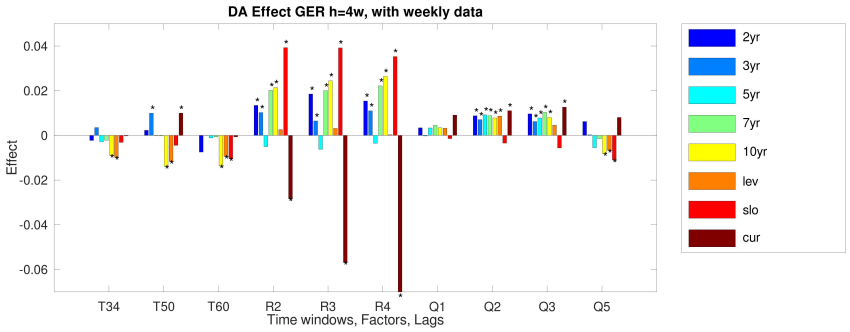
A simple comparison of which data frequency produces a minimum MSFE or maximum DA value is provided for the German term structure in Table 2. Daily data used for $h = 21$ days are compared to $h = 4$ weeks with weekly data and $h = 1$ month from monthly data. In Table 2 the respective minimising MSFE and maximising DA values for each data frequency and maturity are listed. The minimising or maximising values over the three data frequencies are in bold. We find no values in bold for daily data. This highlights that weekly and monthly data can produce better forecasts according to this rough comparison.

We now investigate the effects of alternative time window sizes, factor numbers, and lag numbers across data frequencies. For this purpose we analyse the ANOVA results for the German term structure under the DA criterion for daily data with $h = 21$, weekly data with $h = 4$, and monthly data with $h = 1$. They are depicted in the Panels (a), (b), and (c) of Figure 7.

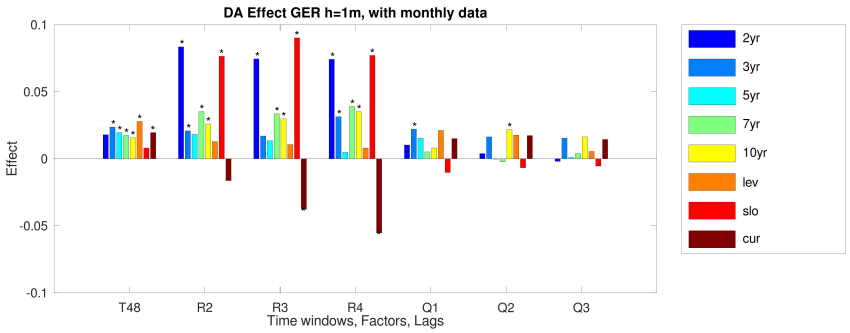
In Panel (a) the results for daily data are depicted. Here, the effects of for larger time windows vary across the term structure and linear combinations. Additional factors significantly improve forecasts for all maturities and linear combinations with the $5yr$ maturity and curvature being notable exceptions. Models that us autoregressive lags for the factors improve forecasts in contrast to those models without lags. Specifically, a lag number of ten appears ideal. Panel (b) shows the results for weekly data at the $h = 4$ weeks forecast horizon. Similar to the daily data the effects for alternative time window sizes vary. Additional factors improve forecasts with $5yr$ and curvature again being the exceptions. Here, a lag number of $q = 2$ or $q = 3$ appears ideal. In Panel (c) we find the results for monthly data. We can observe that the alternative larger time window for factor estimation improves forecasts. For additional factors the




(a) GER DA D21




(b) GER DA W4




(c) GER DA M1

Figure 7: Parameters of the ANOVA regressions for Germany with daily, weekly, and monthly data. Here, DA effects at the $h = 21$, $h = 4$, and $h = 1$ forecast horizons respectively with daily data, weekly data, and monthly data in Panels (a), (b), and (c) are depicted. For further explanation see the description in Figure 5

forecasts for curvature are the exception to general improvement. Here, models with lags generally improve forecasts although the effects are not significant.

## 5 Conclusions

Dynamic factor models allow for a large variety of forecast strategies. Factor analysis supports the idea that a level, slope, and curvature factor underly the yield curve. In a data set with all term structures we find evidence of a global level factor. But further factors are difficult to interpret. Briefly summarising the forecasting results we note that at longer forecast horizons larger time window sizes will usually improve forecasts. For the US the effects of additional factors can vary across the term structure. Using lags improve forecasts mainly under economic criteria. Data with larger time frequency can produce better results. For model selection across data frequencies results for the effects of additional factors are stable.

**2**

# Bibliography


Abbritti, M., Dell'Erba, S., Moreno, A., Sola, S. et al. (2018). Global factors in the term structure of interest rates. *International Journal of Central Banking*, *14*, 301–340.

Armesto, M. T., Engemann, K. M., Owyang, M. T. et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, *92*, 521–36.

Blaskowitz, O., & Herwartz, H. (2009). Pca-based ex-ante forecasting of swap term structures. *International Journal of Theoretical and Applied Finance*, *12*, 465–489.

Blaskowitz, O., & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International Journal of Forecasting*, *27*, 1058–1065.

Breitung, J., & Eickmeier, S. (2006). Dynamic factor models. *Modern econometric analysis*, (pp. 25–40).

Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica (pre-1986)*, *51*, 1281.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of econometrics*, *130*, 337–364.

Diebold, F. X., Li, C., & Yue, V. Z. (2008). Global yield curve dynamics and interactions: a dynamic nelson–siegel approach. *Journal of Econometrics*, *146*, 351–363.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*, 134–144.

Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The journal of Finance*, *46*, 555–576.

Hartzmark, M. L. (1991). Luck versus forecast ability: Determinants of trader performance in futures markets. *Journal of Business*, (pp. 49–74).

Litterman, R. B., & Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income*, *1*, 54–61.

Matthies, A. B. (2014). Validation of term structure forecasts with factor models. *The Journal of Risk Model Validation*, *8*, 65.

Matthies, A. B. (2018). *Modelling risk in financial economics*. Ph.D. thesis lmu.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of business*, (pp. 473–489).

Stock, J. H., & Watson, M. (2011). Dynamic factor models. *Oxford Handbook on Economic Forecasting*, .

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*, 1167–1179.

Yu, W.-C., Salyards, D. M. et al. (2009). Parsimonious modeling and forecasting of corporate yield curve. *Journal of Forecasting*, *28*, 73.

Yu, W.-C., & Zivot, E. (2011). Forecasting the term structures of treasury and corporate yields using dynamic nelson-siegel models. *International Journal of Forecasting*, *27*, 579–591.

# Methods of Detection of Non-Technical Energy Losses with the Application of Data Mining Techniques and Artificial Intelligence in the Utilities.

Marco Toledo[1], Carlos Álvarez [2], Boris Trelles[3] and Diego Morales [4]

[1,2] Universitat Politècnica de València - Institute for Energy Engineering (IIE), Spain
[3,4] Universidad Católica de Cuenca- Smart University 2.0 Project, Ecuador

martoor@doctor.upv.es – calvarez@die.upv.es – btrelles2013@gmail.com – dmoralesj@ucacue.edu.ec

**Abstract.** This research use the technical, commercial and social information of consumers of electric power distribution utilities through the application of data mining techniques and artificial intelligence proposes to determine groups of potential consumers who present infractions or damages in the measurement equipment, in such a way that the theft of energy is minimized and the income in the utilities is increased.

**Keywords:** Artificial Intelligence, Algorithm k-medias, Electric Power Distribution Utilities –EPDU-, Non-Technical Energy Losses, Multivariate Analysis.

## 1 INTRODUCTION

The globalization, the accelerated technological advances and the economic situation, lead the Electric Power Distribution Utilities -EPDU´s-, to make great efforts to reduce their non-technical energy losses, with projects focused on the planning of revisions to the systems of measurement, replacement of bare conductors to isolated concentric conductors or anti-theft, among other actions; however, they lack the use of ad-hoc techniques to determine fraud and/or errors in measurement systems through the use of software and specialized algorithms in data management. For this reason, this research proposes the analysis of the information of the electrical sector through the use of technical, economic, social and commercial variables, such as input data to the algorithm that will be constructed through the use of tools of multivariate analysis, based on the methodology of correlation analysis and automatic classification for pattern recognition, also known as cluster analysis, through the k-means algorithm, to determine the number of homogeneous groups, k-means cluster, and finally the training of a neural network through artificial intelligence.

The proposed algorithm will allow determining the group of consumers that should be reviewed by specialized groups in the measurement system in order to quantify the energy consumed and not billed, either by damage to the measurement systems or by the intervention of third parties in the systems of Distribution.

## 2 OBJECTIVE

Determine non-technical electrical energy losses through the application of data mining techniques and unsupervised detection methods and artificial intelligence through the location of potential consumer groups in EPDU´s.

## 3 CURRENT SITUATION

The problem of the utilities lies in the lack of control of the administrative and technical processes of the EPDU´s, in public policies, on regulatory and judicial issues to minimize their incidence, which causes the losses of electric power to reduce the economic income of the EPDU´s for energy consumption not invoiced. From the last statistics obtained in the Electrification Master Plan -PME, decennial planning document used by the electricity sector throughout its supply chain, it is established that energy losses at the distribution stage are in the order of 11,49%, in the following way.

**Table 1.** ENERGY LOSSES IN THE EPDU´s

| Company | Technical Losses (MWh). | Non-technical losses (MWh). | Losses of the System (%). |
|---|---|---|---|
| CNEL-Manabí. | 207056,9 | 202267,0 | 23,6% |
| CNEL-Esmeraldas. | 56793,3 | 83782,6 | 22,6% |
| CNEL-Los Ríos. | 30762,8 | 47080,9 | 17,3% |
| CNEL-Milagro. | 51304,1 | 62086,7 | 15,8% |
| CNEL-El Oro. | 109741,9 | 70960,5 | 15,6% |
| CNEL-Santa Elena. | 50376,4 | 52781,6 | 15,2% |
| CNEL-Guayas Los Ríos. | 206907,2 | 124544 | 15,1% |
| CNEL-Sucumbios. | 34996,3 | 14990,4 | 12,4% |
| CNEL-Santo Domingo. | 59656,5 | 22865,4 | 11,4% |
| CNEL-Guayaquil. | 380604,7 | 191343,6 | 10,3% |
| E.E. Riobamba. | 28949,0 | 11786,3 | 10,3% |
| E.E. Sur. | 25252,1 | 11804,3 | 10,2% |
| E.E. Norte | 39083,8 | 18185,7 | 9,3% |
| E.E. Cotopaxi. | 42699,1 | 9621,9 | 8,7% |
| E.E. Galapagos. | 3479,0 | 822,6 | 8,0% |
| CNEL-Bolivar. | 7258,9 | 57,8 | 7,9% |
| E.E. Centrosur | 66319,8 | 3950,3 | 6,3% |
| E.E. Ambato. | 36623,7 | 671,3 | 5,6% |
| E.E. Quito. | 222251,9 | 23363,8 | 5,4% |
| E.E. Azogues | 4420,9 | 624,9 | 4,6% |

Currently, Ecuador has 4,6 million consumers in 20 EPDU´s, investments made through the Ministry of Electricity and Renewable Energy in the period 2009 – 2018 in programs for the mitigation of non-technical losses and technical losses in the different stages of the distribution chain led to the replacement of more than 2'000.000 measurement systems allowing greater control in the consumption and billing of consumers.

### 3.1 Losses of electrical energy in the EPDU´s.

In the last decade, the utilities has been carrying out an important management for the reduction and control of non-technical energy losses, which are caused by the manipulation of the measurement systems, by the administrative management in the taking and recording of readings, billing, collection and the erroneous application of the tariff in the final use of energy.

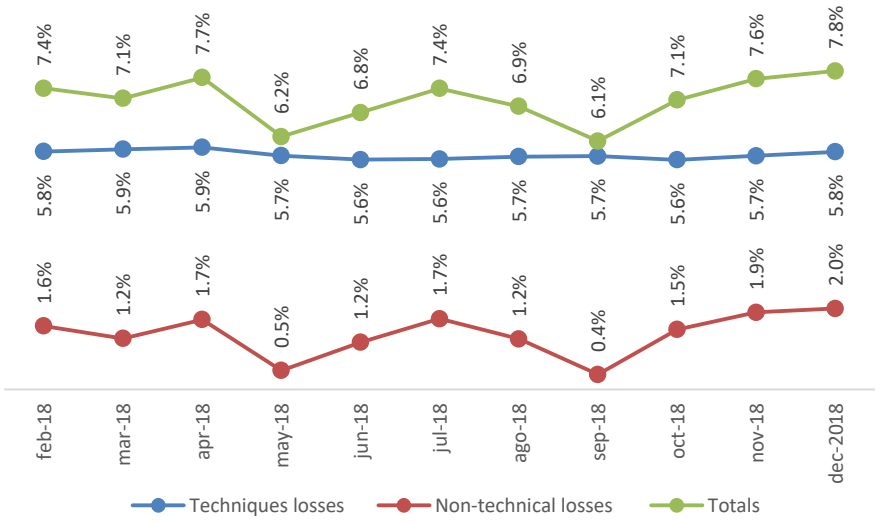



**Fig. 1.** Losses in the CENTROSUR (CENTROSUR is the Utilities of the Cuenca city in Ecuador and case of Study).

The annual losses maintained from 2006 to 2018 are on average at 7.27%, a significant value within the framework of efficiency in one utility, adjusted to Latin American levels, where Peru is the country with the highest lower energy losses in the region, this way CENTROSUR is positioned as a benchmark in Latin America.

## 4 RESEARCH AND DEVELOPMENT

The EPDU have robust computer systems for monthly billing of power and during a chain of supply request, the consumer provides information of a personal, technical and geographical nature prior to the installation of the new measurement system,

which is validated with the information provided by entities of the sector.

Ecuador through the Law of the National System of Public Data Registration-LSNRDP same that ... "creates and regulates the system of registration of public data and its access, in public or private entities that administer said bases or registers...".

Based on the foregoing and complying with what is indicated in the LSNRDP this research takes the information from the databases of the Regional Electric Utility, whose variables are of social, economic and technical order of the consumers connected to the System of Electric Power Distribution -SEPD- to develop a predictive model for analyzing information from different sources through the use of data mining techniques, unsupervised classification, grouping methods, predictive analysis, all these techniques organized under an Artificial Intelligence process -IA.

These techniques allow the creation of models with the purpose of extracting the most relevant information applied to large sets of data sources that can be transactional, social, economic, technical and geographic data that come from different sensors. I, innovation often arise from the combination of data from various sources, to be analyzed through the use of tools to extract knowledge and trends by applying different techniques of machine learning in order to locate patterns in the data and create models that predict future results. There is currently a wide range of machine learning algorithms, including linear and non-linear regression, neural networks, support vector machines, decision trees, etc.

Predictive analysis (Cluster Obtaining) is the process of using data analysis to make predictions based on the data, in this process the data is used together with analytical, statistical and machine learning techniques in order to create a model predictive (Consumer group) to predict future events (Failed data on consumer patterns). The workflow of the model consists of the following steps:

- Import data from various sources or databases.
- Clean the data by eliminating the outliers.
- Identify the peaks of data, the missing data or the anomalous points that should be eliminated from the bases.
- Develop an accurate predictive model based on aggregate data through statistics, curve fitting tools or machine learning.
- Predicting non-technical losses is a complex process with many variables, so neural networks are used to create and train a predictive model.
- Integrate the model into a non-technical loss prediction system in a production environment.

To extract the knowledge they hold, a precise predictive model is needed through mathematical methods and calculation to predict an event or a result (Conglomerate method). Through an iterative process, the model is developed through a set of training data and then tested and validated to determine its accuracy in order to make predictions.

Currently the EPDU´s seek to improve their technical and economic indicators from the reduction of non-technical energy losses, which is why this research is particularly important, since, with the application of the model based on the predictive analysis of potential consumers with theft or who present errors in the measurement systems, these are detectable with greater certainty, located and reviewed by the offi-

cials responsible for operating and maintaining the measurement systems in order to determine if they are in acceptable conditions and if the existence of failure or theft is determined, this energy is billed in such a way that the ED recovers the investment made in the purchase of energy in the Ecuadorian Electricity Market -MEE-.

Figure 2 describes the workflow of the minimum algorithm necessary to perform the forecast of energy losses.

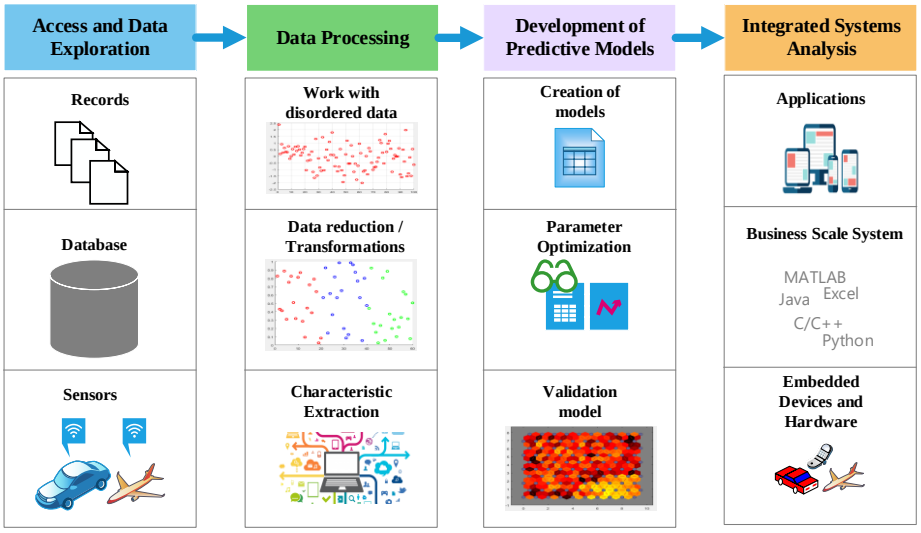



**Fig. 2.** Step-by-step workflow to predict energy losses.

## 5 RELATIONSHIP WITH ACTIVITES AND APPLICATION IN THE COMPANY OR SECTOR

Measurement systems are an essential part of the EPDU´s, from these depends on the economic income received by the billing of the electric power, which is why it is of the utmost importance to keep these measurement systems in the best technical conditions to guarantee the consumer a real billing of electricity consumption. However, for EPDU´s it becomes increasingly complex to determine and control non-technical losses, since the consumer seeks to evade or alter the consumption of electricity, violating the integrity of the measurement system.

Based on the variables obtained from the commercial system and of the geographic information system -GIS-, as showed in the next table 2, said variables geographical, technical, commercial and from socioeconomic groups, are denominated "Matrix of Variables". This matrix has a dimension of [38 X 17.345]. Where 38 are the variables and 17.345 are individuals.

**Table 2.** MATRIX OF VARIABLES

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **GEOREFERENCING VARIABLES.** | | | | | | | | |
| CODE | RATE | PROVINCE | CANTON | PARISH | SECTOR | ROUTE | SEQUENCE | ADDRESS |
| **VARIABLES ENERGY CONSUMPTION (12 MONTHS)** | | | | | | | | |
| | | JANUARY | FEBRUARY | MARCH | APRIL | MAY | | |
| **VARIABLES TECHNICAL DATA.** | | | | | | | | |
| METER | TYPE OF METER | YEAR METER | Nº SERIES | F.M. DEM | VOLT. | AMP. | N° FASES | V. CONSTR. |
| **VARIABLES DATA COMMERCIALIZATION.** | | | | | | | | |
| INST. DATE | REV. DATE | RE-BILLING DATE | RE-LIQUID. DATE | FACT. MULT. | FP | M. DEBT | STATE | WRIT IDENTITY |
| **SOCIOECONOMIC GROUPS.** | | | | | | | | |
| | | SOCIOECONOMIC CATEGORY. | | CONSUMPTION kWh/MONTH | | | | |

### 5.1 Exploratory analysis of the data

The matrix [X] is constructed based on the data of the commercial system of the EPD, this consolidation defines the characteristics technical, economic and social of the EPD, based on the linear dependence between variables, so initially obtained 38 variables and after executing a first data exploration, 19 variables were eliminated, due to the redundancy of their data; The values of the p scalar variables in each of the n elements can be represented in a matrix [X], which for the study will be [19x1,526], where n = 1,526 individuals and 19 variables, arranged as follows:

The consumers of the EPDU´s will represent the n elements, these can vary in quantity according to the block of analysis that executed in the model. In this research, the variables obtained from the different sources of information are of a social, economic and technical nature as described below; 1) Client Code, 2) Rate, 3) Province, 4) Canton, 5) Parish, 6) Sector, 7) Route, 8) Sequence, 9) Direction, 10) Variables of Energy Consumption of the 12 months before analysis (kWh / month), 17) Meter Mark, 18) Meter Type, 19) Meter Year, 20) Serial No., 21) Demand Multiplication Factor, 22) Voltage, 23) Amperage, 24) No. Phases, 25) Type of Home Construction, 26) Date of Installation, 27) Date of Review for Control of Measurement, 28) Date of Re-billing, 29) Date of Re-liquidation, 30) Factor of Multiplication, 31 ) Power Factor, 32) Months of Debt, 33) Status (active-not active), 34) Socio-Economic Category, 35) Variance, 36) Deviation, 37) Average, 38) Coefficient of Variation. This matrix is the basis of the study.

### 5.2 Analysis of atypical data

The analysis of atypical data is of great importance in a study of data mining, since this depends on the detection of atypical data or groups of data which could in some cases distort the data of the covariance matrix [1] for this reason it is necessary to study to eliminate all suspicious points from the sample, so that we avoid confusing

and calculate the vector of means and the covariance matrix without distortions. To determine these possible values we use simple rule is to consider those observations as suspicious that (|Xi-med(X)|)/(media(X))>4,5, where med(X) is the median of the observations, and media(X) is the median of the absolute deviations |Xi-med(X)|, which is a robust measure of the dispersion.

### 5.3 Univariate analysis.

The descriptive analysis of a variable involves calculating its mean, which is the center of gravity of the data, defining the standard deviation and calculating the measure of variability in relation to the mean, averaging the deviations between the data and its mean. To compare the variability of different variables, we construct measures of relative variability that do not depend on the units of measurement. One of these measures is the coefficient of variation as observed in equation 1.

$$CV_j = \sqrt[2]{\frac{{S_j}^2}{{X_j}^2}} \tag{1}$$

Where ${S_j}^2$ is the variance and ${X_j}^2$ is the mean squared.

Once the original matrix is analyzed, important data is observed such as; Since variable 10 to 16 has a coefficient of mean asymmetry with respect to the rest of the variables, variable V35 has a high standard error due to the magnitudes of unequal energy consumption.

### 5.4 Non-linear transformation.

The objective of applying this mathematical technique is to linearize the data, given that the matrix has magnitudes in different units, it is essential to apply the cosine (Cos) to each variable, so the variability of the transformed variable is independent of the variables. units of measurement initially considered as shown below.

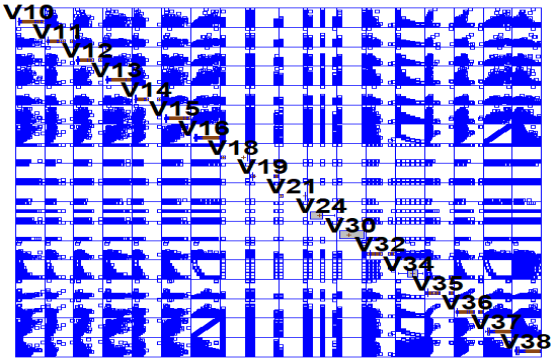



**Fig. 3.** Dispersion diagram of the Linearized Matrix.

### 5.5 Correlation of variables.

This matrix shows Pearson's product of correlations between each pair of variables. The range of these correlation coefficients ranges from -1 to +1they measure the strength of the linear relationship between the variables. The P-value tests the statistical significance of the estimated correlations P-Values below 0.05 indicates correlations significantly different from zero, with a confidence level of 95.0%. The following pairs of variables have P-values below 0.05.

In figure 4, the variables V10, V11, V12, V13, V14, V15, V16, have great similarity, demonstrating the existing correlation between the energy variables (kWh/month). Also, V36 and V37, represent the statistical analysis of the deviation and the average of the monthly energy, V32 V33 and V35 represent the technical and social characteristics of the consumers. However, we observe that the variable V34 (Socioeconomic Category), does not show similarity with the rest of the variables, because its parameters are a function of the geographical stratification for the calculation of the projected maximum unit demand.

| | V10 | V11 | V12 | V13 | V14 | V15 | V16 | V18 | V19 | V21 | V24 | V30 | V32 | V34 | V35 | V36 | V37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V10 | | | | | | | | | | | | | | | | | |
| V11 | 0,79 | | | | | | | | | | | | | | | | |
| V12 | 0,81 | 0,82 | | | | | | | | | | | | | | | |
| V13 | 0,79 | 0,77 | 0,86 | | | | | | | | | | | | | | |
| V14 | 0,72 | 0,71 | 0,78 | 0,83 | | | | | | | | | | | | | |
| V15 | 0,73 | 0,72 | 0,79 | 0,83 | 0,86 | | | | | | | | | | | | |
| V16 | 0,89 | 0,88 | 0,93 | 0,93 | 0,90 | 0,91 | | | | | | | | | | | |
| V18 | 0,01 | 0,00 | 0,01 | 0,00 | 0,01 | 0,01 | 0,01 | | | | | | | | | | |
| V19 | -0,02 | -0,02 | -0,02 | -0,01 | -0,02 | -0,02 | -0,02 | -0,42 | | | | | | | | | |
| V21 | -0,09 | -0,08 | -0,06 | -0,09 | -0,07 | -0,08 | -0,09 | -0,01 | 0,01 | | | | | | | | |
| V24 | 0,04 | 0,03 | 0,03 | 0,04 | 0,04 | 0,05 | 0,04 | 0,29 | -0,22 | 0,04 | | | | | | | |
| V30 | 0,11 | 0,11 | 0,10 | 0,11 | 0,10 | 0,10 | 0,12 | -0,19 | 0,21 | -0,02 | -0,66 | | | | | | |
| V32 | 0,04 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,02 | -0,02 | 0,01 | 0,01 | 0,02 | | | | | |
| V34 | -0,74 | -0,74 | -0,79 | -0,79 | -0,77 | -0,77 | -0,85 | -0,02 | 0,03 | 0,04 | -0,05 | -0,07 | -0,07 | | | | |
| V35 | 0,17 | 0,34 | 0,18 | 0,15 | 0,29 | 0,18 | 0,24 | 0,01 | 0,00 | -0,02 | 0,02 | 0,04 | 0,02 | -0,14 | | | |
| V36 | 0,40 | 0,50 | 0,43 | 0,42 | 0,50 | 0,48 | 0,50 | 0,03 | -0,02 | -0,05 | 0,06 | 0,06 | 0,05 | -0,39 | 0,75 | | |
| V37 | 0,89 | 0,88 | 0,93 | 0,93 | 0,90 | 0,91 | 1,00 | 0,01 | -0,02 | -0,09 | 0,04 | 0,12 | 0,05 | -0,85 | 0,24 | 0,50 | |
| V38 | -0,18 | -0,16 | -0,20 | -0,19 | -0,14 | -0,10 | -0,18 | 0,03 | 0,00 | 0,00 | 0,06 | -0,01 | 0,01 | 0,29 | 0,17 | 0,38 | -0,18 |

**Fig. 4.** Dispersion diagram of the Linearized Matrix.

### 5.6 Analysis of main components -ACP

The technique of -ACP allows representing optimally in a space of small dimension, observations of a general p-dimensional space with minimal loss of information, facilitating the interpretation of the data. In this sense, the study uses this technique to determine more precisely which are the most useful variables to the group and find the possible offenders and/or damage of the measurement systems, for that, the software tool STATGRAPHICS Centurion is used.

The criterion for selecting the number of Principal Components -PC-, is given by the value that explains more than 70% of the variability of the original data, with the purpose of obtaining a reduced number of linear combinations of the 18 variables that explain the greater variability in the data. In this case, 5 components have been extracted since they had eigenvalues greater than or equal to 1.0 together they explain

77.76% of the variability in the original data. This information is corroborated with the criterion of the fall in predictive capacity of the increasing eigenvalues indicated in figure 5.

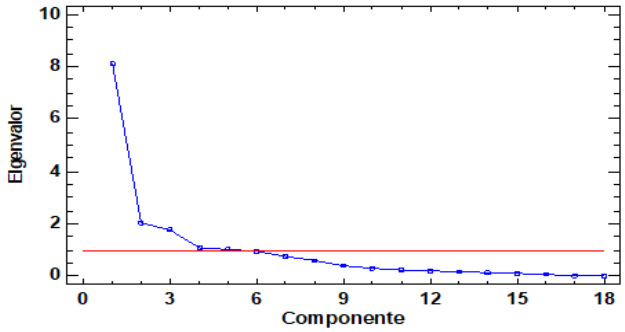



**Fig. 5.** Main Components in STATGRAPHICS

**Table 3.** CP SUMMARY – VARIABLES - INDIVIDUALS

| | **VARIABLES (+)** | **VARIABLES (-)** | **CHARACTERISTIC** |
|---|---|---|---|
| **CP1** | V10, V11, V12, V13, V14, V15, V16, V36, V37, V18, V19, V21, V24, V30, V32, V34, V35, V38 | V34 | Administrative, Commercial, Economic, Technical, Social. |
| **CP2** | V10, V11, V12, V13, V14, V15, V16, V36, V37 | - | Energy variables. |
| **CP3** | V35, V36, V37 | - | Statistical Variables, Deviation, Variance, Variation Coefficient and Mean. |
| **CP4** | V34 | - | Social Variables. |
| **CP5** | V18, V19, V21, V24, V30, V32 | - | Technical and Economic Variables. |

Table 3 shows the summary of the Principal Components analysis, the variables that intervene in each of its components and the coefficients according to the characteristics and correlations.

### 5.7 Reduction of the variables by the criterion of the expert.

Consumers take different forms to evade a correct measurement, or seek direct connection alternatives, in order to reduce or avoid the payment of electric power values. For many years researchers spent time, technical and administrative resources, to record and store the information that comes from the losses of non-energy techniques, so that. For the realization and obtaining of the groups, scenarios are used for the elaboration of the decision tree with the most relevant rules defined as indicated below. This research uses around 40 causes for the criteria of the review of measurement systems, which could be increased in the proposed mathematical model.

**Table 4.** RULES FOR DECISION TREE

| **Scenarios for the elaboration of rules in the decisión tree.** |
|---|
| a.-Stratification of consumption (<10; >500 kWh) |
| b.-Measurement systems not reviewed in the last 6 months. |
| c.-Date of installation of the measurement system. |
| d.-Year of manufacture of the energy meter. |
| e.-Manufacture mark of the energy meter. |
| f.- State (Active, For Cut, Cut). |
| g.- Type of Housing Construction. |
| h.- Socioeconomic level. |
| i.- Type of Rate. |
| j.- Location by geographical area. |
| k.- Voltage Level. |
| l.- Consumption deviation level. |
| m.- Coefficient of variation. |

### 5.8 Conglomerate analysis implementation of the k-means algorithm.

The objectives of this research is to group consumers according to their similar characteristics, to automatically classify the observations made by the cluster analysis using the k-means algorithm, to test of the variability reduction test F, comparing the sum of squares within the groups -SDCG-, and calculating the relative reduction of the variability when increasing an additional group, by means of equation 2.

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1)/(n-G-1)} \quad (2)$$

We obtain that 5 groups are optimal, because the variability reduction test has a value of $F = -0.5$, and relates the n individuals according to the investigator's expertise. To apply the k-Medias algorithm, we use rules to technically reduce the initial data matrix, for example, it is considered that the year of manufacture of the energy meter is less than 2015, that the months of debt are less than 2 months. months, consumers are not considered who are in quintiles 3 and 4 because they possibly maintain consumption correlated with the standard of living, culture of the population, and average statistics correlated between the variables of the model, this rule represents 22.80 % in the consumption stratum between 181 and 310 kWh/month, and in the 4 quintile, 49.30% in the stratum between 111 and 180 kWh/month. The matrix was reduced from [38X17.435] to a matrix of [19X1.526], taking into account that the variables they provide with consumer location information were not executed in the algorithm. Once the model is applied, 5 homogeneous groups are obtained.

**Table 5.** CRITERIA FOR REVIEW OF MEASUREMENT

| Clúster k-Medias | Members | Percentage |
|---|---|---|
| **G1** | 58,0 | 3,8 |
| **G2** | 365,0 | 23,9 |
| **G3** | 326,0 | 21,3 |
| **G4** | 18,0 | 1,1 |
| **G5** | 759,0 | 49,7 |

The group G1, represents the variation in the statistical magnitudes with 3.8%. In addition, variables of a socioeconomic and geographical nature are included. This group becomes the source of potential revisions.

G2, is formed by 365 consumers, contributing with 23.92% and represents the energy variables of the database studied.

G3, is formed by 21.36% and represent more than the energetic variables are geographic variables (location).

G4, contains the variable of the socioeconomic category; from the beginning of the analysis we had seen that this does not have a high correlation coefficient with respect to the rest of the variables.

G5 represents the group clearly of the energy variables, with 49.74%.

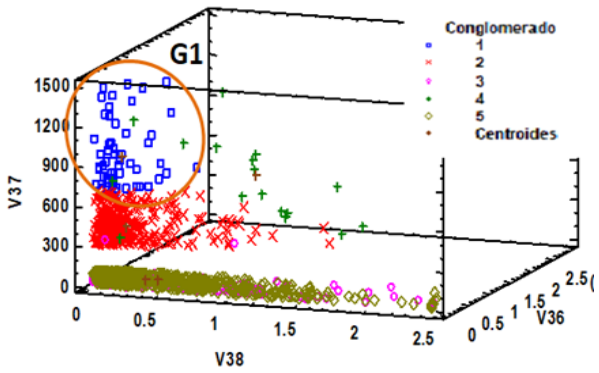



**Fig. 6.** Cluster Dispersion Chart k-Medias.

Within the data analysis there are classification techniques in Artificial Intelligence -IA, such as the Neural Networks RN, so this research applies the no supervised classification tool of the STATGRAPHICS software, from which it is obtained a based on 19 input variables the 1,526 cases with 11.40% correctly classified by the network.

For example, the nearest neighbor for row 1 was V38 = 1.13 and the second closest neighbor was V38 = 0.92. In fact, the true value of V38 was 1.05. Among the 1526 cases used to train the model, 167 individuals were classified correctly, which means that they should be reviewed, since they correspond to the group of potential consumers with novelties.

Once the determined the group and using the graphical tool of GIS, the location of the different consumers that should be re-viewed is represented, in order to determine

if the operation of the measurement sys-tem is correct, or the reason by which the consumer corresponds to the group that presented the greatest variability in the data of its variables.

## 6 RESULTS OBTAINED AND EXPECTED

These goals led to investigate new techniques for the detection of fraud and/or damage to measurement systems, which is why this methodology was applied during the 2016 year with 3 groups, specialized in measurement systems perform the review of 906 measurement systems in the field and obtained an energy recovery no billed of 377.860 kWh/month. To evaluate the results of the applied methodology, the study determining that of 18.160 revisions they were detected 785 with anomalies between altered and damaged measurement systems, giving an efficiency of 4.35%, relatively optimistic value, since the losses no technical not exceed 1,20% of the total losses.

## 7 CONCLUSIONS.

The proposed Methodology determined the optimal form to recover the greater economic amount due to non-technical losses in EPDU´s, through data mining techniques, multivariate analysis methods accompanied by automatic monitoring algorithms and the use of artificial intelligence models, based on information from the commercial system, socioeconomic information and GIS. The analysis of these variables was carried out using , classification algorithms and groupings (k-means) to obtain a definitive list of the potential revisions of the measurement systems, in such a way that it can be located in the geographic information system to optimize the revisions and their routes, recovering the greatest amount of Unclaimed Energy. The projects that originate through the use of this methodology will allow for an economic return in the very short term.

## References

1. D. Peña, “Análisis de Datos Multivariante”, 2da. Edición, Madrid, McGraw-Hill/Interamericana de España, S.A.U., 2002 pp. 13-243.
2. Agencia de Regulación y Control de Electricidad – ARCONEL, “Plan Maestro de Electrificación 2013-2022 y 2016-2025, Estadística del Sector Eléctrico Ecuatoriano.
3. O. Yakubu, N. Babu C., and O. Adjei, “Electricity theft: Analysis of the underlying contributory factors in Ghana,” Energy Policy, vol. 123, no. November 2017, pp. 611–618, 2018.
4. M. Toledo, P. Cando, P. Mendez, C. Álvarez and D. Morales, “Determination of energy losses in distribution transformers using a compensation algorithm in energy meters, ITISE 2018 ” pp. 1963–1702, 2018.
5. G. M. Messinis and N. D. Hatziargyriou, “Review of non-technical loss detection methods,” Electr. Power Syst. Res., vol. 158, pp. 250–266, 2018.

# End of charge detection of batteries with high production tolerances

Andre Loechte, Ole Gebert, and Peter Gloesekoetter

University of Applied Sciences Muenster,
Department of Electrical Engineering and CS,
Stegeweraldstr. 39, 48565 Steinfurt, Germany
a.loechte@fh-muenster.de
http://www.fh-muenster.de

**Abstract.** During the development of new battery technologies high production tolerances can occur due to the manual manufacturing which is not as precise as machine-made. When putting these prototypes into operation, one of the most important exercise is finding a criterion of a full battery. This can be challenging when parameters like the capacity or the end of charge voltage are not precisely known due to the tolerances. In the majority of cases overcharging should be avoided as it harms the battery. This paper proposes a new criterion for detecting the end of charge that is based on the rate of change of electrochemical impedance spectra of the examined batteries. Device parameter fluctuations influence every measurement. Therefore, using the rate of change offers the possibility to not depend on these fluctuations.

**Keywords:** Electrochemical impedance spectroscopy, battery analysis, state of charge

## 1 Introduction

When developing new battery technologies, fundamental research means assembling new batteries by hand since a production line is not worthwhile for building and testing individual cells. This causes high production tolerances to occur because manual manufacturing is not as precise as machine-made.
When putting these prototypes into operation, problems can arise due to the varying parameters. One of the most important exercise is finding a criterion of a full battery. This can be challenging when parameters like the capacity or the end of charge voltage are not precisely known due to the tolerances. Furthermore, new battery types do not necessarily rely on the same stopping criteria. For example zinc-air secondary batteries do not offer an end of charging voltage. Its charging current is not going to decrease when the battery is full and the charging voltage is held at a fixed value. But instead of de-oxidising zinc oxide, hydrogen is produced.
In the majority of cases overcharging should be avoided as it harms the battery[1]. Another even more dangerous consequence is the possibility of an explosion. Especially lithium based batteries are known for their need of compatible

ambient and charging parameters[2].
This paper proposes a new criterion for detecting the end of charge that is based on the rate of change of electrochemical impedance spectra of the examined batteries. Device parameter fluctuations influence every measurement. Therefore, using the rate of change offers the possibility to not depend on these fluctuations.

## 2 Implementation

As there is no criterion for estimating the state of charge of zinc-air batteries, these type of cells are used to develop the new end of charge detection system. Therefore, impedance spectra were measured regularly each 30 min with a setup that is based on a low-cost pc oscilloscope[3]. Applications in electric vehicles can even use an already existing electric motor to become even more inexpensive[4]. The measurement interval is based on the capacity and on the charging current and correspond to a difference of state of charge of 1 %. The interval need to be adjusted for batteries with different capacity or charging current configurations. The resulting spectra of one charging cycle are shown in Fig. 1. The colour of each characteristic specifies the state of charge of the battery. Red spectra indicate an empty battery (SoC=0 %) while blue spectra belong to a fully charged or even an overcharged battery (SoC=100 %). As one can see most variation can be found for impedance values building the semicircle on the right-hand side. This semicircle is formed by low frequency impedance values. The proposed method uses the radius of the semicircles. Therefore, circle models for each spectrum have to be generated.

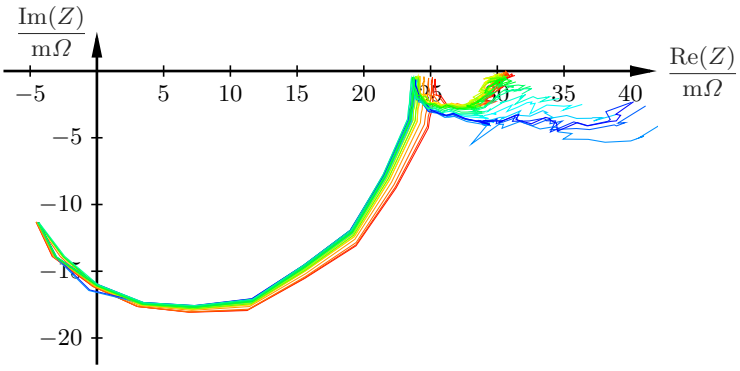



**Fig. 1.** Impedance spectrum colour weighted from red (SOC=0) to blue (SOC=100)

There are two challenges when creating a circle model based on a spectrum. On the one hand, the impedance values building the left semicircle needs to be cut

away. Here, the angle of the impedance values is a good feature for determining the splitting point. The gradient of angles is much higher for the left semicircle. More precisely the index of the splitting point is chosen to be

$$k_{split} = \frac{\partial angle(Z_k)}{\partial k} > \left(5 \cdot RMS(\frac{\partial angle(Z_k)}{\partial k})\right)$$

where $Z_k$ represents the impedance at index $k$.
On the other hand, especially the semicircle on the right-hand side is interfered with a lot of noise due to the low-cost measuring setup. However, the RANSAC (random sample consensus) algorithm gives good results by finding outliers that are not used to calculate the circle model[5]. First, 3 impedance points are selected randomly and used to create an initial circle model. Then the other impedance values are tested against that model. If the distance of an impedance value and the model is lower than 2 % of the maximum absolute value of that spectrum, it is considered as an inlier. These steps are repeated for 15 sets of randomly chosen starting values. Finally, the optimised circle model is calculated by using the impedance values of the biggest set according to Bucher[6]. The relation of a circle is given by

$$(x - x_c)^2 + (y - y_c)^2 = r^2$$

where $x_c$ and $y_c$ denote the centre of the circle and $r$ the radius. Substituting

$$A = x_c^2 + y_c^2,$$

$$B = 2 \cdot x_c,$$

$$C = 2 \cdot y_c$$

results in a linear system of equations:

$$\begin{bmatrix} 1 & -x_1 & -y_1 \\ 1 & -x_2 & -y_2 \\ 1 & -x_3 & -y_3 \\ \vdots & \vdots & \vdots \end{bmatrix} \cdot \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ x_3^2 + y_3^2 \\ \vdots \end{bmatrix}$$

that is solved using the least-squares solution of the system. The resulting circles are shown in Fig. 2.

## 3 Results

Since the radius of spectra is device depended, its derivative with respect to the charged energy is used for the criterion. The development of the derivative during one charging cycle is shown in Fig. 3. While the variation of the radius is quite small at the beginning of the charging cycle, it increases rapidly after 60 A h of charge. After 85 A h of charging the variation becomes smaller once

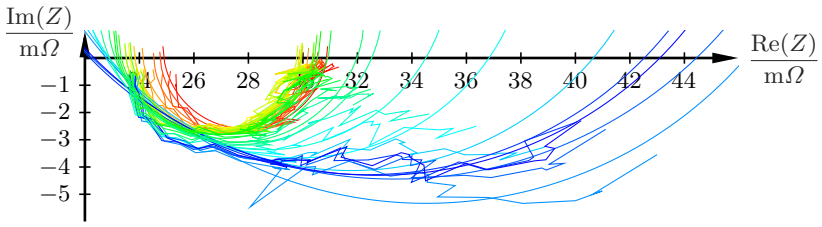



**Fig. 2.** Circles fitted into the colour weighted from red (SOC=0) to blue (SOC=100) spectrum

again.
These 3 phases correspond to different chemical phases of the charging process. During the first phase, zinc is being de-oxidised which increases the state of charge of the battery. When the charging process is close to finish, an attending electrolysis process takes place which decomposes the electrolyte. During the second phase the ratio of de-oxidising zinc becomes smaller while the electrolysis process becomes stronger. This change of process types leads to an increased variation of the impedance spectra resulting in higher derivatives of the radius values. In the last phase the de-oxidising stops completely. Since the rations of the two processes is not changing anymore, the derivative becomes smaller.

**Fig. 3.** Determination of charge to overcharge change at 60 A h

The criterion uses the root mean square of the derivative values at the beginning of the charging cycle to determine the comparison value

$$\Delta r_{limit} = \sqrt{\left(\frac{3}{n}\sum_{k=1}^{n/3}\frac{\partial r(k, Q_{charged})}{\partial Q_{charged}}\right)} \cdot 8.$$

The resulting threshold value is also plotted in Fig. 3. Now a battery is considered to be full if the derivative of the radius is greater than the comparison value:

$$\frac{\partial r(Q_{charged})}{\partial Q_{charged}} > \Delta r_{limit} \rightarrow \begin{cases} True & \text{Battery is full} \\ False & \text{Battery is not full} \end{cases}$$

Fig. 4 shows the classification of each spectrum. As expected the spectra of a charging battery are located densely in a small area. By contrast, the spectra of an overcharging battery are characterised by a strong variance between the spectra.

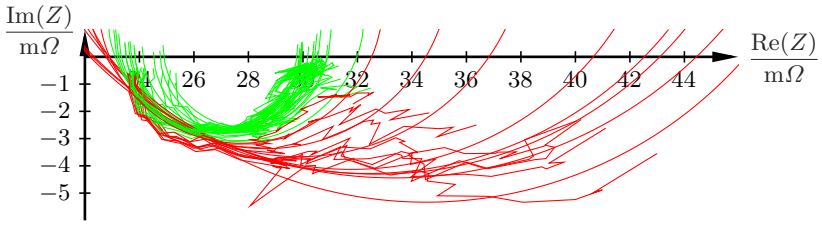



**Fig. 4.** Impedance spectrum split in charging (green) and overcharging (red) - transition at roughly 60 A h

In this example the anticipated capacity of the produced cell was 100 A h. However, although the battery was charged for 50 h at 2 A, only 60 A h could be extracted in the following discharging cycle. Thus, the criterion withstands the practical measurements.

## 4 Conclusion and outlook

This paper proposed a new criterion for detecting the end of charge that is based on the rate of change of electrochemical impedance spectra of the examined batteries. Using the rate of change neglects the dependence on these fluctuations. The criterion was successfully applied to zinc-air batteries and verified by massively overcharging a battery and comparing the state of charge at the estimated end of charge point with the actual extractable energy during the following discharge cycle.
So far, the criterion was tested with zinc-air batteries only. Further testing is necessary to evaluate possibilities for making use of this criterion for other cell technologies as well.

## References

1. Sexton, E. D., Nelson, R. F., Olson, J. B.: Improved charge algorithms for valve regulated lead acid batteries. Annual Battery Conference on Applications and Advances. 15 (2000)
2. Kaypmaz, T. C., Tuncay, R. N.: Diagnosing overcharge behavior in operation of Li-ion Polymer batteries. 2012 IEEE International Conference on Vehicular Electronics and Safety (2012)
3. Kiel. M.: Impedanzspektroskopie an Batterien unter besonderer Bercksichtigung von Batteriesensoren fr den Feldeinsatz. Aachener Beitrge des ISEA, Aachen (2013)
4. Howay, D. A., Mitcheson, V. Y., Offer, G. J., Brandon, N.P.: Impedance measurement for advanced management systems. World Electric Vehicle Symposium and Exhibition. 27 (2013)
5. Fischler, A.M., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM. 24, 381–395 (1981)
6. Bucher, I.: Circle fit, `https://de.mathworks.com/matlabcentral/fileexchange/5557-circle-fit?focused=5059278\&tab=function` (2004)

# Climate change: climate missing data processing, modeling rainfall variability of Soummam watershed (Algeria)

Amir AIEB[a,b], Khalef LEFSIH[a], Marco SCARA[b], Brunella BONACORSO[b], Khodir MADANI[a]

[a] *Laboratory of Biomathematics, Biophysics, Biochemistry, and Scientometry (L3BS), University of Bejaia, 06000 ,Algeria*

[b] *Department of Engineering, University of Messina, Italy*

## Abstract

The time series of monthly rainfall during 51 years of observations (from 1967 until 2018) for 24 stations in Soummam watershed of Algeria were analyzed for trends and aridity state of the area. The choice of the Weibull distribution law was justified by comparing the fitting of different probability distributions law used in literature reviews. This paper proposed a new imputation algorithm to fill missing climate data, based on the optimization of some regression methods, which are hot deck, k-nearest-neighbors imputation, weighted k-nearest-neighbors imputation, multiple imputation, linear regression and simple average method. The choice of these methods was justified by qualitative and quantitative statistical tests analysis. However, the reliability of obtained results depends mainly on percentage of missing data, choice of neighboring stations and data missingness mechanism, which should be missing at random.

**Keywords:** Missing data , Modeling climate change, Weibull law.

## Introduction

The aim of this work was to study the behavior of climatic variability in Soummam watershed (Algeria), using an adequate distribution law on a monthly rainfall measurements series of 51 years, in order to describe the space-time assessment of the climate over the entire surface of the watershed. The paper is organized in the following way, in addition to this introduction:

- The validation of climatic data availability obtained by applying our new approach that uses hybrid methods to solve problem of reliability about filling gaps results.

- A comparison between (Weibull, Gumbel land Gamma) laws which could be applicable in our case of study, according to literature review (Boudrissa et al., 2017; Husak et al., 2007). The results of density function and survival regression curve of Weibull are shown in section 2. De Martonne aridity index was used to express the change of bioclimatic watershed levels during the period of study. All of that are presented in Section 2, then the results of the work are shown in Section3.finally some conclusions are given in Section 4.

## 1. Study Area

Soummam watershed is one of the 17 major hydrological watershed of Algeria, whose an identification number of 15 according to the classification obtained by the National Hydraulic Resources Agency (Zouggaghe and Moali, 2009), the watershed located in the northeast of the country between (3.60 °, 5.55 °) of longitude and (35.75 ° and 36.75 °) of latitude. It has a very irregular shape, extending over an area of 9125 km2, from Hodna Mountains in the south to the Mediterranean Sea. On the north from Djurdjura Massif and the coastal chains of Bejaia (Taourirt Ighil and Toudja Mountains). The watershed border in the West is occupied by the tray of Buira city, while in the East is closed by Babors chains and the tray of Setif (Fig.1). The climate ranging between wet and continental with an extension of semi-arid conditions. Rainfall variability depends on geographic parameters; it increases with the altitude under humid winds in a W-E direction and decreases far from coastal areas (Lounaci, 2005). The occurrence of precipitation decreases from the North to the South, showed the protective role of the high relief of Djurdjura from humid NE winds (Turki et al., 2016).

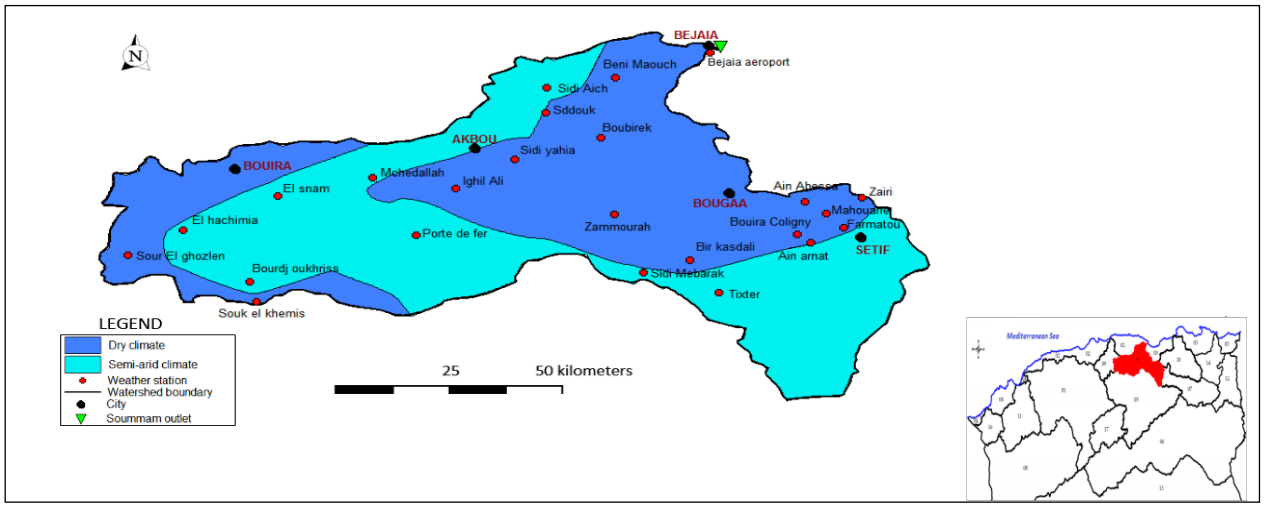



**Figure.1**. Soummam watershed bioclimatic floors map showed the meteorological stations location, followed by the watershed position on northeastern of Algerian (medallion map).

# 2. Materials and methods

## 2.1 Materials

### Data description

The study of climatic variability on Soummam watershed (Northeastern Algeria) were applied on a monthly time series of rainfall and temperature parameters from January 1967 to December 2018 over 24 climatic stations, which are positioned on the whole surface of the watershed (Fig.1). The data were obtained from National Water Resources Agency (A.N.R.H) and National Centers for Environmental Information (NCEI-NOAA), https://www.ncdc.noaa.gov/, however the availability of data is very limited since the difficulties of government in Algeria during the period of study. The Table.1 showed that the rainfall missing data observed over the 24 stations during 51 years had an interval range of (1.1%, 14.2%), whereas the temperature values are less frequent, which had an interval range of (0.2%, 4.1%).

### Determining bioclimatic floors of Soummam watershed

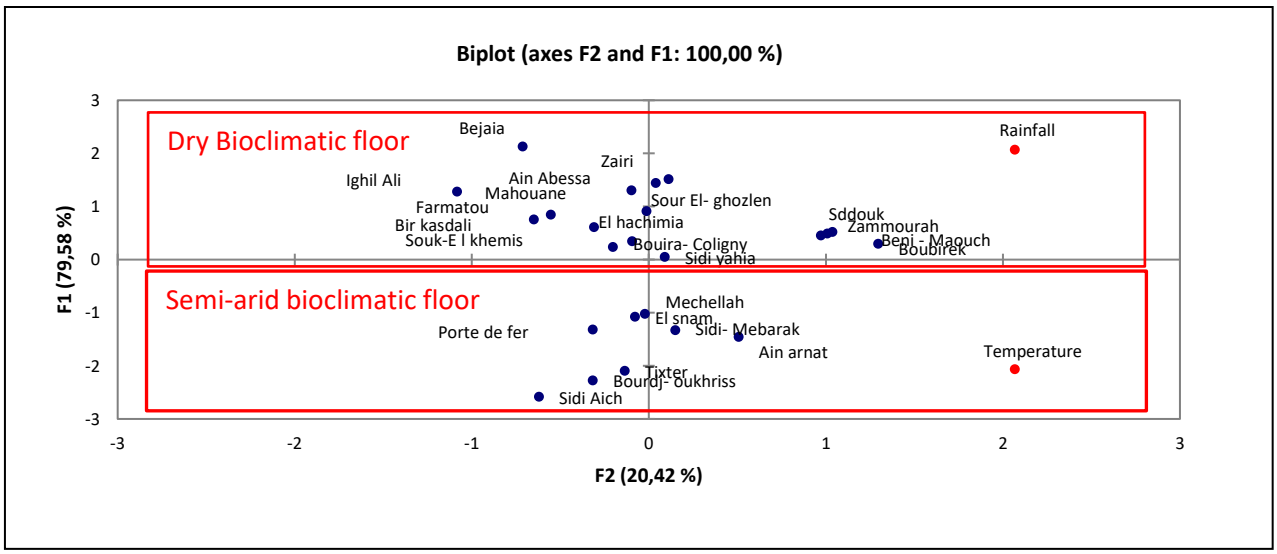



**Figure.2.** Principal Component Analysis (PCA) score plot show the bioclimatic floors of the Soummam watershed of Algeria, used inter-annual (rainfall and temperature) of 24 weather stations during 51 years

According to (Martínez,1980), the graph shows the existence of two bioclimatic floors, which are the dry and the semi-arid, following to the obtained intervals, which are respectively [222.7mm, 350mm[ and [350mm, 502.4mm]. Knowing that the dry bioclimatic floor contains (Sour El ghozlen, El hachimia, Souk el khemis, Ighil Ali, Bejaia airport, Bouira-Coligny, Farmatou, Mahouane, Zairi, Boubirek, Ain Abessa, Bir kasdali, Zammourah, Sidi yahia, Beni Maouch, Sddouk), on the other hand the semi-arid floor includes (Bourdj oukhriss, Ain arnat, El snam, Mechellah, iron gate, Tixter, Sidi Mebarak, Sidi Aich).

## 2.2 Methods

### 2.2.1 Missing climate data analysis

Algorithm description

The various methods used to fill the missing climate data series showed by (Aieb et al., 2019) were summarized in the flowchart of the new algorithm (Fig. 3). Our new approach was applied on space time continuous data (i.e. rainfall time series referred to different monitoring stations belonging to the same network). The data was ranged in matrices with the following structure: each matrix represents one year of each climatologically station, and each column is uniquely associated to month.

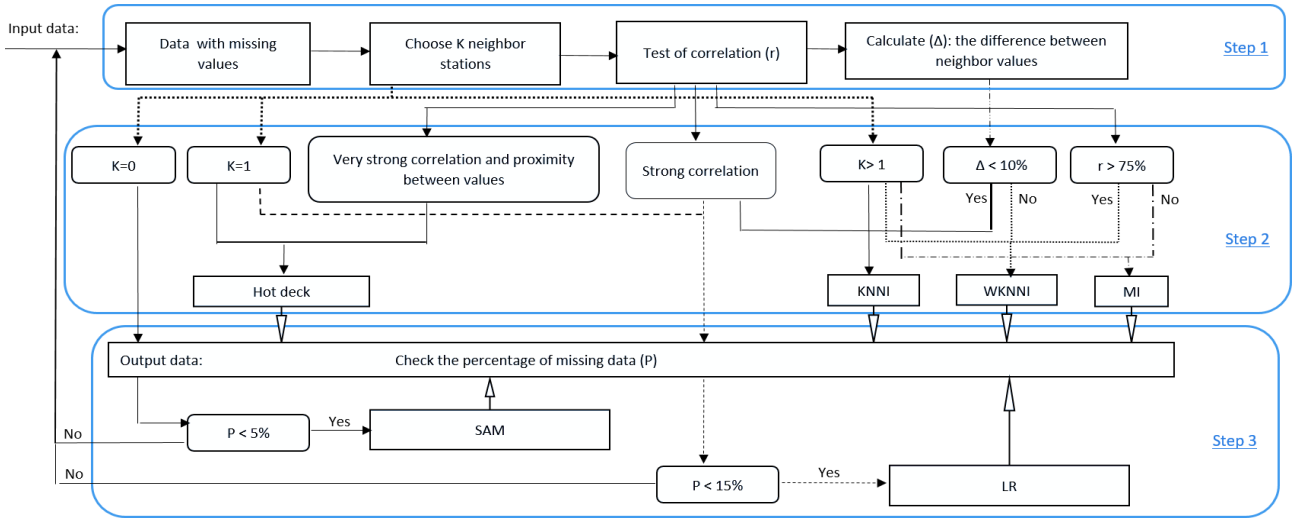



**Figure. 3.** Flowchart summarizing the filling daily rainfall dataset. Pretreatment of climate data bases (Step1), filling missing data with the appropriate method (Step2), checking the percentage of missing data (Step3).

### 2.2.2 Precipitation probability distribution models

**Weibull distribution law**

The Weibull distribution with two parameters, α and β denote the shape and scale parameters, respectively, it is expressed by the function (1) (Wilks, 1989).

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)\ exp\ \left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \qquad \alpha > 0, \beta > 0 \tag{1}$$

### 2.2.3 Test of goodness fitting

The $D_{KS}$ test can be applied to evaluate the compatibility between empirical and theoretical cumulative distribution function (Cdf), which are (F(x)) and (G(x)) respectively. The $D_{KS}$ statistic value is based on a maximum vertical difference of the both function (Justel et al., 1997). The $D_{KS}$ statistic parameter is:

$$D_{ks} = max|F(x) - G(x)| \tag{2}$$

The $D_{KS}$ distribution ($F_n$), which denote the Cdf of $D_{KS}$ under the null hypothesis $H_0$ that the *n* observations are independent and have Cdf (*F*), that is:

$$F_n = P[D_{ks} \leq x] \quad \text{for} \quad x \in [0,1] \tag{3}$$

The critical values of $D_{ks}$ test regarding the tested statistical distribution is rejected when the *P*-value (*P*) of statistical test is greater than the significance level of 5%. The *P*-value of the $D_{ks}$ test is:

$$P = P[D_{ks} > x] = 1 - Fn(x) \tag{4}$$

### 2.2.4 The De Martonne aridity index

It used with good results worldwide in order to identify dry/humid conditions of different regions (Adnan and Haider, 2012), this index is given by equation (5):

$$I_{DM} = \frac{P}{T+10} \quad (5)$$

Where $P$ and $T$ are the annual amount of precipitation and mean annual surface temperature in millimeter and in degree Celsius, respectively.

## 3. Results and discussion

This section highlights the important findings of this research and describes rainfall variability in Soummam watershed of Algeria; a both of graphical and statistical approaches were applied to examine trends of monthly rainfall series during 51 years of the observations.

### 3.1 Trend analysis of rainfall and temperature parameter

#### 3.1.1 Interannual precipitation analysis of the Soummam watershed bioclimatic floors

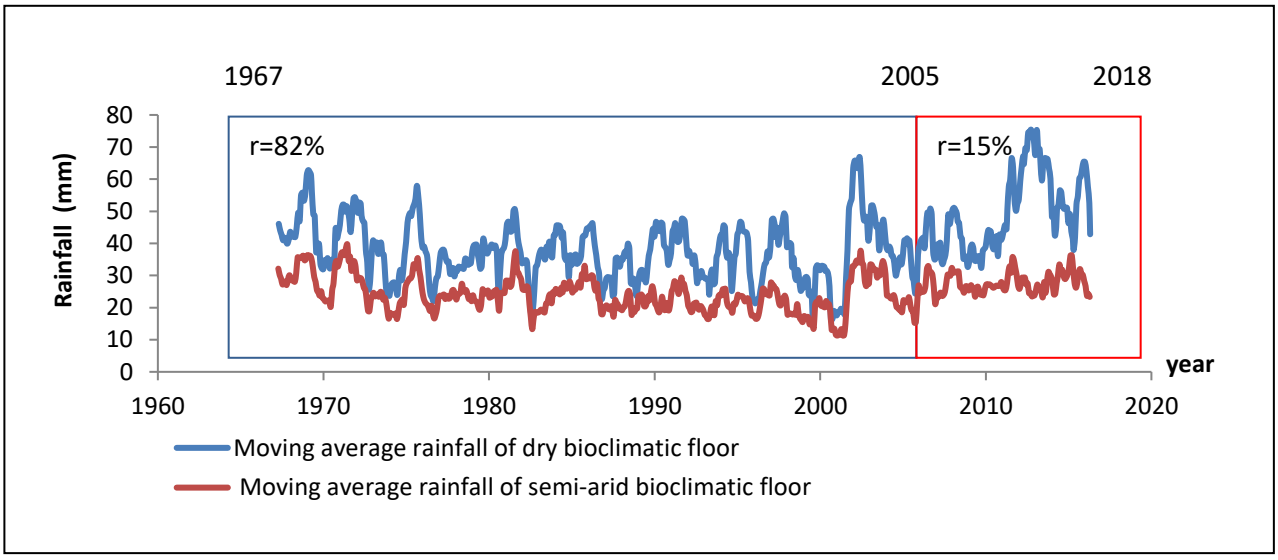



**Figure.4** Moving average curves show the monthly rainfall of dry and semi-arid bioclimatic floors during 51 years of observations at Soummam watershed.

The graph showed that in the period between (1967-2005), the rainfall trends on each floor were homogeneous, varying within a range of (3%, 60%). In this period, the rainfall in the watershed was followed by a correlation of 82%. On the other hand, between 2005 and 2018, the dry bioclimatic floor noted an increase in trend, which reached in 2004 a maximum of 72%. The

rainfall of this period had a very high variability compared to the rainfall measurement obtained on the semi-arid floor, which gave a low correlation of 15%.

## 3.2 Modeling climate variability

### 3.2.1 Model evaluation.

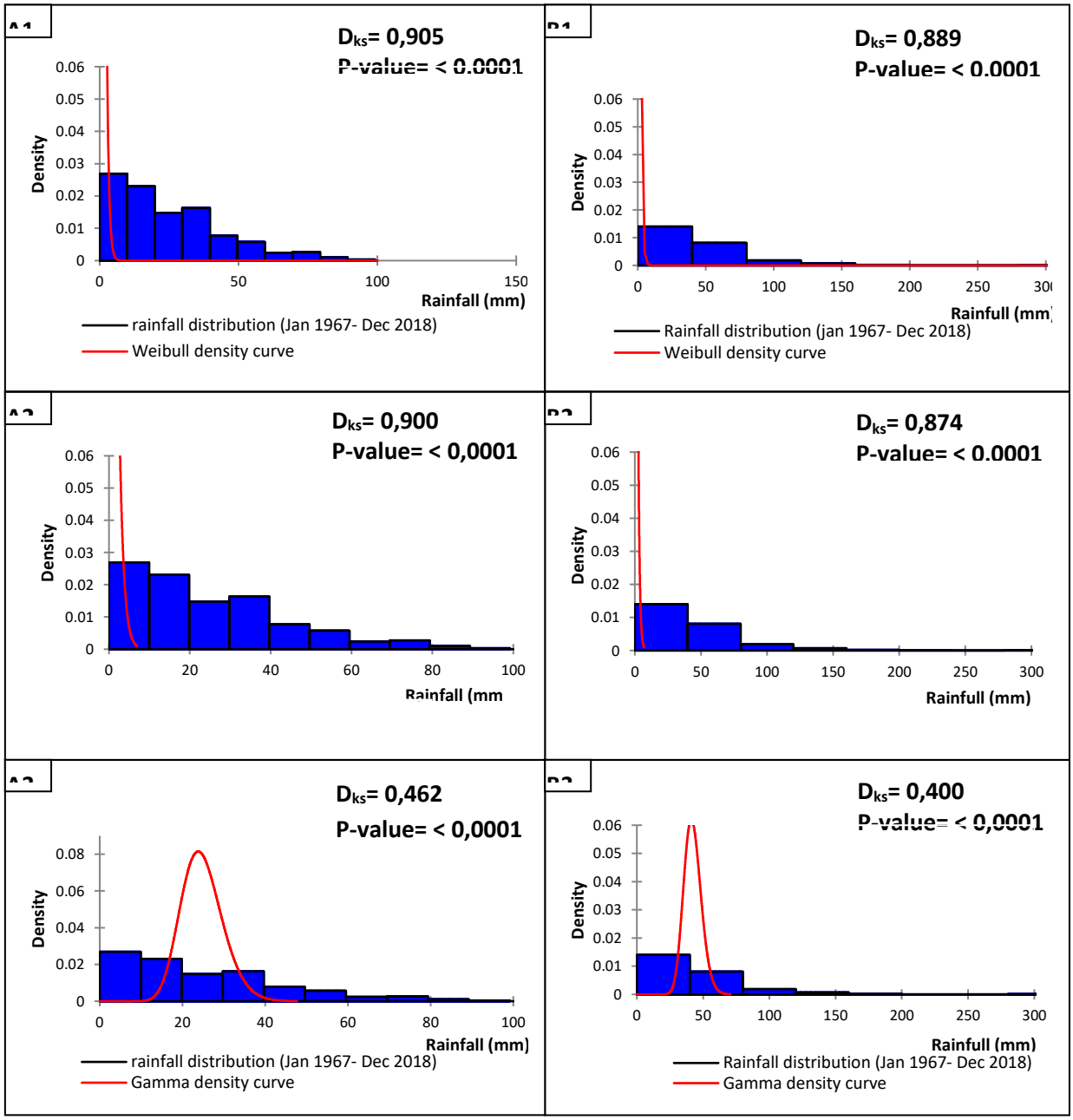



**Figure.5** Monthly rainfall distribution histogram, followed by fitting distribution curves of Weibull (1), Gumbal (2) and Gamma (3), applied for 51 years of observations at Semi-arid (A) and dry (B) bioclimatic floors of Soummam Watershed, (Dks) Kolmogorov-Smirnov.

### 3.2.2 Time modeling of Soummam watershed rainfall variability

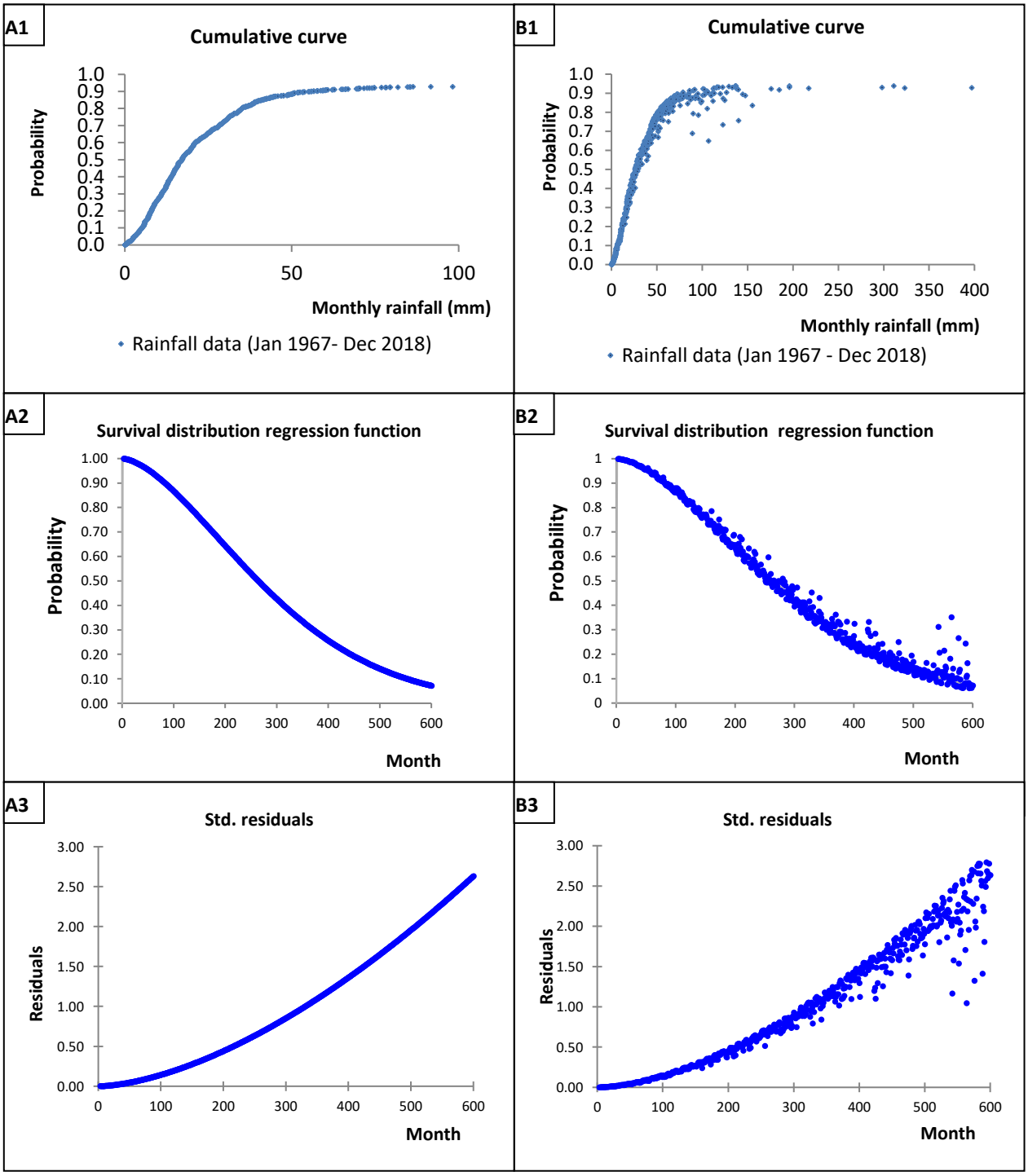



**Figure. 6** Density function, survival regression and residual curves of Weibull for modeling monthly rainfall distribution during 51 years (1967-2018) at Soummam watershed. (A) Semi-Arid bioclimatic floor, (B) Dry bioclimatic floor.

### 3.2.3 Space-time modeling of rainfall variability in Soummam dry bioclimatic floor

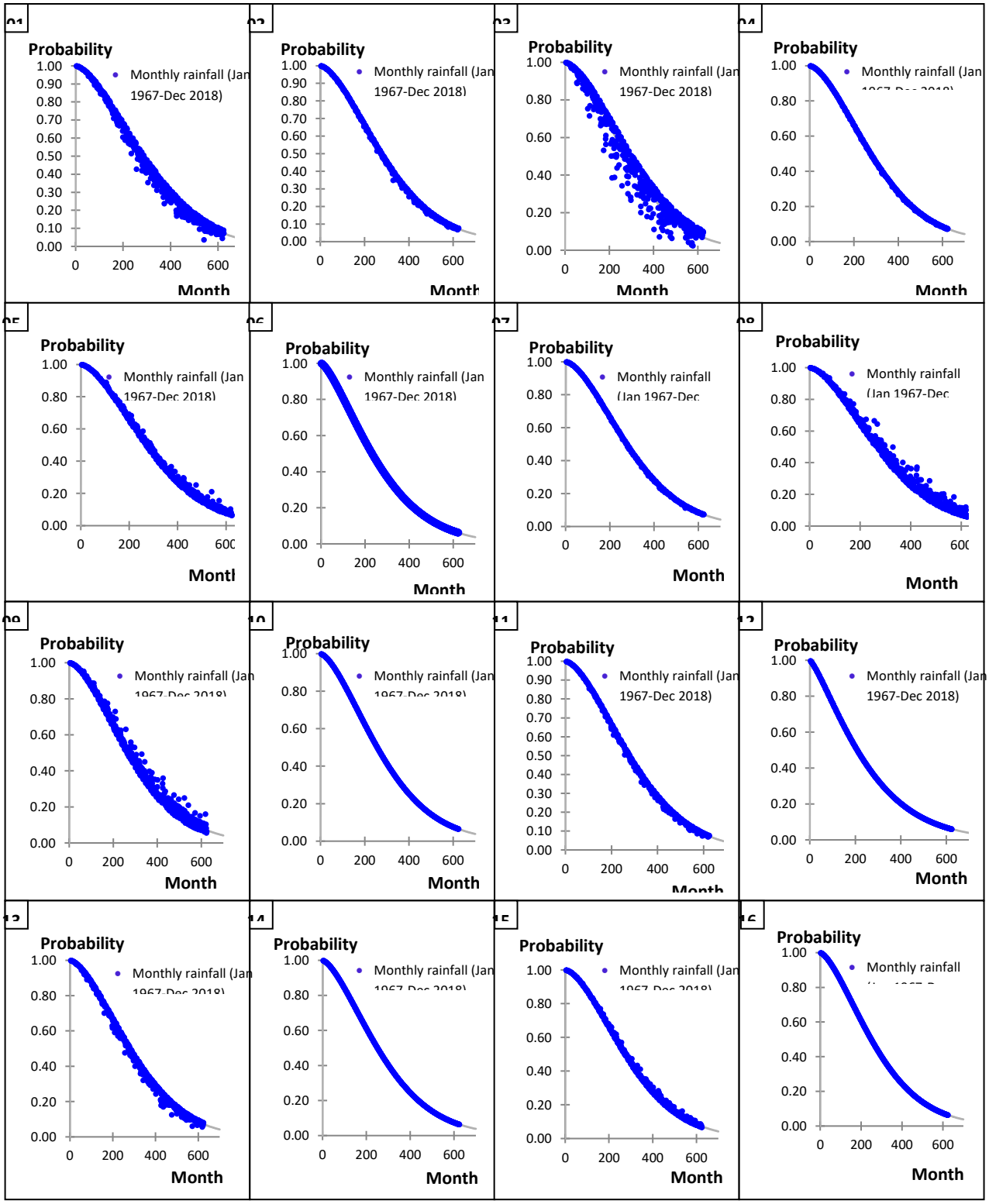



**Figure. 7** Weibull survival regression curves of dry weather stations of Soummam watershed, showed a monthly rainfall distribution of 16 stations (1-16) during 51 years (1967, 2018).

**Table. 1:** Statistical description results of space-time modeling monthly rainfall at 16 weather stations in dry bioclimatic floor of Soummam watershed during 51 years of observations.

| Number | Station | Observations | Min | Max | Mean | Std. deviation | $Chi^2$ |
|---|---|---|---|---|---|---|---|
| 01 | Sour El ghozlen | 624 | 0.000 | 230.400 | 33.149 | 27.812 | 3.833 |
| 02 | El hachimia | 624 | 0.000 | 230.300 | 31.991 | 31.226 | 0.252 |
| 03 | Souk el khemis | 624 | 0.000 | 218.400 | 32.853 | 36.862 | 20.481 |
| 04 | Ighil Ali | 624 | 0.000 | 294.500 | 35.066 | 34.946 | 0.047 |
| 05 | Bejaia airport | 624 | 0.000 | 204.000 | 29.972 | 26.615 | 0.272 |
| 06 | Bouira Coligny | 624 | 0.000 | 204.000 | 40.496 | 40.612 | 0.001 |
| 07 | Farmatou | 624 | 0.000 | 204.000 | 33.891 | 29.967 | 0.057 |
| 08 | Mahouane | 624 | 0.000 | 269.000 | 39.055 | 37.842 | 3.953 |
| 09 | Zairi | 624 | 0.000 | 237.100 | 47.389 | 42.893 | 3.839 |
| 10 | Boubirek | 624 | 0.000 | 262.700 | 46.705 | 46.824 | 0.001 |
| 11 | Ain Abessa | 624 | 0.000 | 276.800 | 47.395 | 44.684 | 0.331 |
| 12 | Bir kasdali | 624 | 0.000 | 304.700 | 34.150 | 31.170 | 0.001 |
| 13 | Zammourah | 624 | 0.000 | 271.900 | 43.581 | 46.958 | 0.428 |
| 14 | Sidi yahia | 624 | 0.000 | 204.840 | 33.063 | 33.339 | 0.001 |
| 15 | Beni Maouch | 624 | 0.000 | 223.400 | 41.218 | 42.998 | 0.526 |
| 16 | Sddouk | 624 | 0.000 | 204.100 | 41.805 | 39.453 | 0.001 |

### 3.3 The Aridity trend

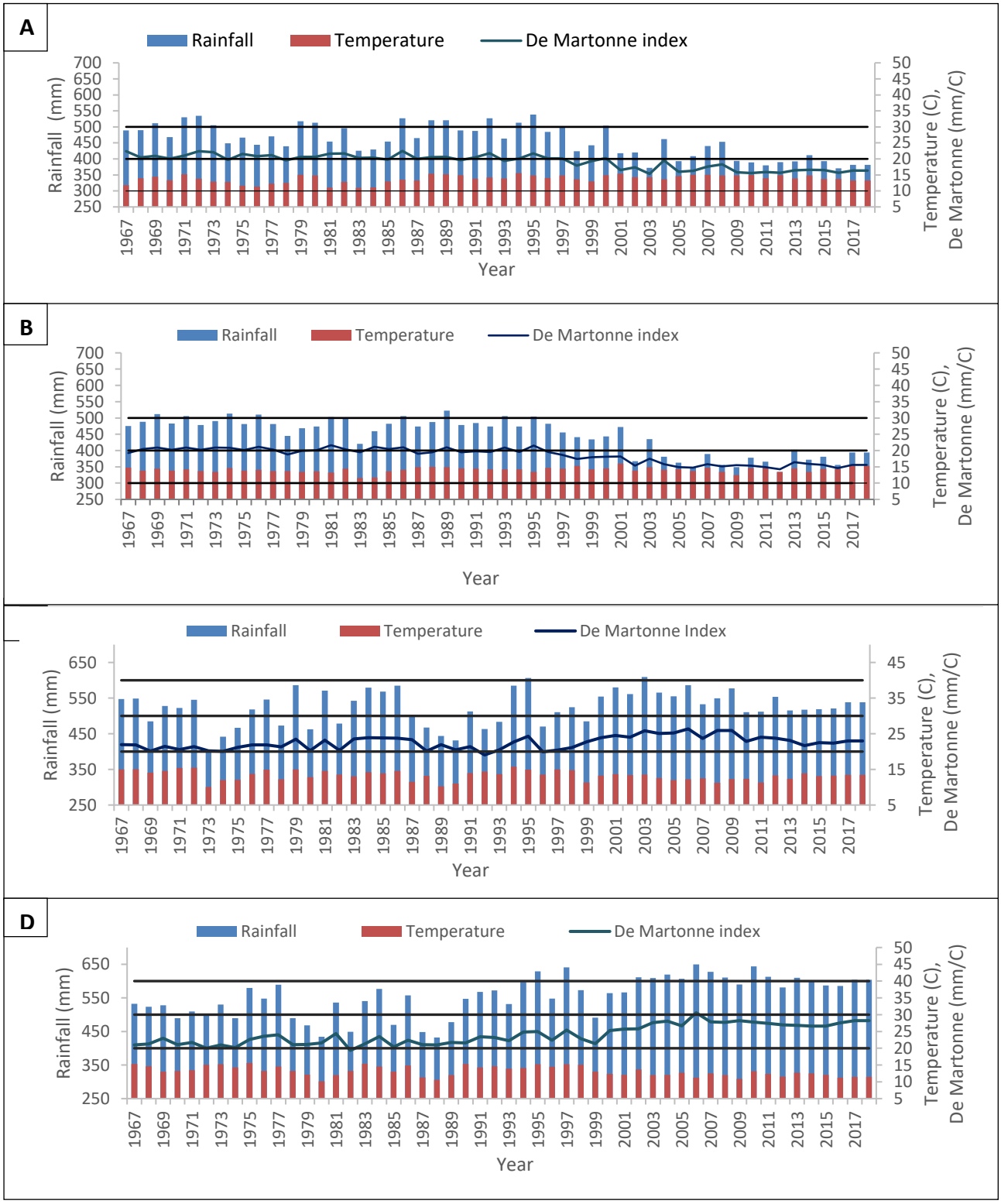



**Figure. 8** Histogram of annual rainfall and temperature data, followed by annual De Martonne curve of Sour El ghozlen (A), Souk El khemis (B), Mahouane (C) et Zairi (D) Station during 51 years (1967, 2018).

## Conclusion

- In this study, the monthly rainfall of each station of the watershed during the 51 years are modeled by using the Weibull law, according to the (Dks ) obtained results of each bioclimatic floor, which equals 0.905 and 0.889, respectively.
- The density graph and the Weibull regression curve show the space-time rainfall variability of the watershed. The breaks of the rainfall frequency are varying between (100 mm, 180 mm) between 2000 and 2018 for some dry bioclimatic stations, which are Sour elghozelen, souk el khemis, Mahouane and Zairi, respectively.
- The aridity index shows that climate change in the southwestern part of the watershed changed from a dry to semi-arid.

## References

***Adnan, S., Haider, S., Classification and assessment of aridity in Pakistan by using different aridity indices ftp://ftp. wmo. int/Documents/PublicWeb/arep/Weather_Mod_Bali/ENV% 20bruntjes. chalon, ENV. Adnan_Pakistan_paper1. pdf. Accessed 21(2012).***

***Aieb, A., Madani, K., Scarpa, M., Bonacorso, B., Lefsih, K., A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria, Heliyon 5(2019), p. e01247.***

***Boudrissa, N., Cheraitia, H., Halimi, L., 2017. Modelling maximum daily yearly rainfall in northern Algeria using generalized extreme value distributions from 1936 to 2009. Meteorological Applications, 24(1): 114-119.***

***Husak, G.J., Michaelsen, J., Funk, C., Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications, International journal of Climatology 27(2007), pp. 935-944.***

***Justel, A., Peña, D., Zamar, R., A multivariate Kolmogorov-Smirnov test of goodness of fit, Statistics & Probability Letters 35(1997), pp. 251-259.***

***Lounaci, A., Recherche sur la faunistique, l'écologie et la biogéographie des macroinvertébrés des cours d'eau de Kabylie (Tizi-Ouzou, Algérie). Thèse de doctorat d'état en biologie. Université Mouloud Mammeri de Tizi … (2005).***

***Martínez, S.R., Les étages bioclimatiques de la végétation de la Péninsule Ibérique. Anales del Jardín Botánico de Madrid, Real Jardín Botánico (1980), pp. 251-268.***

***Turki, I. et al., Hydrological variability of the Soummam watershed (Northeastern Algeria) and the possible links to climate fluctuations, Arabian Journal of Geosciences 9(2016), p. 477.***

***Wilks, D.S., Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff, Journal of Applied meteorology 28(1989), pp. 52-58.***

***Zouggaghe, F., Moali, A., Variabilité structurelle des peuplements de macro-invertébrés benthiques dans le bassin versant de la Soummam (Algérie, Afrique du Nord), Revue d'écologie(2009).***

# CONVERSION OF GEOLOGICAL MODEL (FINE-MESH) TO DYNAMIC (COARSE-MESH) HYDROCARBON MODEL WITH THE NATURE APPROACH IN SIMULATION OF THERMAL RECOVERY IN A FRACTURED RESERVOIR

Mehdi Foroozanfar[1], Mohammad Reza Rasaei[2]

**Abstract**

Operation and proper management of reservoirs requires the prediction of reservoir performance. This prediction is generally done by computer simulation. Since simulation software is capable of generating static models with the number of millions and even billions, using simulation methods and dynamic simulation on these models is difficult and sometimes impossible, multi-scale generation is necessary. In this study, we introduce a methodology which inspired by Earth's grid to make multi-scale grid generation on a multi-phase, heterogeneous reservoir, the enhanced oil recovery process is steam injection. The multi-scale grid generation that has been introduced in this study is base-on Earth's model. From a reservoir engineer's point of view Earth is equivalent to a multi-scale grid model that could be a pattern for multi-scale grid generation in a hydrocarbon reservoir to minimize CPU-Time for dynamic simulation. The principle of multi-scale grid generation in hydrocarbon reservoir is; regions with high Darcy velocities should remain fine-scale and other segments with low intensity of heterogeneity regions could resize into up-scale. Intersection of latitude and longitude lines causes Earth's discretization surface could be a practical pattern for dynamic simulation of hydrocarbon reservoirs. The interpretation of Earth's discretization is; intersection of latitude and longitude lines create segments with fine-size like South and North Poles that equal to injection and production wells in a reservoir model and other segments have verity intensity of coarse-size. Earth's magnetic field which enters from South and exits from North Pole leads to we can consider earth as a hydrocarbon model. The results of the multi-scale grid generation method which inspired by Earth's pattern were compared with the fine-mesh (Geological Model) model; these results show that the proposed method predicts close accuracy of the fine-mesh network model with less run-time.

***Keywords:*** *Multi-Phase, Simulation Speed, Multi-scale Grid Generation*

## INTRODUCTION

Interest in modeling heat and fluid transfer in formations having high permeability streaks or fractures started in 1960's with the use of thermal recovery methods for heavy oils and bitumen. Thomas [1] presented a mathematical model for conduction heating of a formation with limited permeability. He assumed that heat is introduced by a non-condensable gas through a horizontal fracture. Heat transfer from the fracture was assumed to be by vertical conduction, and heat transfer by convection was neglected. Thomas presented an example calculation for this process in an oil shale. Based on the distanced moved by the isotherm, the volume of rock heated and the oil recovered were estimated. An example calculation was presented, but because of the lack of experimental data, the model was not validated. Lesser et al. [2] formulated a similar model to represent the conduction heating of a rock with no permeability. A hot condensing gas was introduced through a horizontal fracture. The model consisted of one heat equation for the matrix and heat and fluid flow equations for the fracture.

[1] *Department of Petroleum Engineering, Kish International Campus, University of Tehran, m.foroozanfar@ut.ac.ir*

[2] *Institute of Petroleum Engineering (IPE), School of Chemical Engineering, College of Engineering, University of Tehran*

Temperature histories were obtained by finite difference solutions for both fracture and matrix. The model was applied of oil shale. They investigated heating rate effects of shale thermal diffusivity, fluid pressure in the fracture and fracture spacing. A higher injection pressure resulted in a slower heating rate. Doubling the thermal diffusivity of the formation resulted in a more rapid rise in the formation temperature. Decreasing the fracture spacing caused an important in heating rate. Again in this study an application of the model was shown for the oil shale heating by steam injection, but no temperature data were available to compare with the model. The two models described only considered conduction in the formation, convection was neglected. Wheeler [3] developed three analytical solutions to model the heat transform from the fracture to the matrix by taking into account the effects of both conduction and convection in the reservoir and heat loss to the overburden. The validity of the model was demonstrated by matching the numerical solution developed by lesser et al. [2]. Applications of these solutions were presented to determine the fracture orientation from field temperature measurements. The works described above involved mostly analytical models. Geshelin et al. [4] presented a unique numerical study on the transport of injected and reservoir water through fractures induced during steam simulation of tar sands and heavy oil deposits. Fractures created during the simulation process acted as channels through which injected fluids flowed. The heat was assumed to be transferred from the fracture to the matrix by convection and conduction. The rate of fluid transfer from the fracture to the surrounding block was a function of shape factor and the pressure difference between the fracture and the surrounding block. The shape factor was estimated by assuming a single narrow fracture instead of the double porosity assumption. The fracture model was incorporated into a conventional thermal simulator. Pruess and Narasimhan [5] presented a multiple interacting continua model (MINC) to simulate the heat and two-phase flow of steam and water in multidimensional fractured porous media. The flow domain was partitioned to computational volume elements by assuming thermodynamic equilibrium in each element. Transient flow of fluid and heat between the matrix and the fractures was treated numerically. The model was verified by comparing it the analytical solution given by Warren and Root [6]. The model was applied to different problems in geothermal reservoir engineering, such as flow to a well penetrating a fractured reservoir with low matrix permeability, boiling depletion of a fractured geothermal reservoir and production and injection in a fractured geothermal five-spot pattern.

Simulation of fluid flow with the highest accuracy and minimum time has always been the attention of experts in the oil industry. Reservoir engineers have always tried to use and learn various software and upgrade their hardware to achieve this goal. Numerical simulations of reservoirs require reservoir grid blocks with the same characteristics. Since the simulator software is based on the separation of analytical fluid flow equations on the reservoir network, increasing the blocks of this network will increase the number of unknowns and thus increase the time. On the other hand, increasing the simulation accuracy requires the use of models with the number of millions and even billions of blocks. However, the use of this number of blocks is virtually impossible due to computer speed and memory limitation. Limiting the simulation with these networks is due to two reasons: Firstly, loading the software with this number of blocks requires a lot of memory, and secondly, solving the pressure equations, which is the result of the algebra of differential fluid flow equations, requires the use of robust and parallel processors that the procurement of such equipment due to high prices and it's not possible for most companies to be available. In addition to hardware advancements in the field of software development, effective steps have been taken to address the various methods of scale up with different networking as well as different resolution algorithms. In the meantime, the use of a network with the number of optimal blocks is one of the most commonly used methods. Such a network is achieved by increasing the scale of the geological model. The choice of a scale increase method is considered an integral part of the simulation of the reservoir. The best methods are methods that first compute the value of the approximate block so that it behaves in the same way as the fine-mesh and, secondly, does not cause excessive homogeneity of the primary network. In the first case, we can use scaling methods that act on the basis of flow concepts. In the second case, the use of a non-uniform method that the coarse size is determined by the level of homogeneity of the reservoir segment. Because

geologists do not require processing speed so much, create models with a lot of blocks, reservoir engineers due to the accuracy required and their computing equipment scale-up the geological model. Scaled incremental data can never replace the original values and also upscaled blocks may don't show a similar trend in compare with the fine-mesh blocks therefore, in order to maintain vital information in the scale increment section, it is necessary to increase the scale in less heterogeneous regions.

Considerable improvement was made when Durlofsky et al. [7] offered a procedure whereby fine-mesh resolution is used in the segments of high-Darcy velocities, and coarse-mesh description is used for the rest of the domain. In their approach no upscaling layout is used for the relative permeability, as the main rock curves are utilized for the up-scaled grid blocks, hence making the technique process-independent. Pseudo-function generation for the first time presented by Chappelear and Hirasaki [8] and King et al. [10] for relative permeability upscaling in multi-phase models. Durlofsky et al. [9] constructed a procedure for calculating the transmissibility for single-phase model based on the solution to the local well-driven flow. Streamline simulation, in order to recognize the places of the grid blocks in the fine-mesh model through which most of the fluids pass, were also presented by Verma and Aziz [11] and Castellini et al. [12]. Ebrahimi and Sahimi [13] presented a multiresolution approach that performs as an automatic grid generator at all the relevant length scales that are incorporated in the geological model. They illustrated the accuracy and efficiency of the method by using it to simulations of unstable miscible displacements and also they extended their procedure for a three dimensional reservoir.

## MULTI-PHASE FLOW UPSCALING

In the case of single-phase flow, the most important parameter to increase the scale is absolute permeability, but when multiphase flow takes place, in addition to calculating the absolute permeability, it is necessary that to scale up relative permeability of each phase. The use of pseudo-function is perhaps the most commonly used method for increasing the scale for multiphase flow functions [14]. On a large scale, the physical mechanisms that occur during a multiphase flow in the reservoir are such as gravity due to the difference in the density between phases, the capillary force due to the competition between the phases and finally viscose force. Depending on which of these three forces is overcome, the results of the scale-up method in a given reservoir can lead to different curves for relative permeability and the capillary pressure [15]. The primary complexity for multi-phase upscaling is saturation distribution for fine-mesh and correspond to coarse cell therefore need to define some pseudo-functions to determined rock and fluid properties in coarse blocks. The pseudo functions may be generated in various ways; (1) analytically based on details of the reservoir properties and the size and shape of the grid blocks. (2) By using a program to process the flows, pressures, saturations etc. from a fine grid model to generate dynamic functions. (3) A vertical equilibration simulation run generates pseudo relative permeability internally. In this research we use the second method to redefine relative permeability of phases in the reservoir for coarse-size block. The primary term for relative permeability upscaling is definition of phase saturation for coarse-size block that is defined by equation 4, relative permeability upscaling (eq.14) affected by phase viscosity (eq.11), phase formation volume factor (eq.12), phase fractional flow (eq.6), phase pressure (eq.7), phase density (eq.10) and cell center depth (eq.5); all the mentioned terms need to be transformed from fine-size cells into coarse-size cells. After computing the local pore volume, replaced the pore volume of the host cell with summation of pore volumes of the fine cells (eq.13).

$$DX_c = \frac{\sum_f DX_f}{(J_2 - J_1 + 1)(K_2 - K_1 + 1)} \quad (1)$$

$$DY_c = \frac{\sum_f DY_f}{(I_2 - I_1 + 1)(K_2 - K_1 + 1)} \quad (2)$$

$$DZ_c = \frac{\sum_f DZ_f}{(I_2 - I_1 + 1)(J_2 - J_1 + 1)} \quad (3)$$

$$S_{pN} = \left(\frac{\sum_{n\epsilon N} v_n S_{pn}}{V_N}\right) \quad (4)$$

$$D_N = (\sum_{n\epsilon N} v_n d_n)/V_N \quad (5)$$

$$F_{pI} = \sum_{ijk} f_{pijk} \quad (6)$$

$$P_{pN} = \frac{\sum_{n\epsilon N} w_n p_n^{center}}{\sum_{n\epsilon N} w_n} \quad (7)$$

$$p_n^{center} = p_n + g\rho_n(D_N - d_n) \quad (8)$$

$$w_n = \left(k_{pn}.k_n.dy_n.dz_n\right)/(dx_n) \quad (9)$$

$$\rho_{pN} = (\sum_{n\epsilon N} v_n \rho_{pn})/V_N \quad (10)$$

$$\mu_{pN} = (\sum_{n\epsilon N} v_n \mu_{pn})/V_N \quad (11)$$

$$B_{pN} = (\sum_{n\epsilon N} v_n B_{pn})/V_N \quad (12)$$

$$V_N = \sum_{n\epsilon N} v_n \quad (13)$$

$$k_{pN}^d = \frac{F_{pN}^d.\mu_{pN}.B_{pN}}{T_{pN}^d}\left[P_{pN} - P_{pM} + \frac{1}{2}\left(\rho_{pN} + \rho_{pM}\right)(D_N - D_M)\right] \quad (14)$$

Where

$S_{pn}, S_{pN}$: Fine and coarse grid saturations for phase p

$v_n, V_N$: Fine and coarse grid volumes

$d_n, D_N$: Fine and coarse grid cell center depths

$f_{pN}^d, F_{pN}^d$: Fine and coarse grid phase flows in the positive d direction

g: Newton's constant

$k_{pn}, k_n$: Phase relative permeability values for fine cell and rock permeability value respectively

dx, dy, dz: Dimensions of a cell in the x, y and z directions

$\rho_{pn}, \rho_{pN}$: Fine and coarse grid phase densities

$\mu_{pn}, \mu_{pN}$: Fine and coarse grid phase viscosities

$B_{pn}, B_{pN}$: Fine and coarse grid formation volume factors

DX, DY, DZ: Coarse cell definition in the x, y and z direction

I, J, K: Indices for x, y and z direction

$F_{pI}, f_{pijk}$ : Fine and coarse Fractional flow for phase p in I, j and k direction

$P_{pN}$: Coarse pressure for phase p

$p_n$: Fine pressure

$\rho_n$: Fine density

$k_{pn}$: Fine relative permeability for phase p

$k_n$: Fine absolute permeability for rock

$P_{pN}, P_{PM}$: Coarse pressure for phase p in two adjacent coarse cells

$\rho_{pN}, \rho_{pM}$: Coarse density for phase p in two adjacent coarse cells

$D_N, D_M$: Coarse grid cell center depths for two adjacent coarse cells

$T_{pN}^{d}$: Coarse transmissibility for phase p in direction d

c: Coarse mesh

f: Fine mesh

## MULTI-SCALE GRID GENERATION OF HYDROCARBON RESERVOIRS INSPIRED BY EARTH'S GRID

The coordinates on the Earth are determined by latitude and longitude, the longitude indicating the coordinates based on its distance from the meridian, and the latitude pointing the coordinates of the location based on its distance from the equator. The Earth is a magnetism field, and Earth's magnetic field can be represented by lines that come out of the North Pole and come back to the South Pole (figure 1). From a reservoir engineer's point of view, the Earth could be considered equivalent to a hydrocarbon reservoir, in that the North and South Pole are respectively equivalent to the production and injection wells in a reservoir, and the Earth's magnetic field, equivalent to the flowlines in the reservoir and the latitude and longitude of the Earth are respectively equivalent to the discretization along the X and Y reservoirs. Earth is in fact a multi-scale grid reservoir (Figure 2).

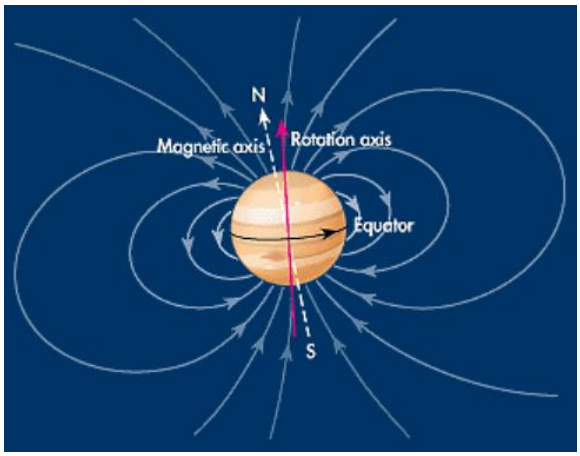



Figure 1. Earth's Magnetic Field

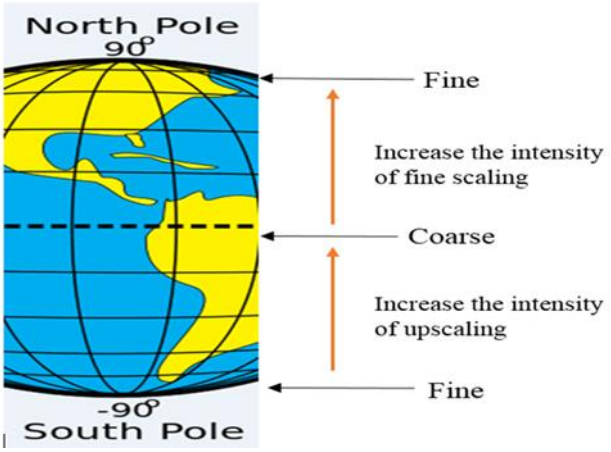



Figure 2. Earth's Grid Block

Earth's grid pattern can be an effective model for building a multi-dimensional grid scale of hydrocarbon reservoirs to reduce simulation time. To implement this kind of gridding in the model requires programming in the MATLAB software that creates the mentioned pattern in figure 2 based on the location of the production and injection wells and the distance between the two wells, which are the significant computational reservoir points. The schematic schema derived from this code is as follows (figure 3).

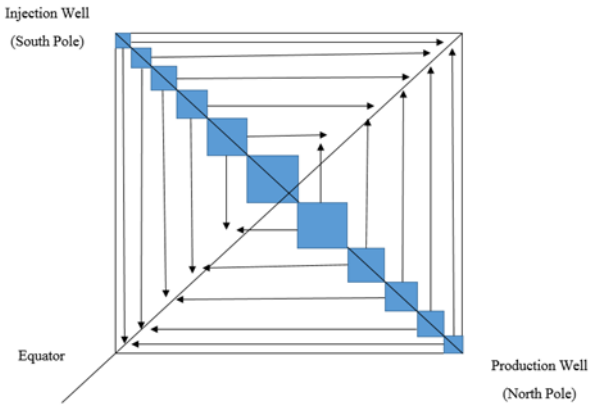



Figure 3. Schematic Schema Inspired by Earth's Grid

## FINE-MESH AND COARSE-MESH (UPSCALED) OF FRACTURED MODEL

Flow processes occur in a 3D heterogeneous reservoir (figure 4) with a fine-mesh composed of [40 × 40 × 3](total cell number: 4,800).The permeability field is a distribution of low and high values, 5 and 100 milidarcy respectively. The porosity is constant and equal to 0.3 and initial reservoir pressure is 2,900 psi. Initial reservoir temperature is 125°F and fracture permeability is 300 milidarcy.

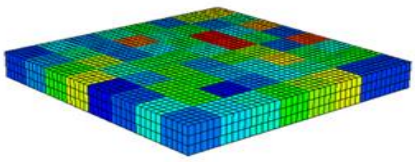


Figure 4. Fine-Mesh View of Studied Model

The upscaling level is determined by the following formula:

$$\epsilon = \frac{Upscaled\ Grid\ Block\ Numbers}{Fine\ Scale\ Grid\ Block\ Numbers}$$

$$0 < \epsilon < 1$$

The boundary of upscaling level is between zero and one for instance, number of blocks of fine-mesh model is 4800 if we minimize just one block from fine-mesh model (4799) the amount of this term is 0.99 and on the other hand if we minimize 4799 to remain just one block the amount is $2.08 \times 10^{-4}$. In this case study $\epsilon = 0.11$.

Figures 5 and 6 represent up-scaled view of studied reservoir with the feature of wells and fracture location which need to be remained fine-size in dynamic model in order to minimize numerical dispersion and the other sectors which are far from the wells and fracture could experience different intensity of coarse-size block according to Earth's pattern that is represented in figure 7.

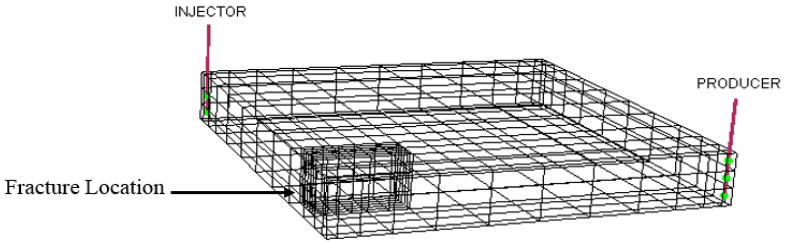



Figure 5. Coarse-Mesh View of Studied Model

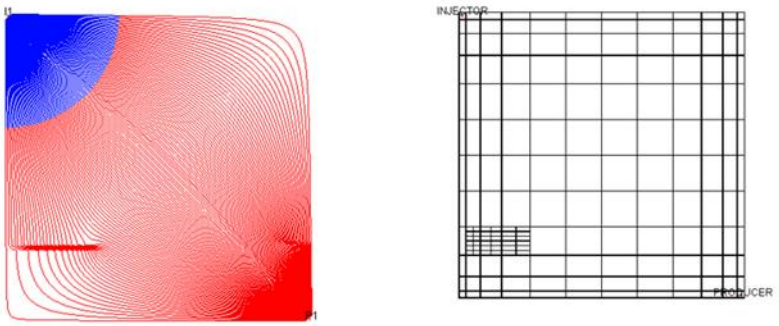



Figure 6. Coarse-Mesh View (Top View) of Studied Model

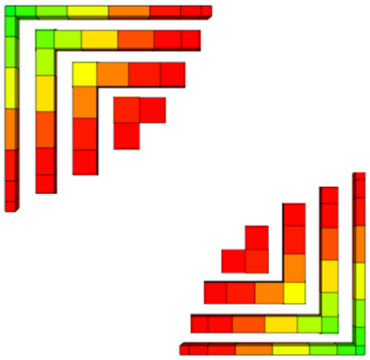


Figure 7. Discretized View of Coarse-Mesh (Up-scaled) which Inspired by Earth's Grid

The dynamic model which designed in this article has one production (sink) and injection (source) well which perforated in 3 layers, the production well controlled by oil rate target (400 STB/Day) and injection well controlled by surface flow rate target (300 STB/Day), the steam quality and its temperature are 0.7 (Dimensionless) and 450°F, respectively.

A typical unit of steam generator is shown in figure 8, treated water is diverted to the generator at the pressure necessary to inject steam into the wells by the feed pump, which usually operates at a constant rate. Water is bypassed to the upstream side of the pump if the wells take less water than is fed by the pump. Raising the temperature to the specified value when it enters the top part of a heat exchanger. Preheating of the water is not necessary where the heat exchanger in the convection system is designed to be contacted directly by cold water. In the convection section, heat present in the hot flue gas is used to further increase the temperature of the feed water. The preheated water next enters the radiant section of the steam generator, in which the arrangement of the heat exchange tubes are designed to maximize the enthalpy of the steam and reduce transverse heat loses.

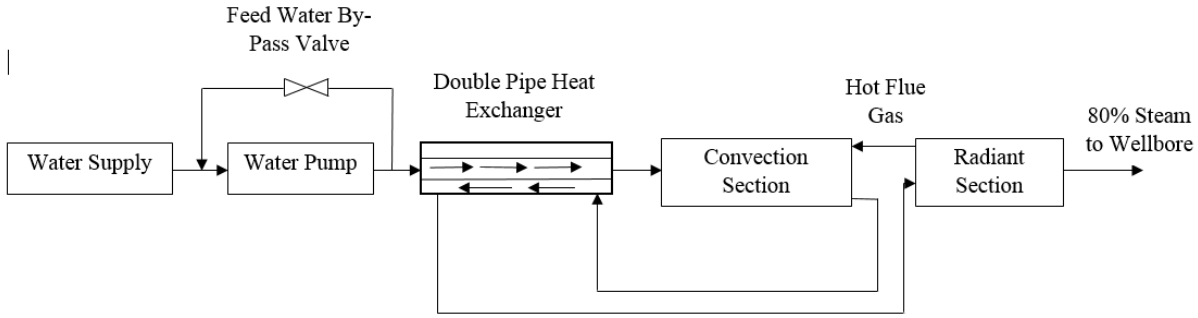



Figure 8. Steam Production System for an Oil Field

Figure 9 represents hierarchy of variable solutions such as; pressure, Darcy velocity, saturation and gravity segregation for coarse-size (global) grid.

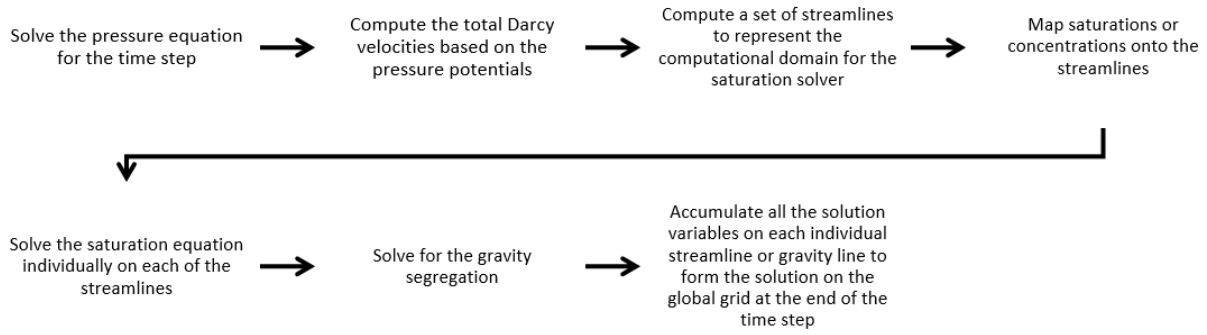



Figure 9. Algorithm of Equations

## RESULTS AND DISCUSSION

At this stage, the model is compared in terms of simulation time, energy percentage in water, oil and rock before and after steam injection, to measure the accuracy and speed of simulation. The injection pattern in this study is diagonal and the purpose of steam injection is to increase the production of oil in heavy oil reservoirs. The simulation time is 2,990 days. In order to measure energy percentage in water, oil and rock we need to evaluate the manner of water and oil because of temperature difference between steam injection temperature and reservoir temperature. Therefore, we simulated three reservoirs in three different temperature; 200°F, 250°F and 300°F and steam injection temperature is constant 450°F . Figures 10 and 11 present the higher temperature difference increases the amount of water saturation and oil in vapor phase in the reservoir. The reason for increasing the amount of water saturation is that steam with temperature 450°F enters a reservoir with lower temperature; temperature difference leads to steam condensation, and high temperature of steam causes the evaporation of oil components. Oil saturation is reduced due to production and evaporation therefore, a large part of the thermal energy is allocated to water. Energy trend in oil is declining due to decrease in oil saturation and small portion of thermal energy is allocated to the reservoir rock. In this simulation study the temperature difference is 325 °F .

In aspect of accuracy and simulation speed of up-scaled model which inspired by Earth's pattern the results are as follow:

Figure 12 shows the simulation time, the simulation time at the end of the period for the fine-mesh is equal to 71.65 seconds and for the coarse-mesh model is 9.062 seconds. Figure 13 presents energy percentage in water, before steam injection energy percentage in fine-mesh and coarse mesh is 62.42% and after steam injection for fine and coarse mesh is 64.62% and 64.52% respectively, the percentage error is 0.15%. Figure 14 presents energy percentage in oil, before steam injection energy percentage in oil for fine-mesh and coarse-mesh is 36.72% and after steam injection for fine and coarse mesh is 34.08% and 34.18% respectively, the percentage error is 0.29% and figure 15 illustrates energy percentage in rock, the amount before injection is 0.85% and after injection for fine-mesh and coarse mesh is 1.14% and 1.13% respectively, the percentage error is 0.88%. This method works in such a way that firstly, Darcy velocity of fine-mesh is calculated to identify the points affecting the flow, such as the production and injection wells, and the existence of high-permeability heterogeneity. The highest percentage of increase in scale is related to the areas that have the greatest distance from production and injection wells and the pore volume of up-scaled grid is equal to the total volume of the cavities forming it and the solution of the saturation and pressure equation of the up-scaled block is based on the information of its constituent blocks, which can be achieved higher accuracy and speed.

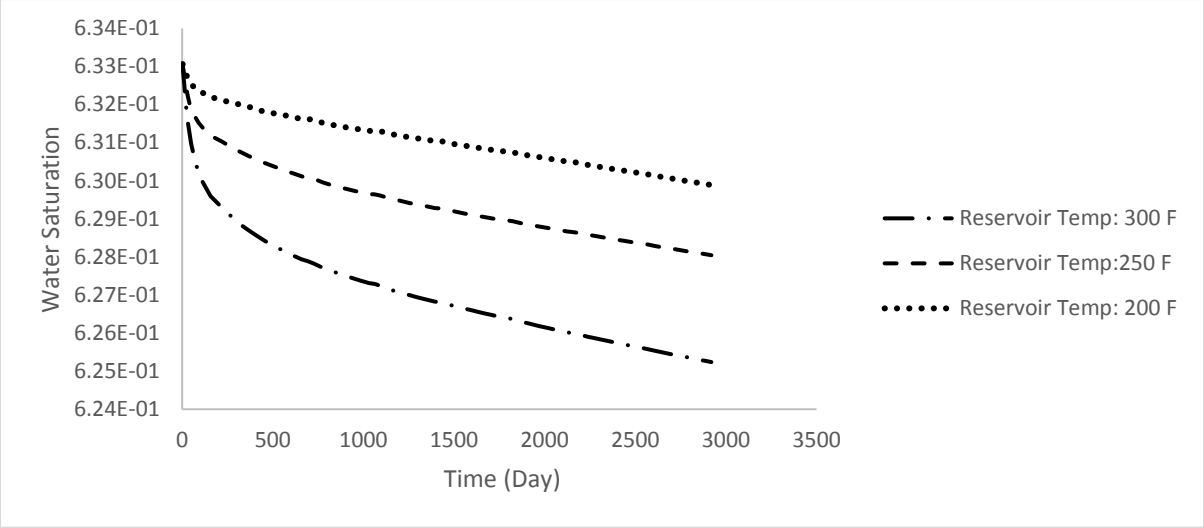



Figure 10. Effect of Temperature Difference between Reservoir Temperature and Steam Injection Temperature on Water Saturation in the Reservoir

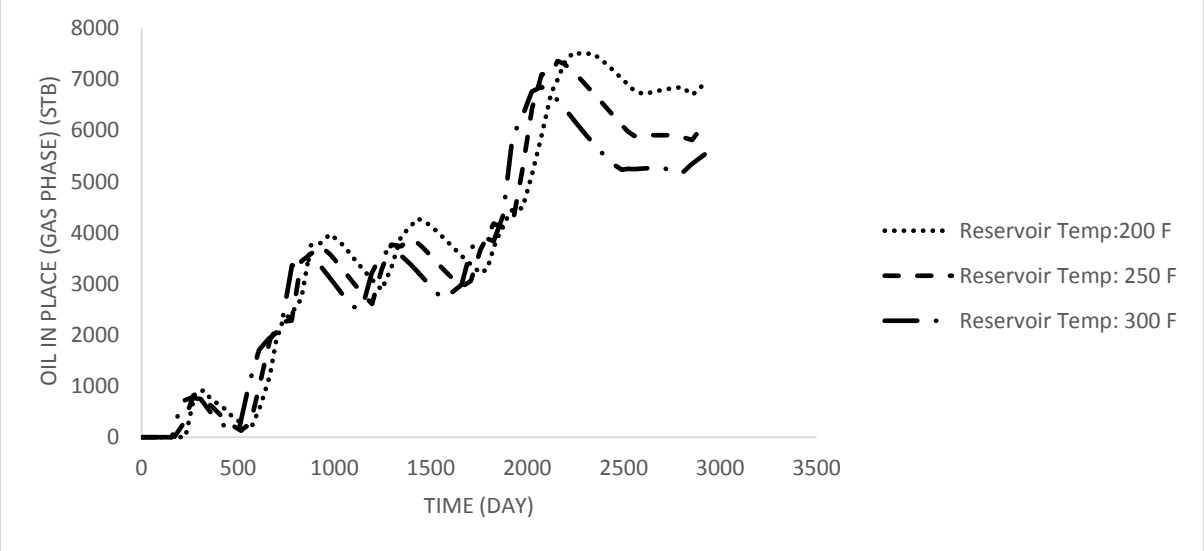



Figure 11. Effect of Temperature Difference between Reservoir Temperature and Steam Injection Temperature on Oil in Place (Gas Phase) in the Reservoir

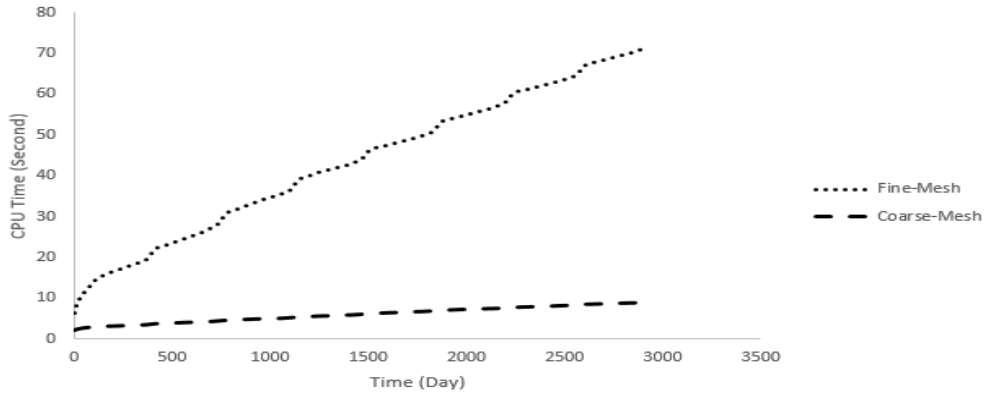



Figure 12. CPU Time

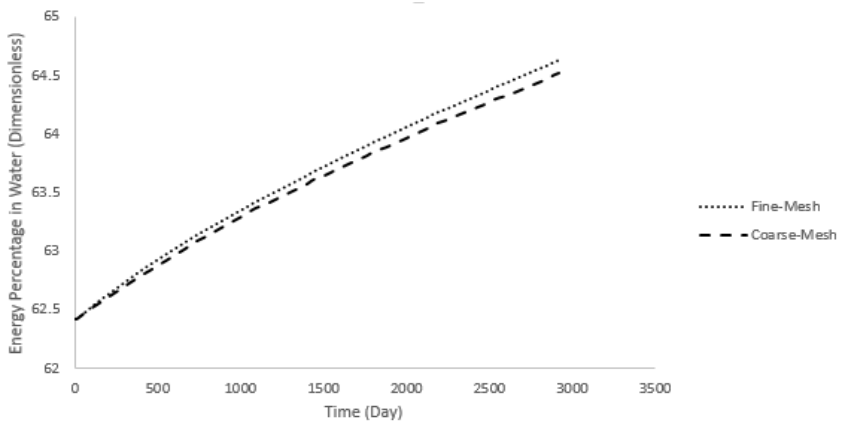



Figure 13. Energy Percentage in Water

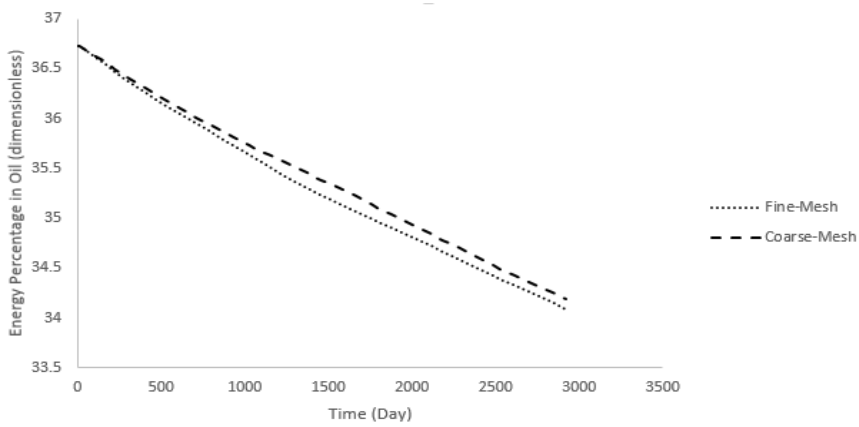



Figure 14. Energy Percentage in Oil

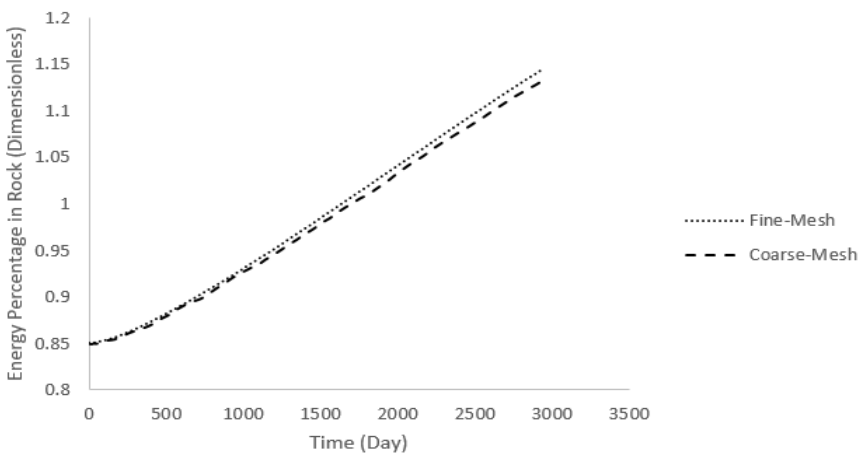



Figure 15. Energy Percentage in Rock

## CONCLUSION

According to the results obtained in the previous section of this study, the accuracy and speed of the introduced method can be found. The speed up factor in coarse-mesh in compared with fine-mesh is 7.9 times faster, and in terms of accuracy, the results obtained in the previous section are as good as

the low error compared to the fine-mesh model. There are many basic and practical issues that come with a multi-level solution like composite materials, porous media and flow with high Reynolds numbers, solving these problems even with the use of supercomputers is very difficult due to its computational cost. The method proposed in this study can be a way to reduce the computational cost of complex issues.

## REFERENCE


[1]. Thomas, G.W.: "A Simplified Model of Conduction Heating in Systems of Limited Permeability," Soc. of Pet. Eng. J. (December 1964) 335-344.

[2]. Lesser, H.A., Bruce, G.H. and Stone, H.L.: "Conduction Heating of Formations with Limited Permeability by Condensing Gases." Soc. of Pet. Eng. J. (December 1966) 372-382.

[3]. Wheeler, J.A.: "Analytical Calculations for Heat Transfer from Fractures," No. SPE 2494 (1969).

[4]. Geshelin, B.M., Grabowski, J.W. and Pease, E.C.: "Numerical Study of Transport of Injected and Reservoir Water in Fractured Reservoirs During Steam Injection Stimulation." Paper SPE 10322 Presented at the 1981 SPE Annual Fall Technical Conference and Exhibition, San Antonni, TX, October 5-7.

[5]. Pruess, K. and Narasimhan,T.N.: "A Practical Method for Modeling Fluid and Heat Flow in Fractured Porous Media." Soc. of Pet. Eng. J. (February 1985) 14-26.

[6]. Warren, J.E. and Root, P.J.: "The Behavior of Naturally Fractured Reservoirs," Soc. Pet. Eng. J. (September 1963) 245-255.

[7]. Durlofsky, L.J., Jones, R.C., Milliken, W.J.: A non-uniform coarsening approach for the scale up of displacement processes in heterogeneous porous media. Adv. Water Resour (1997).

[8]. Chappelear, A., Hirasaki, G.J.: A model of oil-water coning for two dimensional areal reservoir simulation,SPE paper 4980 (1976).

[9]. Durlosfky, L.J., Milliken, W.J., Bemath, A.: Scaleup in the near-well region, SPE paper 61855 (2000).

[10]. King, P.R., Snyder, D.E., Obut, T.S., Perkins, R.L.: A case study of the full-field simulation of a reservoir containing bottomwater, SPE paper 21203 (1991).

[11]. Verma, S., Aziz, K.: Two and three dimensional flexible grids for reservoir simulation, Proceedings of the Fifth European Conference on the Mathematics of Oil Recovery, Leoben, Austria (1996).

[12]. Castellini, A., Edwards, M.G., Durlofsky, L.J.: Flow based modules for grid generation in two and three dimensions. Proceedings of Seventh European Conference on the Mathematics of Oil Recovery (2000).

[13]. Ebrahimi, F., Sahimi,M.: Multiresolutionwavelet coarsening and analysis of transport in heterogeneous porous media. Physica A **316**, 160 (2002).

[14]. Peddibhotla S, Gupta A.D, Xue, G.: Multiphase streamline modeling in three-dimensions: further generalizations and a field application, SPE paper 38003 (1997).

[15]. Sahimi M, Rasaei M.R, Ebrahimi F, Haghighi, M.: Upscaling of unstable displacements and multiphase flows using multiresolution wavelet transformation, SPE paper 93320 (2005).

# ANALYSIS OF PERIODICITIES OF COSMIC RAY TIME SERIES LOCATED AT DIFFERENT GEOMAGNETIC LOCATIONS

Jose F. Valdes and Marni Pazos
Universidad Nacional Autonoma de Mexico

## Abstract

We studied the neutron monitor data bases of Mexico City, Oulu, Finlad and Moscow, Russia from 1990 to 2017 to find periodicities in the intensity variations of the cosmic ray flux. We used the wavelet transform to identifiy mid-term variations present in the records. The corresponding confidence levels are given to the periodicities, as well as the contribution to the total power spectrum of such variations. Results are consistent with previous analysis done for other cosmic ray detectors, showing the relevance of mid-term variations, probably related to phenomena occuring below the solar atmosphere. As a reference, we compare these results with those of classical Fourier analysis based on the discrete Fourier transform, and a fractal analysis, giving consistent results. To the best of our knowledge, this is the first time that a comparative analysis of this kind is done for these three neutron monitor series representing low, medium and high geomagnetic latitudes.

# Wind-power intra-day multi-step predictions using polynomial networks solutions of general PDEs based on Operational Calculus

Ladislav Zjavka, Stanislav Mišák and Lukáš Prokop

VŠB-Technical University of Ostrava, ENET Centre, Ostrava, Czech Republic
ladislav.zjavka@vsb.cz

**Abstract.** Precise intra-day predictions of wind-power are challenging due to its intermittent nature and high correlation with large-scale atmospheric chaotic circulation processes. NWP systems solve sets of differential equations to predict a time-change of each 3D-grid cell in several atmospheric layers. Their surface forecasts of wind speed are not entirely adapted to specific local characteristics and anomalies, which largely influence its temporal-flow. AI methods using historical observations can convert and refine the daily forecasts in consideration of wind farm siting, terrain asperity and ground level (hub height). Their independent wind-power predictions in horizon of several hours are also more precise then NWP model forecasts as these are usually produced every 6 hours. The designed method uses Polynomial neural networks to decompose and substitute for the general linear Partial Differential Equation being able to describe n-variable functions of unknown complex dynamic systems. It solves specific 2-variable 2nd order PDEs, formed in PNN nodes, using a polynomial conversion based on Operational Calculus. The inverse Laplace transformation is applied to the resulting rational terms to obtain the originals of node functions whose sum gives the complete PDE model. The composite PDE models are developed with data samples from the estimated optimal numbers of training days to represent spatial data relations in current weather, necessary for applicable predictions. They can predict wind power up to 12 hours ahead according to a trained data inputs->output time-shift. Intra-day multi-step predictions using the PDE models are more precise than those based on NWP model forecasts or statistical techniques allowing using local time-series of several variables only.

**Keywords:** Polynomial Neural Network; General Partial Differential Equation; Polynomial PDE substitution of Operational Calculus; External Complement

## 1 Introduction

Direct and indirect wind power forecasting methods, which attempt to predict at first wind speed and then transform it to wind power forecasts, can be generally classified into 2 major approaches [4]:

- Physical, solving the energy and momentum equations
- Statistical, based mainly on historical data series

adfa, p. 1, 2011.

Continuous fluctuations in atmospheric circulation processes impose unstable wind speed temporal-behavior causing difficulties in forecasting using Numerical Weather Prediction (NWP) systems [3]. These solve sets of primitive equations to simulate the global atmospheric dynamics in the 3D deterministic NWP models. Statistical methods can adapt these local wind speed forecasts to the near-ground layer, considering the parametrized wind relief and profile, wind farm allocation and hub heights. Artificial Intelligence (AI) adaptive techniques are usually employed to model stochastic correlations between inputs->output weather observations and wind power measurements. Predictions of stand-alone statistical models using historical time-series are usually worthless beyond several hours, while combined solutions using in addition 24-48 hour NWP outputs can show very strong interdependence on the quality of the local forecasts. Meso-scale NWP models are usually produced each 6 hours, so that post-processing methods must wait for the forecasts [8]. Independent AI solutions are to transform local weather information into the optimal power model to minimalize average predictions errors. AI methods can select from relevant weather data to represent the current spatial relation-ships necessary for the optimal predictions in a trained time-horizon. Their predictions are usually more precise than those of NWP models in more or less settled weather though if an atmospheric front has passed through the target area the AI models do not represent changed new patterns. Each of the direct, indirect or combined approach has its own advantages, generally independent statistical models without using NWP outputs are helpful in intra-day predictions.

Composite models of an appropriate complexity can represent the complex dynamics in local weather. Standard computing techniques require data pre-processing to significantly reduce the number of input variables, which leads usually to the models over-simplification. Polynomial Neural Networks (PNN) overcome this problem using polynomial regression where the number of parameters grows exponentially along with the number of input variables. PNN decompose the general connections between inputs->output variables, expressed by the Kolmogorov-Gabor polynomial (1).

$$Y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \ldots \quad (1)$$

*n - number of input variables $x_i$* *$a_i$, $a_{ij}$, $a_{ijk}$, ... - polynomial parameters*

Group Method of Data Handling (GMDH) evolves a multi-layer PNN structure in successive steps, adding one layer at a time to calculate its node parameters and select the best ones to be applied in the next layer. PNN nodes decompose the complexity of a system into a number of simpler relationships, each described by low order polynomial transfer functions (2) for every pair of input variables $x_i$, $x_j$ [5].

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \quad (2)$$

*$x_i$, $x_j$ - input variables of polynomial neuron nodes*

Differential Polynomial Neural Network (D-PNN) is a new class of computing networks which apply adapted procedures of Operational Calculus (OC). It decomposes the general Partial Differential Equation (PDE), being able to describe unknown complex dynamic systems of n-variables [7]. D-PNN combines the best 2-inputs in

PNN nodes, analogous to GMDH but in contrast to it producing applicable PDE components to solve particular 2nd order sub-PDEs. These are OC converted into rational terms, produced in PNN nodes, which represent the Laplace images of unknown node functions. The inverse L-transformation is applied to them to obtain the node originals whose sum gives the complete PDE model of a searched separable output functions. D-PNN uses GMDH principles of External Complement to develop its models, which usually lead to the optimal representation of a problem [1]. The composite PDE models are formed for each 0.5-12 hour inputs->output time-shift of spatial data observations to predict the target power output [6]. The D-PNN models, allowing complex representation of local weather patterns, can compute more precise intra-day predictions than those based on adapted middle-term NWP wind speed forecasts or standard statistical techniques using only a few input variables in the simplified solutions [10].

## 2 Intra-day multi-step wind power prediction

D-PNN using the statistical approach need to pre-estimate the optimal numbers of the last days, whose data are used to develop its prediction PDE models. The optimal daily training periods were initially estimated by additional D-PNN assistant models according to their best approximation of power output in the last 6-hour test. Their formation is analogous to that one of prediction models but the D-PNN output is permanently compared with the reserved latest measurements of the desired power. The lowest testing errors indicate the optimal training parameters [8].

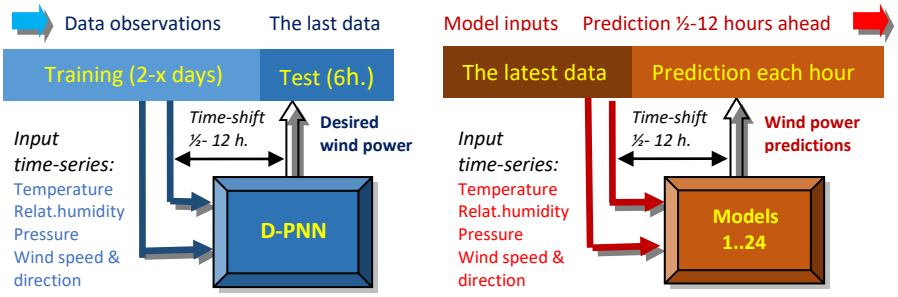



**Fig. 1.** D-PNN is trained with spatial data observations from the estimated period of the last few days for each inputs->output time-shift 0.5-12 hours (blue-left) to develop PDE models which apply the latest data inputs to predict wind power in the trained time-horizon (red-right)

D-PNN is trained with the estimated optimal number of daily data samples. The resulting PDE models apply the latest morning input time-series to predict wind power in the trained inputs->output time-shift ½-12 hours ahead. A separate model is developed for each half-hour prediction horizon (Fig.1). Periods of more or less settled weather over the last few days with similar data patterns tend to prevail till they are broken up by an overnight change in conditions [29]. Converted NWP model forecasts of wind-speed could be applied in these days as training data patterns do not correspond to those in the prediction day. The statistical model is flawed and cannot represent the current data relations [6].

## 3 A polynomial PDE substitution using Operational Calculus

D-PNN defines and substitutes for the general linear PDE (3) which can describe unknown complex dynamic systems. It decomposes the n-variable PDE into 2-variable 2nd order sub-PDEs in PNN nodes. These can be solved using OC to model unknown node functions $u_k$ whose sum gives the complete *n*-variable *u* model (3).

$$a + bu + \sum_{i=1}^{n} c_i \frac{\partial u}{\partial x_i} + \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \ldots = 0 \qquad u = \sum_{k=1}^{\infty} u_k \tag{3}$$

*$u(x_1, x_2, \ldots, x_n)$ - unknown separable function of n-input variables*
*$a, b, c_i, d_{ij}, \ldots$ - weights of terms* *$u_i$ - partial functions*

Specific 2nd order PDEs, formed in PNN nodes, can be expressed in the equality of 8 variables (4), including derivative terms formed with respect to variables corresponding to all the GMDH polynomial members (2).

$$F\left(x_1, x_2, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \frac{\partial^2 u}{\partial x_2^2}\right) = 0 \tag{4}$$

*$u_k$ - node partial sum functions of an unknown separable function u*

The polynomial conversion of 2nd order PDEs (4) using procedures of Operational Calculus is based on the proposition of the Laplace transform (L-transform) of the function $n^{th}$ derivatives in consideration of the initial conditions (5).

$$L\left\{f^{(n)}(t)\right\} = p^n F(p) - \sum_{k=1}^{n} p^{n-i} f_{0+}^{(i-1)} \qquad L\{f(t)\} = F(p) \tag{5}$$

*$f(t), f'(t), \ldots, f^{(n)}(t)$ – originals continuous in $<0+, \infty>$* *p, t - complex and real variables*

This polynomial substitution for the *f(t)* function $n^{th}$ derivatives in an Ordinary Differential Equation (ODE) leads to algebraic equations from which the L-transform image *F(p)* of an unknown function *f(t)* is separated in the form of a pure rational function (6). These fractions represent the L-transforms *F(p)*, expressed with the complex number *p*, so that the inverse L-transformation is applied to them to obtain the original functions *f(t)* of a real variable *t* (6) described by the ODE [2].

$$F(p) = \frac{P(p)}{Q(p)} = \sum_{k=1}^{n} \frac{P(\alpha_k)}{Q_k(\alpha_k)} \frac{1}{p - \alpha_k} \qquad f(t) = \sum_{k=1}^{n} \frac{P(\alpha_k)}{Q_k(\alpha_k)} e^{\alpha_k \cdot t} \tag{6}$$

*$\alpha_k$ - simple real roots of the multinomial Q(p)* *F(p) - L- transform image*

Specific 2nd order pure rational terms (6), with the lower multinomial degree in the numerator $P(x_1, x_2)$ than the denominator $Q(x_1, x_2)$, are produced in D-PNN node blocks (Fig.2) for each particular 2nd order sub-PDE (4), using the OC conversion (5). The inverse L-transformation is analogously applied to the converted sub-PDEs (6), i.e. Laplace images, to obtain the originals of node functions $u_k$ whose sum gives the model of the unknown separable output function *u* (3). Each node block uses GMDH polynomial (2) to produce its output applied in the next layer nodes. It contains 2 vectors of parameters ***a***, ***b*** to form rational functions for neurons, i.e. specific sub-PDE solutions (7), using GMDH output and reduced polynomials. One of the neurons in node blocks can be selected to be directly included in the network output sum [6].

$$y_i = w_i \frac{b_0 + b_1 x_1 + b_2 sig(x_1^2) + b_3 x_2 + b_4 sig(x_2^2)}{a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 sig(x_1^2) + a_5 sig(x_2^2)} \cdot e^{\varphi} \tag{7}$$

*$\varphi$ = arctg($x_1/x_2$) - phase representation of 2 input variables $x_1$, $x_2$*
*$a_i$, $b_i$ - polynomial parameters* *$w_i$ - weights* *sig - sigmoidal*

The inverse L-transformation of converted sub-PDEs uses complex variables $p$ (6) so that the phase of complex representation of 2-variables in Eulers's notation (8) is applied $e^{\varphi}$ (7). The pure rational fractions correspond to the amplitude $r$ (radius).

$$p = \underbrace{x_1}_{\mathrm{Re}} + i \cdot \underbrace{x_2}_{\mathrm{Im}} = \sqrt{x_1^2 + x_2^2} \cdot e^{i \cdot \arctan\left(\frac{x_2}{x_1}\right)} = r \cdot e^{i \cdot \varphi} = r \cdot (\cos\varphi + i \cdot \sin\varphi) \tag{8}$$

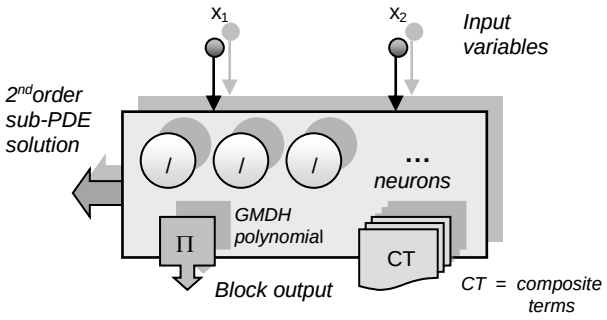



**Fig. 2.** A block of derivative neurons - 2nd order sub-PDE solutions formed in PNN nodes

## 4 PDE decomposition using backward Differential network

Composite polynomial functions are formed in the multi-layer PNN structure (9). The blocks in nodes of the 2nd and next layers can produce in addition Composite Terms (CT) which are equivalent to the simple neurons in calculation of the D-PNN output sum. CTs substitute for the sub-PDEs with respect to input variables of back-connected node blocks of the previous layers (Fig.3) using the product of their Laplace images according to the composite function partial derivation rules (10).

$$F(x_1, x_2, \ldots, x_n) = f(z_1, z_2, \ldots, z_m) = f(\phi_1(X), \phi_2(X), \ldots, \phi_m(X)) \tag{9}$$

$$\frac{\partial F}{\partial x_k} = \sum_{i=1}^{m} \frac{\partial f(z_1, z_2, \ldots, z_m)}{\partial z_i} \cdot \frac{\partial \phi_i(X)}{\partial x_k} \qquad k=1, \ldots, n \tag{10}$$

The 3rd layer blocks, for example, can select from additional CTs using products of sub-PDE converts of 2 and 4 back-connected blocks in the previous 2nd and 1st layers (11). The number of possible CTs in blocks doubles along with each joined preceding layer (Fig.3).

$$y_{31} = w_{31} \cdot \frac{b_0 + b_1 x_{21} + b_2 x_{21}^2 + b_3 x_{22} + b_4 x_{22}^2}{a_0 + a_1 x_{21} + a_2 x_{22} + a_3 x_{21} x_{22} + a_4 x_{21}^2 + a_5 x_{22}^2} \cdot \frac{b_0 + b_1 x_{12} + b_2 x_{12}^2}{a_0 + a_1 x_{11} + a_2 x_{12} + a_3 x_{11} x_{12} + a_4 x_{11}^2 + a_5 x_{12}^2} \cdot \frac{P_{12}(x_1, x_2)}{Q_{12}(x_1, x_2)} \cdot e^{\varphi_{31}}$$

*$Q_{ij}$, $P_{ij}$ = GMDH output and reduced polynomials of n and n-1th degree* (11)
*$y_{kp}$ - $p^{th}$ Composite Term (CT) output* *$\varphi_{21}$ = arctg($x_{11}/x_{13}$)* *$\varphi_{31}$ = arctg($x_{21}/x_{22}$)*
*$c_{kl}$ - complex representation of the $l^{th}$ block inputs $x_i$, $x_j$ in the $k^{th}$ layer*

The CTs are the products of the external function sub-PDE solution, i.e. the $L^{-1}$ transformed image in the starting node block, and selected neurons (i.e. the internal function images) of back-connected blocks in the previous layers (11).

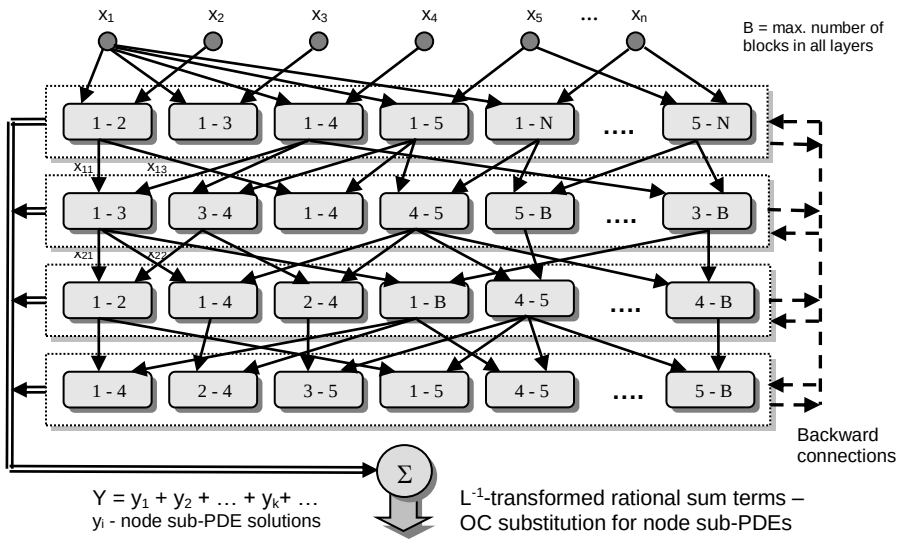



**Fig. 3** D-PNN selects from possible 2-variable combination node blocks to produce applicable sum PDE components (neurons)

The D-PNN output *Y* is the arithmetic mean of the selected active neurons + CTs outputs in node blocks to simplify and speed-up the parameters adaptation (12).

$$Y = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad k = \textit{the number of active neurons + CTs (node PDE solutions)} \qquad (12)$$

Multi-objective algorithms can efficiently perform the "back-production" and output calculation of neurons and CTs in the tree-like structure (Fig.3). D-PNN selects the best inputs of 2-combination blocks in each layer node (analogous to GMDH) to produce applicable sum PDE model components and pre-optimize their polynomial parameters and weights using the Gradient Steepest Descent (GSD) method [12]. This iteration algorithm skips from the actual to the next block, one by one, to select and adapt one of its sub-PDE solutions. D-PNN training error is minimized in consideration of a continual test using the External Complement of GMDH [5]. A convergent combination of selected neurons and CTs can form the optimal PDE solution [11].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\left(Y_i^d - Y_i\right)^2}{M}} \to \min \qquad (13)$$

*$Y_i$ - produced and $Y_i^d$ - desired D-PNN output for $i^{th}$ training vector of M-data samples*

The Root Mean Squared Error (RMSE) is calculated in each iteration step of the training and testing to be minimized (13).

## 5 Forecasting experiments using the estimated data periods

D-PNN applied 2-lagged time-series of 20 data inputs to predict power ½-12 hours ahead in the central wind farm at Drahany, Czech Republic. The last 6-hour data were reserved for the continual test, i.e. their data samples were not used to adapt the model parameters but only to calculate the testing error and control the training. Additional spatial historical data measurements (wind speed and direction) of 3 surrounding wind farms and meteorological observations (temperature, relative humidity, see level pressure, wind speed and azimuth) from 2 nearby airports were used [A] (Fig.4).

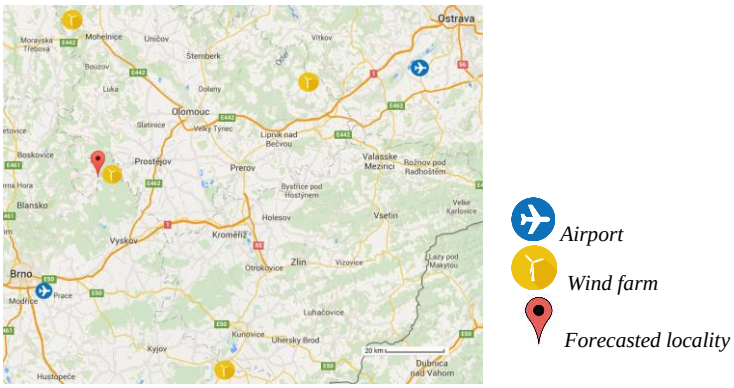



**Fig. 4** The data observation and forecasted locations

Regression SVM using the dot kernel and the GMDH Shell for Data Science, a professional self-optimizing forecasting software, were used to compare wind power predictions. Their training was analogous to that of D-PNN using the spatial data from the estimated optimal lengths of daily periods. The selection of the most valuable 2-inputs in PNN nodes of GMDH [1] is analogous to those of D-PNN and improves both the prediction models accuracy (Fig.5 – Fig.6).

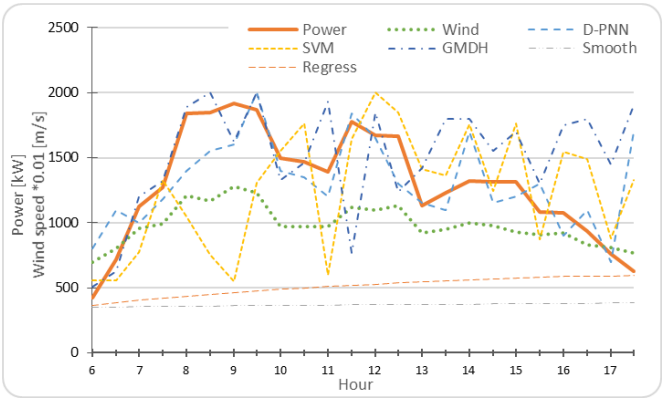



**Fig. 5.** Drahany 15.5.2011 - RMSE:
D-PNN=*314.1*, SVM=*525.4*, GMDH=*495.0*, Smooth=1021.0, Regress=897.9

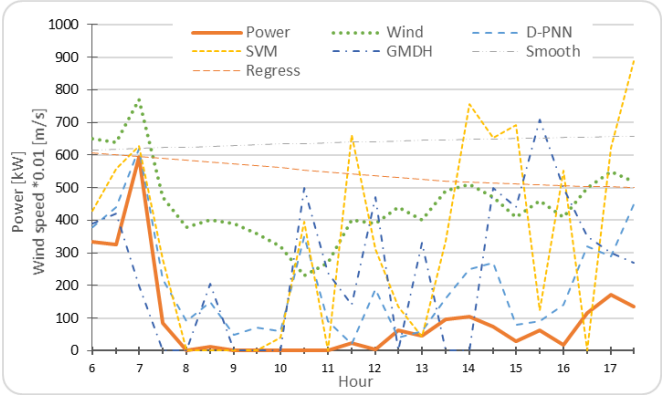



**Fig. 6.** Drahany 17.5.2011 - RMSE:
D-PNN=*140.3*, SVM=*358.9*, GMDH=*284.6*, Smooth=561.6, Regress=467.7

The performance of the soft-computing models was compared with 2 conventional regression methods - Exponential Smoothing (ES) and Linear Regression (LR) in one week 12-hour intra-day predictions, from 12 to 18 May, 2011 (Fig.7 and Fig.8). ES and LR use only the historical local time-series of wind power and their previous time-step predictions.

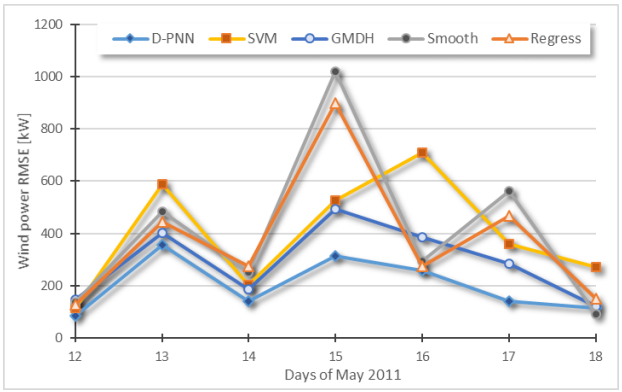



**Fig. 7.** One week daily 12-hour wind power average prediction RMSE:
D-PNN=201.1, SVM=396.1, GMDH=288.8, Smooth=405.9, Regress=377.0

The predictions of soft-computing models can mostly approximate the real power course. Their model formation is problematic, if training data patterns with an alternate output power in changeable weather do not correspond to the latest conditions in the predicted capful days with an intermittent or stable low power output (Fig.6). The actual wind speed can vary under or around the power generation limit (about 400 kW), which causes additional difficulties and failures in the predictions. A NWP analysis can detect these days to extend the training periods and include days with similar data patterns. Their predictions in catchy wind days (Fig.5) usually succeed as the training weather data better characterize predicted wind variations. The SVM

output can alternate in subsequent time-steps (Fig.5 and Fig.6), which can considerably debase the predictions. SVM is more sensitive to the precise estimation of the lengths of training periods than D-PNN or GMDH. These selective methods can apply different numbers of the last days to form the models producing similar predictions.

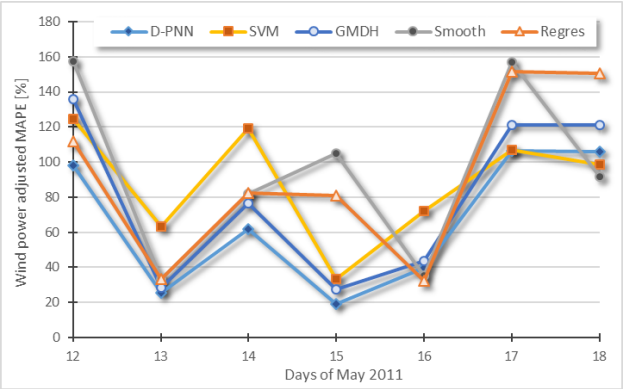



**Fig. 8.** One week daily 12-hour wind power average prediction MAPE: D-PNN=65.3, SVM=88.2, GMDH=79.1, Smooth=94.5, Regress=91.9

ES and LR predictions provide mostly only a simple course or linear trend of in the predictions of power time-series. ES can sparsely predict the rough course of real power. Both the methods can obtain lower prediction errors in calm days with slight wind speed alterations (Fig.6), which follow periods of catchy wind. Their predictions mostly represent only a trend of the predicted time-series, though they can obtain lower errors in calm days with gentle wind speed alterations. ES and LR require also estimations of the optimal daily periods, i.e. the numbers of data samples in the last days, used to calculate the parameters [10].

## 6 Conclusions

D-PNN is a novel neuro-computing method using multi-layer tree-like network structures and adapted mathematical techniques to decompose and solve the n-variable general PDE. Its selective sum PDE models can represent the local atmospheric dynamics. The presented D-PNN predictions approximate time-behavior of the real wind power in catchy or assailing wind days. They are less valuable in calm wind days following a break change in weather patterns. The compared soft-computing and conventional regression methods are not able to model complex local weather patterns in most of the predicted days. D-PNN can analogously predict the intra-day production of the photo-voltaic (PV) energy using additionally clear sky index of solar radiation, cloudiness or sky conditions [9]. NWP model forecasts are necessary in middle-term 24-48 hour predictions of statistical models [B]. The presented wind power intra-day predictions are more precise than AI converted wind speed forecasts of meso-scale NWP systems, which cannot fully consider specific local conditions.

## Acknowledgements

This paper was supported by the following projects: LO1404: Sustainable Development of ENET Centre; CZ.1.05/2.1.00/19.0389 Development of the ENET Centre Research Infrastructure; SP2018/58 and SP2018/78 Student Grant Competition and TACR TS777701, Czech Republic.

## References

[1] ANASTASAKIS, L., AND MORT, N. *The Development of Self-Organization Techniques in Modelling: A Review of the Group Method of Data Handling (GMDH)*. The University of Sheffield, 2001.

[2] BERG, L. *Introduction To The Operational Calculus*, vol. 2 of *North-Holland Series on Applied Mathematics and Mechanics*. North-Holland, New York, 1967.

[3] KIMURA, R. Numerical weather prediction. *Journal of Wind Engineering and Industrial Aerodynamics 90* (2002), 1403–1414.

[4] MONTEIRO, C., BESSA, R., MIRANDA, V., BOTTERUD, A., WANG, J., AND CONZELMANN, G. *Wind Power Forecasting: State of the Art 2009*. Report No.: ANL/DIS-10-1. Argonne National Laboratory, Argonne, Illinois, 2009.

[5] NIKOLAEV, N. Y., AND IBA, H. *Adaptive Learning of Polynomial Networks*. Genetic and evolutionary computation. Springer, New York, 2006.

[6] ZJAVKA, L. Wind speed forecast correction models using polynomial neural networks. *Renewable Energy 83* (2015), 998–1006.

[7] ZJAVKA, L. Numerical weather prediction revisions using the locally trained differential polynomial network. *Expert Systems With Applications 44* (2016), 265–274.

[8] ZJAVKA, L. Multi-site post-processing of numerical forecasts using a polynomial network substitution for the general differential equation based on operational calculus. *Applied Soft Computing 73* (2018), 192–202.

[9] ZJAVKA, L., KRÖMER, P., MIŠÁK, S., AND SNÁŠEL, V. Modeling the photovoltaic output power using the differential polynomial network and evolutional fuzzy rules. *Mathematical Modelling and Analysis 22* (2017), 78–94.

[10] ZJAVKA, L., AND MIŠÁK, S. Direct wind power forecasting using a polynomial decomposition of the general differential equation. *IEEE Transactions on Sustainable Energy 9* (2018), 1529–1539.

[11] ZJAVKA, L., AND PEDRYCZ, W. Constructing general partial differential equations using polynomial and neural network. *Neural Networks 73* (2016), 58–69.

[12] ZJAVKA, L., AND SNÁŠEL, V. Constructing ordinary sum differential equations using polynomial networks. *Information Sciences 281* (2014), 462–477.

[A] Weather underground historical data series: www.wunderground.com/history/airport/LKTB/2016/7/22/DailyHistory.html

[B] Weather underground tabular forecasts: www.wunderground.com/cgi-bin/findweather/getForecast?query=LKMT

# Stochastic Weather Generators in Czechia: 25 Years of Development and Applications

Martin Dubrovsky, Radan Huth, Ondrej Lhotka, Jiri Miksovsky, Petr Stepanek, Jan Meitner and Miroslav Trnka
Global Change Research Institute CAS, Brno; Institute of Atmospheric Physics CAS, Prague

## Abstract

Stochastic weather generators (WGs) are software tools, which can produce synthetic weather time series statistically similar to real-world weather data. This means that the relevant statistics representing probabilistic distributions of individual variables, correlations between variables, and temporal & spatial structure derived from the synthetic series are as close as possible to those derived from observations.

WGs can do different tasks. Most typically, they are used to produce synthetic weather series serving as an input to various models (often agricultural and hydrological models; these will be called "impact models" in next) employed in simulating weather-dependent processes. In this role, WGs are commonly used in cases when observational data are not available or they are not sufficiently long or when we want to assess variability or changes in characteristics simulated by the impact models in response to weather variability or projected climate change. In the later case (climate change impact studies), parameters of the generator are derived from the baseline (present climate) weather data in the first step, and then they are changed according to climate change scenarios derived from the dynamical climate models (either Global Climate Models or Regional Climate Models). WGs may be applied also in other tasks. Specifically, our M&Rfi generator may be linked with the weather forecast and used for a probabilistic crop yield forecasting. Our other generator SPAGETTA (being a multi-site WG) has been used for studying the collective significance of local trends at multiple sites and developing a new test for examining this significance by analysis of data from multiple mutually correlated sites; all WG parameters representing the statistical structure of the series are prescribed by the user in this mode of operation.

Although no WG can produce perfectly realistic weather data, they have important advantages, which make them a favourite tool used in applied climatology: (i) WGs are very fast and may produce a large number of realisations of arbitrarily long weather series in a reasonable time (in contrast with complex but slow dynamical climate models. (ii) WGs may be interpolated so that they can produce data even for sites where there are no weather observations available for WG calibration. (iii) WGs are very flexible and may handle various combinations pf weather variables needed to feed the impact models. (iv) WGs may produce weather series representing both present and future climates (even for emission scenarios, for which outputs from dynamical climate models are not available), as well as the weather series which conform to the probabilistic weather forecast.

Various WGs are available, they may differ in several aspects: (a) The modeling approach may be parametric (most common approach) or semi-parametric. In the parametric WGs, the temporal structure is often based on Markov chains (for precipitation occurrence) and autoregressive models (non-precipitation variables) and the variables are assumed to have a specific distribution function (for example, Gamma distribution or mixed-exponential distribution are used for daily precipitation sums). The non-parametric generators are based on resampling. Apart from that, some generators may be called semi-parametric as they use both parametric and non-parametric approaches (e.g. LARS generator developed by M. Semenov). To make the temporal structure of the synthetic series more realistic, the surface weather may be generated conditionally on large-scale circulation characteristics or on weather series simulated by WG running at longer (e.g. monthly) time step. (b) Number of variables: the generators may simulate only a single variable, but they are mostly multivariate. Specifically, in agrometeorological applications, up to 6 variables are commonly needed: daily temperature minimum and maximum, daily sums of precipitation and incoming solar radiation, and daily means of air humidity and wind speed. (c) Time step ranges from continuous time-scale or very short time intervals (minutes or shorter) models (using different statistical models instead of Markov chains and AR models) to hourly, daily and monthly or even annual generator; most commonly, the daily weather generators are used in both agrometeorological and hydrological applications. (d) Spatial aspect: in agrometeorology, single-site generators are usually used as the simulated processes mostly depend only on the site-specific weather. On the contrary, in hydrological modelling, where the proper simulation of rainfall-runoff processes requires realistically spatially coherent data, multi-site (the sites may be regularly or irregularly distributed in space) WGs are needed.

In our contribution, we will give a brief overview of the weather generators developed since 1994 by the main author in a close co-operation with agrometeorologists, hydrologists, programmers and other statistical climatologists. The WGs mostly (but not only) run at daily time step, they are based on both parametric and non-parametric approaches, and those developed before 2016 are only single-site - only the most recent SPAGETTA generator is multi-site; all of them are multi-variate. The focus will be put on two generators most commonly used by our (but not only) team: (A) M&Rfi is a parametric single-site multi-variate daily generator, which has been involved in many agrometeorological experiments since 1994. In this generator, precipitation occurrence is modelled with the Markov chain, precipitation amount is sampled from the Gamma distribution, and the non-precipitation variables are simulated using the first-order AR model, whose parameters are conditioned on precipitation occurrence. (B) SPAGETTA is a spatial generator, whose development (started in 2016) was motivated by its involvement in hydrological modelling and in studying the collective significance of local trends at multiple mutually correlated sites. SPAGETTA is based on a single-site M&Rfi generator, which was spatialised using the Wilks' approach. As a part of our presentation, we will show selected results obtained while validating the two generators in terms of various climatological indices (for example in terms of occurrence of

hot/cold/wet/dry spells as well as the compound hot-dry/hot-wet/cold-dry/cold-wet spells). Some results obtained with these generators while being employed in climate change impact experiments will be also shown. In addition, we will demonstrate the use of the generators in other applications: application of M&Rfi in seasonal forecasting of crop yields, and application of SPAGETTA in developing the test for examining a collective significance of local trends at multiple sites.

**Acknowledgement:** The development and applications of our weather generators have been funded by several projects since the end of 20th century. The most recent experiments are made within the frame of project SustES ("Adaptation strategies for sustainable ecosystem services and food security under adverse environmental conditions"; CZ.02.1.01/0.0/0.0/16_019/0000797) and project GRIMASA ("Development of high-resolution spatial weather generator for use in present and future climate conditions" project no.18-15958S) funded by Czech Science Foundation..

# Wind and Solar Forecasting for Renewable Energy System using SARIMA-based Model

Marwa Haddad[1], Jean-Marc Nicod[1], Yacouba Boubacar Maïnassara[2], Landy Rabehasaina[2], Zeina Al Masry[1], and Marie-Cécile Péra[1]

[1] FEMTO-ST institute, Univ. Bourgogne Franche-Comté, CNRS/ENSMM France
[2] Laboratory of Mathematics (LMB), Univ. Bourgogne Franche-Comté, France

**Abstract** In order to completely fulfill a datacenter power demand, one important issue is to determine and investigate a reasonable sizing for Hybrid Renewable Energy System (HRES). Usually, in the context of datacenter renewable power supply, the energy production is hybrid and it consists of wind and solar energy production associated with battery and hydrogen energy. To design the electrical energy system, one needs to forecast weather conditions (solar radiation, wind speed) in order to evaluate the energy production yearly. The aim of this paper is to propose a SARIMA-based model for a particular renewable energy system. Indeed, thanks to the wind turbine and solar panel models, it is possible to optimize the overall cost of the global energy system. We finally validate the proposed model on actual data.

## 1 Introduction

The twentieth century witnessed a boom in the number of data centers around the world driven by a rapid growing demand for Cloud services. Consequently, the energy footprint of the IT sector has increased exponentially and reached unprecedented levels. It is, actually, estimated to consume approximately 7% of global electricity in 2007 [1]. Furthermore, in 2016, statistics shows that data centers demand reached 91 billion kWh of electricity which is twice more than New York city consumption [2].

In this scenario, with climatic conditions going for drastic reversals, a global alert concerning the environment, the greenhouse gas (GHG) emissions, air pollution, social concerns and other energy security issues [3,4] is raised. Consequently, the attention of many government and researchers around the world has shifted to find a new alternative energy sources that matches with the environment. One of the most popular solution is the utilization of renewable energy sources as they have been established to be sustainable, economical, nature friendly, abundant, non-polluting and renewable [5,6]. In fact, the European Technology Platform for electricity networks of the future, known as ETP Smart grid expected that, by 2020, approximately 34% of the total electrical consumption will come from renewable energy and will have gone more than that by 2035 [7]

Nevertheless, considering the intermittent nature of solar radiation and wind, and the high capital and operational costs of solar panels and wind turbines

with the necessary energy storage devices, forecasting the next-day outputs of the power generation systems becomes a major issue to evaluate the appropriate power architecture sizing. Thus, a lot of research teams around the world and particularly in the coastal area [8,9,10] mobilize their efforts on either solar prediction or wind speed prediction. As a result, many forcasting methods have been developed by experts around the world [11,12,13] that could be classified following their approach and the time scale of prediction (e.g., physical approach such as Numeric Weather Prediction (NWP)). This model solves complex mathematical models using weather data like temperature, pressure, surface, etc. NWP is usefulness for medium to long-term forecasts ($>$ 6h ahead) [14,15]. Also, statistical approach which is based on training the measurement data by using the difference between the predicted and the actual wind speeds in immediate past to tune model parameters such as neural network (NN) based methods, and Time-Series based models like ARMA [16], ARIMA [17], Grey Predictors, Linear Predictions, etc. Finally, a hybrid approach exists with a combination of different approaches like combining short-term and medium-term models or mixing physical and statistical approaches [18,19,20].

This paper focuses on applying a statistic approach for forcasting wind speed and solar radiation on two different location, Los Angeles and Chicago in the USA. Considering the history of meteorological conditions in terms of solar radiation and wind speed at the two selected sites and the mathematical models of wind turbine and solar panels, one can compute the electrical production during the time horizon considered. Thus, based on this amount of energy production, the sizing of a hybrid renewable energy system composed of wind turbines, solar panels, battery and hydrogen storage systems has to be designed to completely supply a data center whose demand has a peak power less than 500 kW.

The remainder of the paper is organized as follows. Section 2 presents the methodology used in order to forecast weed speed and solar irradiation to be able to designed a hybrid renewable energy systems supplying a datacenter. This sizing is briefly explained in Section 3. Then, the obtained results are presented and discussed in Section 4. A conclusion and perspectives are given in Section 5.

## 2 Forecasting Methodology

Before presenting the whole methodology, we start by introducing the type of data as well as their locations. We here dispose of two types of data: solar radiation and wind speed. The latter could be obtained from various databases online such as the national solar data base (NSRDB) [21], the Modern-Era Retrospective analysis for Research and Applications (MERRA2) [22], the wind prospector from the National Renewable Energy Laboratory (NREL) [23]. In our case, the data are obtained from NSRDB AND NREL. Recall that the aim of this paper is to propose a statistical approach for wind and solar forecasting. For that purpose, based on a review and results obtained by different researchers mentioning the accruacy of the ARIMA model [18,17,24,25], we have selected the SARIMA model [26,27]. In order to verify the robustness of the SARIMA approach on

our application, we will apply the methodology on two distinct locations having different characteristics.

### 2.1 SARIMA model

ARIMA is a statisctical approach widely used in today's world since the evolution of sophisticated statistical software package. ARIMA has four major steps in model building- Identification, Estimation, Diagnostics & Forecast. Then, the general scheme for ARIMA model is translated by:

1. Identification of the model structure.
2. Application of autocorrelation function (ACF) and partial autocorrelation function (PACF) in order to identify the orders of the ARMA model. The parameters of the model are estimated by a maximum likelihood (ML) function
3. Testing the goodness of fit on the estimated model residuals
4. Using the estimated model for forecasting.

ARIMA model uses the historic data and decomposes it into autoregressive (AR), Integrated (I) indicates linear trends or polynomial trend and Moving Average (MA) indicates weighted moving average over past errors. Therefore, it has three model parameters AR($p$), I($d$) and MA($q$) all combined to form ARIMA($p$, $d$, $q$) model where:

- $p$ = order of AR
- $q$ = order of MA
- $d$ = order of I (differencing)

The multiplicatif Seasonal ARIMA model namely SARIMA is actually a variation of the classical ARIMA model. In order to take into account the seasonal effect of the irradiation and the wind speed, this model is generally written as SARIMA($p$,$d$,$q$)($P$,$D$,$Q$) where, as in the ARIMA model, $p$, $d$, $q$ and $P$, $D$, $Q$ are non-negative integers that refer to the polynomial order of the AR, I, MA parts of the non-seasonal and seasonal components of the model, respectively. Mathematically, the SARIMA model is defined as in (1)

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D x_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{1}$$

Where: $x_t$ is the forecast variable (i.e., solar radiation), $\phi_p(B)$ is the regular AR polynomial of order $p()$, $\theta_q(B)$ is the regular MA polynomial of order $q()$, $\Phi_P(B^s)$ is the seasonal AR polynomial of order $P()$, $\Theta_Q(B^s)$ is the seasonal MA polynomial of order $Q$, $\nabla^d$ is the differentiating operator that eliminate the non-seasonal non-stationarity, $\nabla_s^D$ is the seasonal differentiating operator that eliminate the seasonal non-stationarity, $B$ is the backshift operator, which shift one point in time the observation $x_t$ (i.e., $B^k(x_t) = x_{t-k}$) and finally $\varepsilon_t$ follows a white noise process and $s$ defines the seasonal period. These polynomials are described mathematically in Equations (2):

$$\begin{aligned} \theta_q(B) &= 1 - \sum_{i=1}^{q} \theta_i B^i \qquad \Theta_Q(B^s) = 1 - \sum_{i=1}^{Q} \Theta_i B^{s,i} \\ \phi_p(B) &= 1 - \sum_{i=1}^{p} \phi_i B^i \qquad \Phi_P(B^s) = 1 - \sum_{i=1}^{P} \Phi_i B^{s,i} \\ \nabla^d &= (1-B)^d \qquad \nabla_s^d = (1-B^s)^D \end{aligned} \tag{2}$$

In order to get the model that fits the best the data, the Akaike Information Criterion (AIC) is a statistic measure to compare them. In fact, the AIC rewards models for a good fit and penalize others for complexity. It could be written as:

$$AIC = 2k + \ln\left(\frac{RSS}{n}\right) \tag{3}$$

with $k$ the number of free parameters, $n$ the total number of observations equal to 468 and $RSS$ is the residual sum of squares.

Finally, using the obtained valid model, one can proceed to the forecasting of the wished period.

### 2.2 Evaluation of the forecasting performance

The forecasting model is constructed on time series of solar radiation and wind speed for a duration of 9 years weekly (which means 468 values). Once the models are formulated, they are used to forecast wind speed and solar radiation for the last two years . Afterwards, the averages of the statistics for the 2 years forecasting results are computed to analyze the models' accuracy. Several measurement statistics can be used to examine the forecast accuracy of different models. Mean absolute percentage error (MAPE) is used very often to evaluate the performance of the forecasting model. The above-mentioned statistical quantities are computed as in (4):

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{x_t - \hat{x}_t}{x_t} \right| \tag{4}$$

where $\hat{x}_t$ is the forecast value.

## 3 Hybrid renewable energy system sizing

The aim of this work is to determine an the optimal size of the stand-alone hybrid renewable energy sources (HRES) to fulfill the energy demand of the data center. These electrical sources are divided into 2 different subsystems:

– The primary sources: consist in providing the basic power to supply the data center and are composed of photovoltaic panels and wind turbines.

– The secondary sources: are the back up power to supply the data center in times of need and are composed of batteries and fuel cells.

As the datacenter should be autonomous in terms of energy consumption, the totality of the energy comes from primary sources. Moreover knowing that the primary sources operate as intermittent sources in time, we have to balance the lack of energy production (for example in the winter) by an over production (summer) during the year. To achieve this balance, the secondary sources (storage sources) are divided following the type of storage and operate as follows:

– Long-term storage: where the day of overproduction will balance the days of underproduction. The electrical resources used in this case is the hydrogen system
– Short-term storage: where the hours of overproduction will balance the hours of underproduction during the same day (fluctuations between day and night). It means that the production will be smoothed over the day. The electrical resources used in this case is the batteries.

Using the data center demand and the meteorological data downloaded, one needs to understand the models of the first subsystem in order to proceed to the sizing of the primary sources.

**Solar Model** To model the relation between the irradiation data and the output power $P_pv$ of the PV panels, a widely used model [28,29,30,31,32] is described by Equation (5):

$$P_{pv} = I \times A_{pv} \times \eta_{pv} \tag{5}$$

where, $I$ is the hourly solar irradiance in $kW/m^2$, $A_{pv}$ is the area of the PV panels in $m^2$, and $\eta_{pv}$ is their efficiency.

**Wind model** The model of the output power of one wind turbine generator $P_w$ that follows the power curve is shown in Figure 1 [33]. So, the turbine starts generating power at the "cut-in" wind speed $v_{ci}$. Then, the generated output power increases with the increase of wind speed from the "cut-in" $v_{ci}$ to the rated wind speed $v_r$. When the wind speed varies between the rated wind and the "cut-out" wind speed $v_{co}$, which is the maximum wind speed value at which the turbine can correctly work, the turbine produces a constant or "rated power". Once the wind speed goes beyond the "cut-out" speed, the turbine stops generating for safety reasons.

Many other papers, such as [34,35,36,32], have adapted this mathematical model of the wind turbine power output that can be written as in Equation (6):

$$P_w = \begin{cases} 0 & \text{if} \quad v(t) \leq v_{ci} \text{ or } v(t) \geq v_{co} \\ P_r \dfrac{v(t) - v_{ci}}{v_r - v_{ci}} & \text{if} \quad v_{ci} < v(t) < v_r \\ P_r & \text{if} \quad v_r < v(t) < v_{co} \end{cases} \tag{6}$$

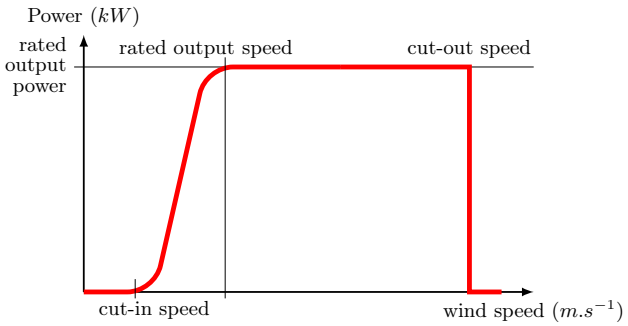



Figure 1: Ideal wind turbine power output

where $v(t)$ is the wind speed $(m.s^{-1})$ at any time $t$ $(s)$ and $P_r$ is the nominal power of the wind turbne.

## 4 Results and discussions

The data representing the solar radiation and wind speed were measured on an hourly scale from January 2004 till December 2012 (more than 6 years) in two different location. To be more precise, the endogenous data of the solar radiation and wind speed time series were measured at Chicago (Latitude: 41.810539, Longitude: -87.643127, Time Zone: -6 ) and at Los Angeles (Latitude: 34.57, Longitude: -118.02, Time Zone: -8). Then, in order to obtain weekly values, we calculated averages per groups of 168 values (168 hours per week). Finally, we have obtained time series of 52 value per year, i.e., 468 values during the nine years. Figures 4d and 4c showed this distribution respectively for solar radiation and wind speed in Los Angeles. The first nine years have been used to setup our models and the last two years to test them. The model has been implemented using R programming language.

### 4.1 Models validation

Based on Figure 4d, the measured solar radiation from 2004 till 2012 is quite seasonal. In fact, the data starts from the first week of January till the last week of December. Each year, the pic of solar radiation is in July that corresponds to the summer season where days are quite long. Contrariwise, the lowest values are obtained in December or in January. This period matches with the winter where days are short. Thus, the solar distribution is intrinsically seasonal and periodic which validates the choice of the SARIMA model.

In Figure 4c, the data also starts from the first week of January till the last week of December. Moreover, it shows a random distribution where data varies from $3\,m/s$ till $14.8\,m/s$. This series presents a seasonality that could be well seen

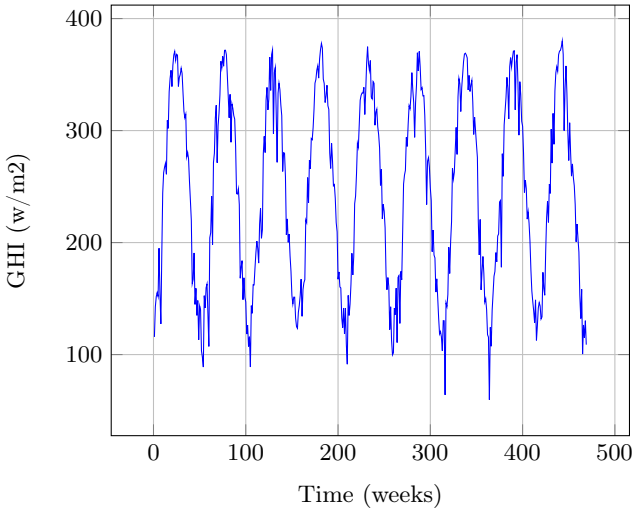



Figure 2: Weekly solar radiation distribution in Los Angeles

especially starting from the week 150. The wind speed is quite low in the winter and increases with the oncoming of summer corresponding to the thermal hot wind of Los Angeles. Nevertheless, the wind can vary from one year to another so we cannot confirm the periodicity.

Table 1: Comparison of the statistic criterion AIC for wind speed in both Chicago and Los Angeles

| ***DATA*** | ***SARIMA Configurations*** | ***AIC*** |
|---|---|---|
| Chicago | SARIMA(21,0,21)(1,1,0) | 1441,393 |
| | SARIMA(11,0,14)(1,1,1) | 1359,254 |
| | SARIMA(11,0,14)(0,1,1) | 1358,813 |
| | SARIMA(18,0,18)(0,1,0) | 1498,97 |
| Los Angeles | SARIMA(9,0,19)(1,1,0) | 1886,61 |
| | SARIMA(6,0,6)(1,1,1) | 1844,582 |
| | SARIMA(6,0,6)(0,1,1) | 1839,007 |
| | SARIMA(9,0,0)(0,1,0) | 2006,21 |

In order to obtain the model that fits the best the data, different configurations of SARIMA have been applied on the distributions for the two locations. In each set, 4 seasonal configurations have been applied such as the seasonal polynomial AR and MA respectively $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are set as explained in Table 1

Based on results given in Tables 1 and 2 in the two different cities, with completely different characteristics, one can see that the best AIC obtained is

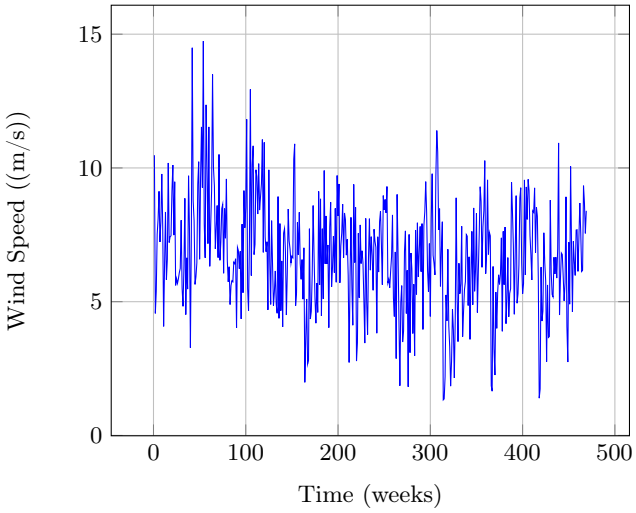



Figure 3: Weekly wind speed distribution in Los Angeles

Table 2: Comparison of the statistic criterion AIC for solar radiation in both Chicago and Los Angeles

| ***DATA*** | ***SARIMA Configuration*** | ***AIC*** |
|---|---|---|
| Chicago | SARIMA(10,0,9)(1,1,0) | 4162,49 |
| | SARIMA(4,0,18)(1,1,1) | 4075,66 |
| | SARIMA(4,0,18)(0,1,1) | 4075,32 |
| | SARIMA(10,0,9)(0,1,0) | 4283,52 |
| Los Angeles | SARIMA(20,0,14)(1,1,0) | 3796,42 |
| | SARIMA(13,0,14)(1,1,1) | 3735,64 |
| | SARIMA(13,0,14)(0,1,1) | 3733,59 |
| | SARIMA(13,0,20)(0,1,0) | 3884,346 |

the one of the model configuration SARIMA($p$,$d$,$q$)(0,1,1) for both solar radiation and wind speed data. For instance, The SARIMA model$(6,0,6)(0,1,1)$ is written during a period of $s = 52$ as in Equation (7)

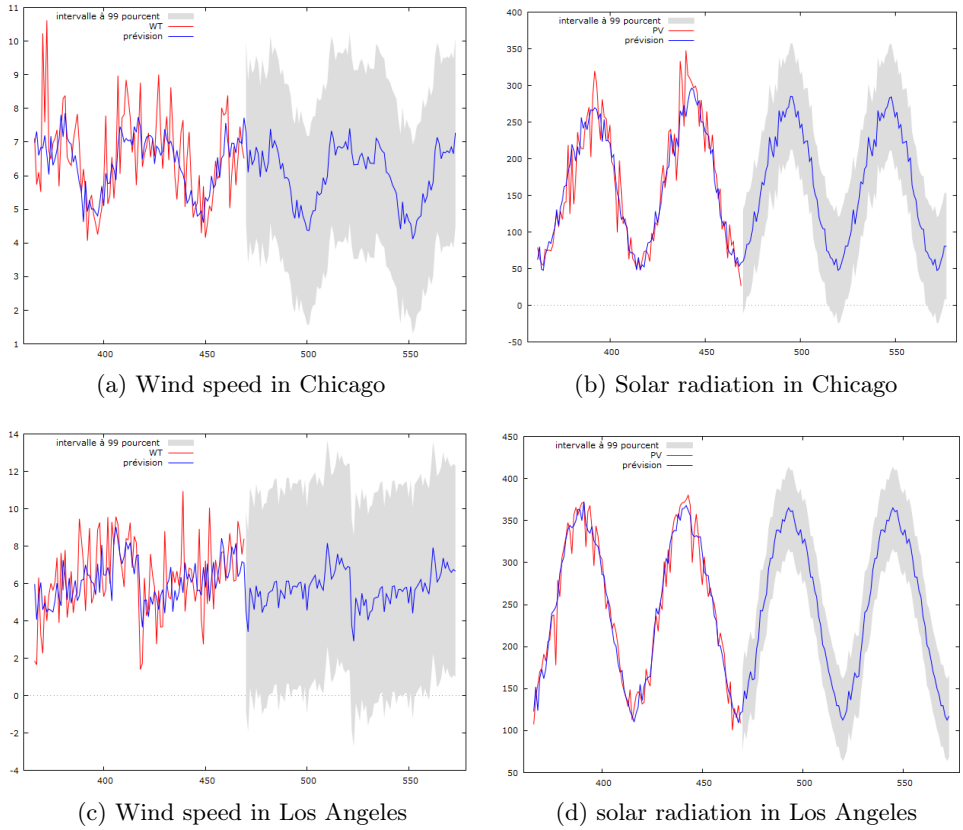
$$(1-\phi_1 B^1-\phi_3 B^3-\phi_6 B^6)x_t = (1-\Theta_1 B^s)(1-\theta_1 B^1-\theta_3 B^3-\theta_3 B^3-\theta_6 B^6)\varepsilon_t \quad (7)$$

Thus, only the latter is maintained as valid models to be used in the forecasting of the solar radiation and wind speed for a duration of two years.

### 4.2 Forecasting evaluation

Now coming back to the objective to predict the future meteorological data with the valid SARIMA$(p,d,q)(0,1,1)$ model obtained in the section before, the

results are shown in Figure 4. Moreover, to investigate the model sufficiency, we summarize the useful statistics about the forecasting results in Table 3 by computing the mean absolute percentage error of all the tested weeks.




(a) Wind speed in Chicago

(b) Solar radiation in Chicago

(c) Wind speed in Los Angeles

(d) solar radiation in Los Angeles

Figure 4: Forcasting results of the wind speed and solar radiation for Chicago and Los Angeles

The forecasting model is applied on four years where it compares with the last two years 2011 and 2012 and continue the forecasts till 2014 weekly.

Based on the solar radiation prevision in Figures 4b and **??**, the forecasting curves in blue are quite fitting the real data of the years 2011 and 2012 and follow the seasonality. Moreover, with a MAPE equal to 7 and 15.60 for the two sites Los Angeles and Chicago respectively, the SARIMA model is quite validated and accurate.

Figures 4a and 4c show the wind speed forecasting starting from 2011 till 2014. The forecasting curves in blue is following the trend of the real data of the

Table 3: The mean absolute percentage error of the used methods

| Data | Locations | Methode | MAPE |
|---|---|---|---|
| Solar radiation | LA | SARIMA(4,0,18)(0,1,1) | 7,03 |
| | Chicago | SARIMA(11,0,14)(0,1,1) | 15,60 |
| Wind Speed | LA | SARIMA(6,0,6)(0,1,1) | 29,83 |
| | Chicago | SARIMA(13,0,14)(0,1,1) | 12,54 |

years 2011 and 2012. Nevertheless, the the 99% confidence interval is quite large and indicates high variability. Thus, these rates should be interpreted with the noise variance estimated. The SARIMA model applied on the wind speed in these two cities, Los Angeles and Chicago, is not as precise as the solar forecasting.

Finally, recall that all interpretations and conclusions presented in this paper are based on data available for the specific areas.

## 5 Conclusion

This paper presents a comparison among four distinct solar radiation and wind speed generation forecasting models. It is shown that in general, SARIMA model is quite good in the forecasting of the solar radiation during years and fits very well the data because of their seasonal distribution. We also pointed out that the performance of the used model in forecasting the solar during the years is more precise than the ones for the wind speed which degrades noticeably for long term previsions. It is hence important to predict wind speed variation as precisely as possible. This shows the interest to consider other models or characteristics such as Markov Switching ARMA [37] to improve the precision of the results in order to get an optimal sizing for the hybrid renewable energy system supplying a datacenter power demand.

## 6 Acknowledgments

This work was supported in part by the ANR DATAZERO (contract “ANR-15-CE25-0012”) project and by the EIPHI Graduate school (contract "ANR-17-EURE-0002").

## References

1. Cook, G., Lee, J., Tsai, T., Kong, A., Deans, J., Johnson, B., Jardim, E.: Clicking clean: Who is winning the race to build a green internet? Greenpeace Inc., Washington, DC (2017) 5
2. Zik, O., Shapiro, A.: Coal computing: How companies misunderstand their dirty data centers. Lux Research White Paper (2016)
3. Li, Y., Xue, B., He, X.: Catalytic synthesis of ethylbenzene by alkylation of benzene with diethyl carbonate over hzsm-5. Catalysis Communications **10**(5) (2009) 702–707

4. Demirbas, A.: Present and future transportation fuels. Energy Sources, Part A **30**(16) (2008) 1473–1483
5. Bajpai, P., Dash, V.: Hybrid renewable energy systems for power generation in stand-alone applications: A review. Renewable and Sustainable Energy Reviews **16**(5) (2012) 2926–2939
6. Guinot, B., Champel, B., Montignac, F., Lemaire, E., Vannucci, D., Sailler, S., Bultel, Y.: Techno-economic study of a pv-hydrogen-battery hybrid system for off-grid power supply: Impact of performances' ageing on optimal system sizing and competitiveness. International Journal of Hydrogen Energy **40**(1) (2015) 623–632
7. Van, T.V., Norton, M., Ivanov, C., Delimar, M., Hatziargyriou, N., Stromsather, J., Iliceto, A., Llanos, C., Panciatici, P.: Organic growth: toward a holistic approach to european research and innovation. IEEE Power and Energy Magazine **13**(1) (2015) 30–37
8. Kalogirou, S.A.: Artificial neural networks in renewable energy systems applications: a review. Renewable and sustainable energy reviews **5**(4) (2001) 373–401
9. Mellit, A., Kalogirou, S.A., Hontoria, L., Shaari, S.: Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews **13**(2) (2009) 406–419
10. Paoli, C., Voyant, C., Muselli, M., Nivet, M.L.: Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy **84**(12) (2010) 2146–2160
11. Mubiru, J., Banda, E.: Estimation of monthly average daily global solar irradiation using artificial neural networks. Solar Energy **82**(2) (2008) 181–187
12. Kaplanis, S.: New methodologies to estimate the hourly global solar radiation; comparisons with existing models. Renewable Energy **31**(6) (2006) 781–790
13. Elminir, H.K., Azzam, Y.A., Younes, F.I.: Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. Energy **32**(8) (2007) 1513–1523
14. Potter, C.W., Negnevitsky, M.: Very short-term wind forecasting for tasmanian power generation. IEEE Transactions on Power Systems **21**(2) (2006) 965–972
15. Lange, M., Focken, U.: New developments in wind energy forecasting. In: 2008 IEEE power and energy society general meeting-conversion and delivery of electrical energy in the 21st century, IEEE (2008) 1–8
16. Torres, J.L., Garcia, A., De Blas, M., De Francisco, A.: Forecast of hourly average wind speed with arma models in navarre (spain). Solar Energy **79**(1) (2005) 65–77
17. Kavasseri, R.G., Seetharaman, K.: Day-ahead wind speed forecasting using f-arima models. Renewable Energy **34**(5) (2009) 1388–1393
18. Voyant, C., Muselli, M., Paoli, C., Nivet, M.L.: Numerical weather prediction (nwp) and hybrid arma/ann model to predict global radiation. Energy **39**(1) (2012) 341–355
19. Chang, P.S., Li, L.: Ocean surface wind speed and direction retrievals from the ssm/i. IEEE transactions on geoscience and remote sensing **36**(6) (1998) 1866–1871
20. Alexiadis, M., Dokopoulos, P., Sahsamanoglou, H.: Wind speed and power forecasting based on spatial correlation models. IEEE Transactions on Energy Conversion **14**(3) (1999) 836–842
21. Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., Shelby, J.: The national solar radiation data base (NSRDB). Renewable and Sustainable Energy Reviews **89** (2018) 51–60
22. Bosilovich, M., Lucchesi, R., Suarez, M.: MERRA-2: File specification. https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich785.pdf (2015)

23. Fingersh, L., Simms, D., Hand, M., Jager, D., Cotrell, J., Robinson, M., Schreck, S., Larwood, S.M.: Wind tunnel testing of NREL's unsteady aerodynamics experiment. In: 20th ASME Wind Energy Symposium. (2001)
24. Hassan, J.: Arima and regression models for prediction of daily and monthly clearness index. Renewable Energy **68** (2014) 421–427
25. Bouzerdoum, M., Mellit, A., Pavan, A.M.: A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. Solar Energy **98** (2013) 226–235
26. Brockwell, P.J., Davis, R.A.: Time series: theory and methods. Second edn. Springer Series in Statistics. Springer-Verlag, New York (1991)
27. Hipel, K., McLeod, A.I.: Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam (1994)
28. Bortolini, M., Gamberi, M., Graziani, A.: Technical and economic design of photovoltaic and battery energy storage system. Energy Conversion and Management **86** (2014) 81–92
29. Sediqi, M.M., Furukakoi, M., Lotfy, M.E., Yona, A., Senjyu, T.: Optimal economical sizing of grid-connected hybrid renewable energy system. Journal of Energy and Power Engineering **11**(4) (2017) 244–53
30. Kaldellis, J., Zafirakis, D., Kondili, E.: Optimum autonomous stand-alone photovoltaic system design on the basis of energy pay-back analysis. Energy **34**(9) (2009) 1187–1198
31. Sinha, S., Chandel, S.: Review of recent trends in optimization techniques for solar photovoltaic–wind based hybrid energy systems. Renewable and Sustainable Energy Reviews **50** (2015) 755–769
32. Tina, G., Gagliano, S.: Probabilistic analysis of weather data for a hybrid solar/wind energy system. International Journal of Energy Research **35**(3) (2011) 221–232
33. Garcia, R.S., Weisser, D.: A wind–diesel system with hydrogen storage: Joint optimisation of design and dispatch. Renewable energy **31**(14) (2006) 2296–2320
34. Dong, W., Li, Y., Xiang, J.: Optimal sizing of a stand-alone hybrid power system based on battery/hydrogen with an improved ant colony optimization. Energies **9**(10) (2016) 785
35. Zeng, J., Li, M., Liu, J., Wu, J., Ngan, H.: Operational optimization of a stand-alone hybrid renewable energy generation system based on an improved genetic algorithm. In: Power and Energy Society General Meeting, 2010 IEEE, IEEE (2010) 1–6
36. Yang, H., Lu, L., Burnett, J.: Weather data and probability analysis of hybrid photovoltaic–wind power generation systems in hong kong. Renewable Energy **28**(11) (2003) 1813–1824
37. Francq, C., Gautier, A.: Large sample properties of parameter least squares estimates for time-varying ARMA models. J. Time Ser. Anal. **25**(5) (2004) 765–783

# Deterministic weather forecasting with a newly developed non-hydrostatic global atmospheric model

Song-You Hong
KIAPS (South Korea)

## Abstract

Numerical weather prediction based on computer programs of weather phenomena is a grand challenge and has been continuously progressed for past half century. Korea Institute of Atmospheric Prediction Systems (KIAPS) has embarked a national project in developing a new global forecast system in 2011. The ultimate goal of this 9-year project is to replace the current operational model at Korea Meteorological Administration (KMA), which was adopted from the United Kingdom's Meteorological Office's operational model. As of January 2019, the 12-km Korean Integrated Model (KIM) system that consists of a spectral-element non-hydrostatic dynamical core on a cubed sphere and the state-of-the-art physics has been launched in a real-time forecast framework, with the initial conditions obtained from the advanced 4-DEnvar over its native grid. A background on the KIAPS mission and the development strategy of KIM toward a world-class global forecast system are described, along with a future plan for operational deployment.

# The Generalized STAR Model with Spatial and Time Correlated Errors to Analyze the Monthly Crime Frequency Data

Utriweni Mukhaiyar, Udjianna Sekteria Pasaribu, Kurnia Novita Sari, and Debby Masteriana

Statistics Research Group, Faculty of Mathematics and Natural Sciences,
Institut Teknologi Bandung,
St. Ganesha 10 Bandung, 40132, West Java, Indonesia
{utriweni,udjianna,kurnia}@math.itb.ac.id
debbymath.itb@gmail.com
http://www.math.itb.ac.id

**Abstract.** Most of space-time series models assume that errors are independent or at least uncorrelated. Here the errors are considered have both spatial and time correlated errors (STCE). This assumption defines that errors in a certain location is affected by its neighbors errors of previous times. Such errors are attached to Generalized Space-Time Autoregressive (GSTAR) model. Due to the errors dependence, generalized least squares is applied to estimate the parameters. For case study, the monthly crime frequency data is used. It is obtained that the GSTAR(1; 1) with STCE give the lower Akaike Information Criteria (AIC) than higher GSTAR model with time correlated errors.

**Keywords:** space-time, correlated errors, autoregressive, Generalized Least Squares, forecasting

# The Generalized STAR Model with Adjacency-Spatial Weight Matrix Approach to Investigate the Vehicle Density in Nearby Toll Gates

Utriweni Mukhaiyar, Kurnia Novita Sari, and Nur Tashya Noviana

Statistics Research Group, Faculty of Mathematics and Natural Sciences,
Institut Teknologi Bandung,
St. Ganesha 10 Bandung, 40132, West Java, Indonesia
{utriweni,kurnia}@math.itb.ac.id
nurtashya.noviana@gmail.com
http://www.math.itb.ac.id

**Abstract.** A spatial weight matrix represents dependency among observed locations. In Generalized Space Time Autoregressive (STAR) modelling, the non-uniform type of this matrix is mostly built based on Euclidean distance, such that it become fixed all the time. This study use adjacency matrix approach which is constructed based on correlation among time series data of each location. From the adjacency matrix, a minimum spanning tree of observed locations is acquired along with the degree of neighbours. This degree determines how many and large the neighbour locations influence the observed locations in various spatial lag. Then the more representative spatial weight matrix which represents the behaviour of real observations, is obtained. This approach is applied to analyze the pattern of number vehicles which enter the Purbaleunyi of toll gates in West Java, Indonesia. It is obtained that the simplest of GSTAR model, GSTRA$(1;1)$, is the most appropriate model to be applied for forecasting. Although adjacency matrix approach does not give the best result, but its performance is alike with the best obtained model.

**Keywords:** spatial weight, adjacency matrix, autoregressive, minimum spanning tree, forecasting

# Landslide Debris-Flow Prediction using Ensemble and Non-Ensemble Machine-Learning Methods

Praveen Kumar[1], Priyanka Sihag[1], Ankush Pathania[1], Shubham Agarwal[1], Naresh Mali[2], Pratik Chaturvedi[3], Ravinder Singh[4], K.V. Uday[2], and Varun Dutt[1]

[1]Applied Cognitive Science Lab, Indian Institute of Technology Mandi, Himachal Pradesh, India-175005
[2]Geohazard Studies Laboratory, Indian Institute of Technology Mandi, Himachal Pradesh, India-175005
[3]Defence Terrain Research Laboratory, Defence Research and Development Organization (DRDO), New Delhi, India-110054
[4]National Disaster Management Authority (NDMA), New Delhi, India-110029
bluecodeindia@gmail.com,priyankasihag8993@gmail.com, ankushpathania.ap@gmail.com,shubhamagarwal223@gmail.com, malinareshmudhiraj@gmail.com,prateek@dtrl.drdo.in, ravinder.ndma@gmail.com,uday@iitmandi.ac.in, varun@iitmandi.ac.in

**Abstract.** Landslides and associated soil movements (debris-flow) are the common natural calamities in the hilly regions. In particular, Tangni in Uttrakhand state between Pipalkoti and Joshimath has experienced a number of landslides in the recent past. Prior research has used certain machine-learning (ML) algorithms to predict landslides. However, a comparison of ensemble and non-ensemble ML algorithms for debris-flow predictions has not been undertaken. In this paper, we use ensemble and non-ensemble machine-learning (ML) algorithms to predict debris-flow at the Tangni landslide. Non-ensemble algorithms (Sequential Minimal Optimization (SMO), and Autoregression) and ensemble algorithms (Random Forest, Bagging, Stacking, and Voting) involving the non-ensemble algorithms were used to predict weekly debris-flow at Tangni between 2013 and 2014. Result revealed that the ensemble algorithms (Bagging, Stacking, and Random Forest) performed better compared to non-ensemble algorithms. We highlight the implications of predicting debris- flow ahead of time in landslide-prone areas in the world.

**Keywords:** Landslides, Sequential Minimal Optimization (SMO), Autoregression, Random Forest, Bagging, Stacking, Voting.

## 1 Introduction

Landslides and soil-movements (debris-flow) are caused due to the rain, mostly in the monsoon season in hilly areas [1]. These landslides and debris-flow are

natural hazards that often happen without warning and cause massive loss of property and life across the world [2]. Landslides are a major problem in India with a staggering 11,000 deaths over the past 12 years in the country [3]. Machine–learning (ML) algorithms, which allow prediction of future outcomes based upon historical data, hold the key for timely alerting people about debris-flow and impending landslides [4]. As mentioned above, the landslide debris-flow causes damages to lives and property [5], and ML algorithms could be utilized in predicting debris–flow in landslide-prone areas [6]. Here, ML algorithms could learn patterns in data collected by sensor systems that are installed in different landslide-prone locations [7].

Although there is a large class of ML algorithms that have been proposed for making predictions about debris-flow [15-23], a comparison of these algorithms with ensemble versions of existing algorithms have been less investigated. In this paper, we use several ensemble and non-ensemble ML algorithms for predicting debris-flow ahead of time. For example, different non-ensemble algorithms (SMO [8], Autoregression (AR) [9]), and ensemble algorithms (Random Forest (RF) [10], Bagging [11], Stacking [12], and Voting [13]) have been used in the past for time series forecasting. SMO is a variant of the support-vector machine (SVM) algorithm that optimizes the training of SVM for regression problems . Autoregression is mostly used for predicting the values of variables based upon the prior values of the same variables. Random Forest, an ensemble algorithm, produces its predictions using a collection of decision trees with different feature sets. Bagging, bootstrapping and aggregation, uses subsamples from a dataset with replacement and creates predictive models from training on different subsamples . Stacking is combining the predictions of multiple models into a single model that makes the final prediction . Voting algorithm is used for combining the prediction of different ML algorithms by averaging their predictions to generate the time-series forecasts .

The primary goal of this paper is to evaluate non-ensemble ML algorithms (SMO and Autoregression) with ensemble versions of these algorithms (Random Forest, Bagging, Stacking, and Voting) to predict debris-flow over time at a real-world landslide location. Specifically, we use weekly debris-flow sensor data collected at the Tangni landslide between 2013 and 2014 and use it to predict debris-flow one-week ahead of time.

In what follows, we first detail the background literature concerning the use of ML algorithms for soil-movement predictions. Next, we detail the study area and data used for our time-series forecasting. Then, we detail different ensemble and non-ensemble methods that we developed on debris-flow data. Finally, we detail our results and discuss the main implications of our results for debris-flow time-series predictions in the real world.

## 2 Background

Prior research has explored landslide susceptibility mapping and debris-flow prediction through ML algorithms [14-22]. Some researcher has used support-vector-

based models for debris-flow predictions [14–18]. For example, reference [14] predicted the slope displacement in the Three Gorges Reservoir, China, using a Particle Swarm Optimization and Support Vector Machine (PSO–SVM) coupling model. Results revealed that the proposed PSO–SVM model could represent the relationship between the causal factors and the cyclic slope displacements. Similarly, reference [15] have found that the Gaussian process performed better than the simple artificial neural network (ANN) models, relevance vector machine, and support vector machine.

Furthermore, reference [16] used a case-study of landslides in the Ecuadorian Andes and compared the prediction power of support vector machines, logistic regression, and bootstrap-aggregated classification trees (Bagging and Double-Bagging). Results revealed that logistic regression with stepwise backward variable selection produced the lowest error rates and provided the best generalization capabilities. Next, reference [17] compared a Least Square Sup- port Vector Machine (LSSVM) model boosted with Genetic Algorithm, namely GA-LSSVM, with a Double Exponential Smoothing (DES) and LSSVM model to empirically forecast landslide displacements. Some researchers have also found that a Support Vector Machine (SVM) regression predicted debris-flow in Baishuihe landslide in Three Gorges Reservoir Area, China, with a small error [18].

Reference [19] have used an autoregressive (AR) model and a detrended fluctuation analysis (DFA) method to model debris-flow. The coefficient and variance of AR and the scaling exponent of DFA were estimated using a slide window. Similarly, reference [20] used autoregressive moving average time-series models to analyze the autocorrelation of landslide triggering factors.

Some researchers have also used tree-based models for predicting debris-flow [21]. For example, reference [21] showed that the Random Forest model was capable of predicting the evolution of daily slope displacements over a 30-day period.

Reference [22] used a hybrid bagging-based method for the prediction of landslides at the district of Mu Cang Chai, Vietnam. In this study, 248 past landslides and fifteen geo-environmental factors were considered for the model construction. Based on the AUC values, bagging outperformed the SVM and NBT models. Thus, ensemble methods like bagging may provide promising methods for landslide debris-flow predictions.

## 3 Study Area

The study was performed in on the Tangni landslide in Chamoli district of Uttarakhand, India. The study area covers an area of 0.72 $km^2$. It is located on the northern Himalayan region at latitudes 30° 27' 54.3" N and longitudes 79° 27' 26.3" E, at an altitude of 1450 meter (Fig. 1A and 1B). As seen in Fig. 1B, the landslide is located on National Highway 58, which connects Ghaziabad in Uttar Pradesh near New Delhi with Badrinath and Mana Pass in Uttarakhand. The geology of this area consists of slate and dolomite rocks [21]. The landslide slope is 30° above the road level and 42° below the road level. There have been

several prior landslides in this area causing roadblocks and economic losses to tourism [23].

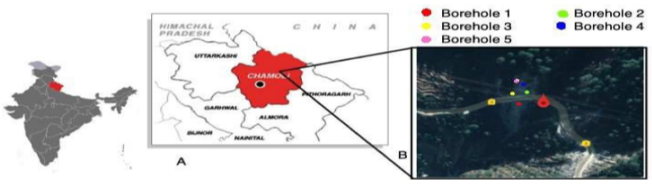



**Fig. 1.** (A) Location of the study area. (B) The Tangni landslide on Google Maps.

Data was collected from the Tangni landslide at a daily scale between 1st July 2012 and 1st July 2014 across five different boreholes (BH). These five boreholes are represented by different colors in Fig. 1B (red – borehole 1, green – borehole 2, yellow – borehole 3, blue – borehole 4, and pink – borehole 5). Each borehole contained five sensors at different depths (3m, 6m, 9m, 12m, and 15m). Data from some of these five boreholes was used for evaluating different ensemble and non-ensemble machine-learning methods.

## 4 Methodology

### 4.1 Data Pre-Processing

Data from Tangni landslide in Chamoli, India was obtained from the Defence Terrain Research Laboratory, Defence Research and Development Organization. The monitoring system in each of the five boreholes at the Tangni landslide contained inclinometer sensors at different depths (3m, 6m, 9m, 12m, and 15m). These sensors measured tilt in mm per m units (essentially the angle the inclinometer tilts). Each inclinometer sensor was a 0.5-meter long sensor that was installed vertically at different depths in a borehole. The monitoring system at Tangni landslide had five sensors per borehole across five boreholes. Thus, in total there are 25 sensors across 5 boreholes. Fig. 2 shows the inclinometer sensor installed in its casing at a certain depth. As shown in Fig. 2, if there was a tilting movement ($\theta$) of the inclinometer of length L, then the horizontal displacement in the tilting direction was $L \sin \theta$. For better understanding, we converted the displacement in mm per m units into a $\theta$ angle (degrees), where 1mm/m displacement equalled 0.05729° . First, we calculated the relative tilt angle of each sensor from its initial reading at the time of installation. Second, we chose only those sensors from each borehole that gave the maximum average tilt angle over a two-year period. Thus, the data was reduced to five time-series, where each time-series represented the relative tilt per borehole from the sensor that moved the most in the borehole across the two-year period. As the daily data was sparse, we averaged the tilt over weeks to yield 78 weeks of average

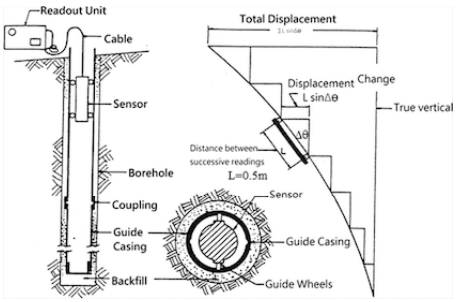



**Fig. 2.** Inclinometer sensor installed in its casing at a certain depth.

tilt data per time-series. Fig. 3A-3E represent the average relative tilt per week from five sensors across five boreholes (one sensor per borehole) that caused the maximum average tilt across 78 weeks. These five time-series were used to compare the ensemble and non-ensemble methods.

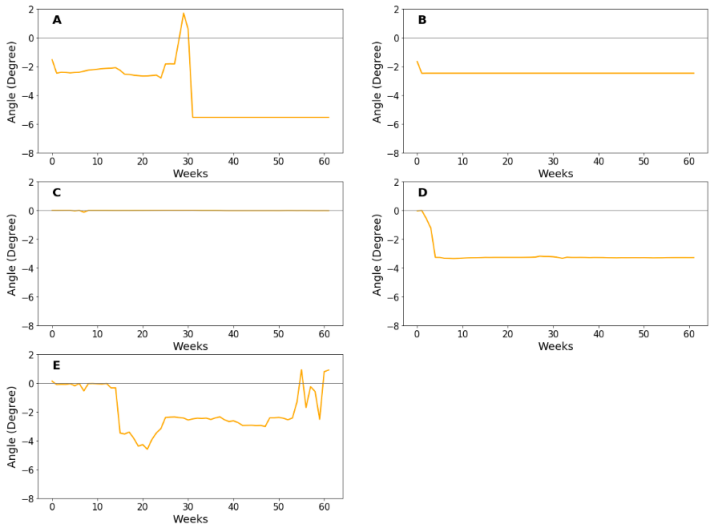



**Fig. 3.** Average tilt angle in degrees across five sensors (one per borehole). (A) Borehole 1 and 3m sensor. (B) Borehole 2 and 12m sensor. (C) Borehole 3 and 6m sensor. (D) Borehole 4 and 15m sensor. (E) Borehole 5 and 15m sensor.

By convention, a negative tilt angle was a downhill motion and a positive tilt angle was an uphill motion. As seen in Fig. 3C, a downhill motion starts from -0.11° in the $73^{rd}$ week and suddenly becomes larger (-4.4°) in the last four weeks. The data was split in an 80:20 ratio (sixty-two weeks for training and the last sixteen weeks for testing) across different machine learning algorithms.

### 4.2 Sequential Minimal Optimization

John Platt proposed the sequential minimal optimization (SMO) in 1998 [8]. It is a widely-used algorithm for solving a quadratic programming (QP) problem that arises during the training of support vector machines. The goal of the SMO algorithm is to return alpha parameters (Lagrange multipliers) that satisfy the following constraint optimization problem:

$$min_{\alpha} \sum_{i} \sum_{j} \alpha_i \alpha_j y_{ij} X_i . K(X_i, X_j) - \sum \alpha_i \tag{1}$$

For a Kernel function:

$$K(X_i, X_j) = \phi(X_i).\phi(X_j) \tag{2}$$

and

$$s.t. \sum_{i} \alpha_i y_i = 0, \alpha_i \in [0, C] \tag{3}$$

### 4.3 Autoregression

Autoregression (AR) is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. This technique can be used on a time-series where input variables are taken as observations at previous time steps, called lag variables. For example, we can predict the value for the next time step (t+1) given the observations at the last two time-steps (t-1 and t-2). As a regression model, this would look as follows ($X_{t+1}$ is the movement to be predicted using the movements ($X_{t-1}$ and $X_{t-2}$) :

$$X_{t+1} = \beta_0 + \beta_1 . X_{t-1} + \beta_2 . X_{t-2} \tag{4}$$

### 4.4 Random Forest

Random Forest algorithm was developed by Leo Breiman and Adele Cutler [9]. Random Forest or random decision forests are an ensemble learning method for classification and regression. At the time of training, the Random Forest algorithm output is the majority class (classification) or the value that is the mean of the prediction of individual trees (regression). By aggregation, the Random Forest algorithm corrects the problem of over fitting in decision trees [10].

### 4.5 Voting

Voting is perhaps the simplest ensemble algorithm and it is often very effective [13]. Voting can be used for classification or regression problems. Voting is one of the ensemble methods that allows to improve accuracy by means of combining several base models. There are several possible ways to organize voting procedure: majority voting, average of probabilities, median of probabilities, etc. In this paper, we consider the average of probabilities model as it is the default assumption in several machine-learning tools.

### 4.6 Bagging

Bagging is a machine learning ensemble Meta algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression [11]. Bagging uses subsamples from the dataset with replacement and trains the predictive model on these subsamples. The final output model is averaged across all models for the better result.

### 4.7 Stacking

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor [12]. In Stacking, the algorithm takes the outputs of sub-modes as inputs and attempts to learn how to best combine the input predictions to make a better output prediction.

### 4.8 Optimization of Model Parameters

**Sequential Minimal Optimization (SMO).** SMO algorithm has two parameters. The first parameter is the complexity parameter (C) that is used to build a "hyperplane" between two classes, which are used for classification, regression, or other tasks. The C parameter controls how many instances are used as "support vectors" to draw a linear separation boundary in the transformed Euclidean feature space. The second parameter of the SMO algorithm is an exponent (E) of the kernel function. We varied the C and E parameters in SMO as per the following: C=0, 1 and E=1, 2, 3, 4 for polynomial kernel; C=0, 1 and E=1, 2 for normalized polynomial kernel; and, C=0 and E = 1 for RBF kernel.

**Random Forest.** In Random Forest, a key parameter is the number of features (nF) to consider in each split point. In this paper, we varied the nF parameter between 0 and 10 to calibrate the Random Forest model.

**Autoregression.** This algorithm has parameters corresponding to the beta coefficients ($\beta_0$, $\beta_1$, $\beta_2$, ... $\beta_n$) and the last n lag terms (t-1, ..., t-n). We were varied ridge parameter value from 1 to 50 with step size=10 [24]. We tried the two options for attribute selection, M5 and greedy, that is known as the Akaike Information Criterion (AIC) to select features for the linear regression.

**Voting.** In voting, we used SMO, Autoregression, and Random Forest as the sub models. A key parameter in voting is how the predictions of the sub models are combined. In this paper, we have assumed the combination rule to be average of the probabilities, minimum probability, maximum probability, and median.

**Stacking.** In Stacking, we tried SMO, Autoregression, and Random Forest as the sub models. Furthermore, IBk, K Star, LWL were used as the meta-learner. The number of folds used for cross-validation (numFold) were varied between 10 to 100 with step size 10 to find best value of parameters.

**Bagging.** In Bagging, we used SMO, Autoregression, and Random Forest one at a time as a sub model. Furthermore, the number of bags (I) were varied between 10 and 150 with step size of 5 to find this parameter's best value.

## 5 Results

Each algorithm was calibrated to each of the five time-series independently. Table 1 shows the root-mean squared error (RMSE) results of applying ensemble and non-ensemble algorithms, AR, SMO, Random Forest, Bagging Voting and Stacking, on the training data across the five boreholes. As can be seen in Table 1, the Random Forest and Voting (ensemble) algorithms performed the best and second best and better compared to the other algorithms.

**Table 1.** The RMSE of different algorithms in the training dataset.

| **Root-Mean Squared Error (RMSE) in angle ∘** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rank | Algorithm | Borehole 1 3m | Borehole 2 12m | Borehole 3 6m | Borehole 4 15m | Borehole 5 15m | Average |
| 5 | SMO | 0.95 | 0.00 | 0.00 | 0.07 | 0.83 | 0.37 |
| 6 | Autoregression | 1.02 | 0.01 | 0.00 | 0.12 | 0.85 | 0.40 |
| 1 | Random Forest | 0.39 | 0.00 | 0.00 | 0.01 | 0.31 | 0.14 |
| 3 | Voting | 0.96 | 0.01 | 0.00 | 0.11 | 0.52 | 0.32 |
| 4 | Bagging | 0.94 | 0.00 | 0.00 | 0.01 | 0.70 | 0.33 |
| 2 | Stacking | 0.60 | 0.00 | 0.00 | 0.02 | 0.70 | 0.26 |

Table 2 shows the optimized values of different parameters of ensemble and non-ensemble algorithms. For example, in Random Forest algorithm, the Bags (I) = 50 and Number of Features (nF) = 5. Similarly, Bagging used I = 5.

**Table 2.** Best set of parameters.

| **Algorithm** | **Parameters** |
|---|---|
| SMO | C=5, E=1, Polynomial Kernel |
| Autoregressive | Ridge=30, Attribute Selection Method=M5 |
| Random Forest | Bags (I)=50, Number of Features (nF)=5 |
| Voting | Combination Rule=[Minimum of Probabilities] |
| Bagging | Bags (I)=25, Classifiers=[SMO] |
| Stacking | Meta Classifier=[SMO], numFold=50 |

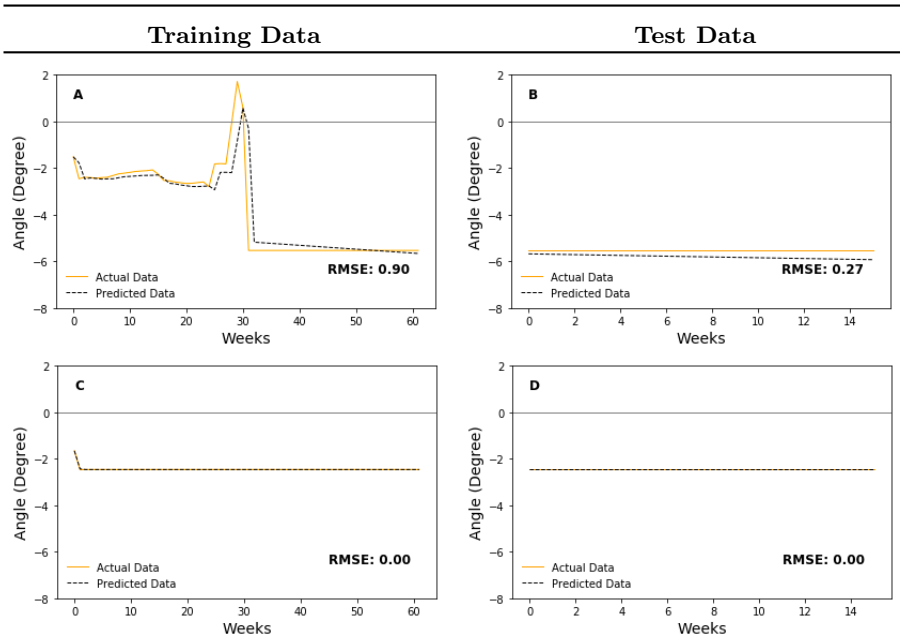
Table 3 shows the RMSEs from different models across different boreholes in the last 16-weeks of test data. As can be seen in the table, Bagging, Random

**Table 3.** The RMSE of different algorithms in the testing dataset.

| Root-Mean Squared Error (RMSE) in angle ∘ | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rank | Algorithm | Borehole 1 3m | Borehole 2 12m | Borehole 3 6m | Borehole 4 15m | Borehole 5 15m | Average |
| 4 | SMO | 0.03 | 0.00 | 1.36 | 0.06 | 1.26 | 0.54 |
| 5 | Autoregression | 0.25 | 0.01 | 1.45 | 0.16 | 1.12 | 0.59 |
| 3 | Random Forest | 0.00 | 0.00 | 1.54 | 0.09 | 1.00 | 0.52 |
| 6 | Voting | 0.03 | 0.18 | 0.58 | 1.23 | 1.29 | 0.66 |
| 1 | Bagging | 0.06 | 0.00 | 0.73 | 0.08 | 1.30 | 0.43 |
| 2 | Stacking | 0.00 | 0.00 | 1.55 | 0.09 | 0.97 | 0.52 |

Forest, and Voting performed the best, second best, and third best among all algorithms.

Fig. 4. shows the fits of the Bagging algorithm to the time-series data across the five boreholes in the training and test datasets. Overall, these results are reasonably good with very small RMSE values.



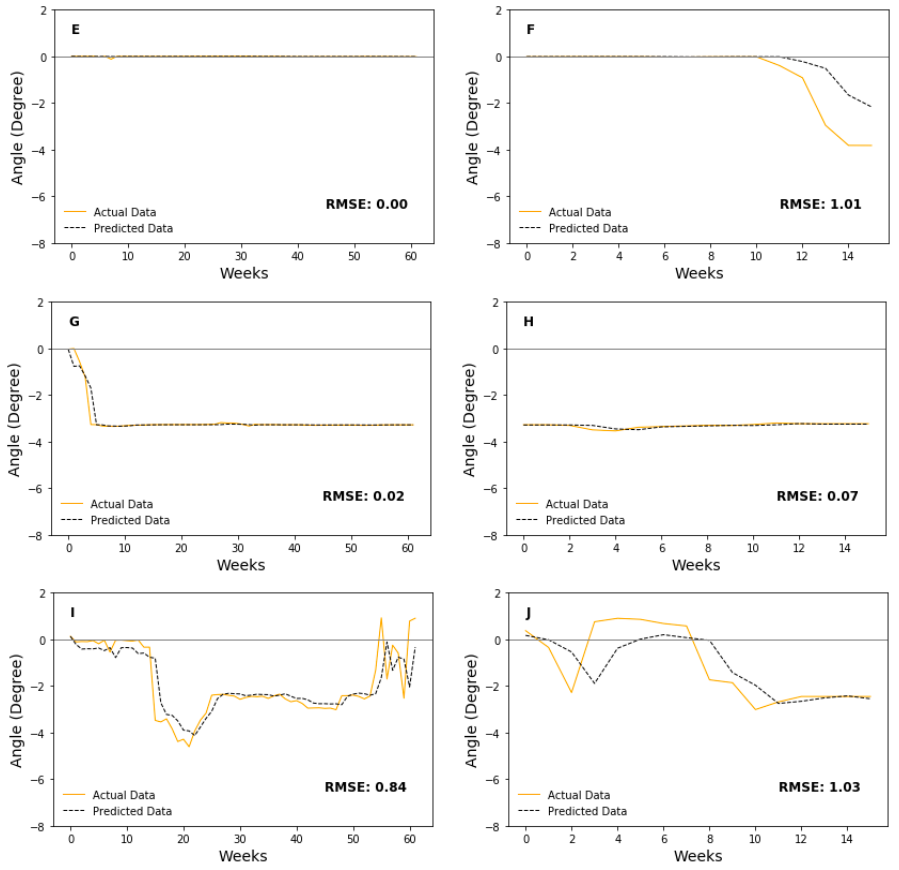



**Fig. 4.** Relative angle (in degree) over training data (62 weeks) and test data (16 weeks) from the best performing Bagging algorithm. (A) and (B): borehole 1, 3m depth. (C) and (D): borehole 2, 12m depth. (E) and (F): borehole 3, 6m depth. (G) and (H): borehole 4, 15m depth. (I) and (J): borehole 5, 15m depth.

## 6 Discussion and Conclusions

A focus of machine-learning (ML) algorithms could be the prediction of debris-flow in advance to timely warn people about impending landslides. In this work, we applied both ensemble and non-ensemble ML algorithms on weekly debris-flow data from the Tangni landslide in Chamoli, India. All models were calibrated on the first 80% of data and tested on the last 20% of data. All models could generate the debris-flow predictions in the following week given the history of movements in prior weeks. Our results revealed that the ensemble algorithms performed better compared to non-ensemble algorithms during both model training and test.

First, we found that the Bagging algorithm performed the best among all algorithms. A likely reason for this result could be that the Bagging algorithm includes two operations as part of its functioning: Bootstrap and Aggregation. Bootstrap involves estimating means from multiple random samples of data with replacement. This mean estimation may be useful especially in cases where we have limited amount of data. In addition, the aggregation involved many bags (= 25), which likely yielded superior performance.

Second, we found that the Random Forest algorithm performed the second best among all algorithms. A likely reason for this result could be that the Random Forest algorithm is an ensemble of several decision trees with varying number of features. Also, this algorithm builds upon the Bagging algorithm and includes features like different bag sizes.

In this paper, we were able to show that both ensemble and non-ensemble algorithms may be used for debris-flow predictions. However, as part of our future research, we plan to extend these analyses to other neural-network-based methods including the use of both artificial neural networks as well as recurrent neural networks (e.g., long short-term memory models). Also, we plan to combine these methods using different ensembling techniques like bagging. Some of these ideas form the immediate next steps in our program on debris-flow predictions using ML techniques.

**Acknowledgments** The project was supported from grants (awards: IITM/NDMA/VD/184, IITM/DRDO-DTRL/VD/179, and IITM/DCoN/VD/204) to Varun Dutt. We are also grateful to Indian Institute of Technology Mandi for providing computational resources for this project.

## References

1. Mathew, D. Babu, S. Kundu, K. Kumar and C. Pant.:Integrating intensity–duration-based rainfall threshold and antecedent rainfallbased probability estimate towards generating early warning for rainfall-induced landslides in parts of the Garhwal Himalaya, India", Landslides, vol. 11(4), pp. 575-588 (2013).
2. Melanie J. Froude and David N. Petley, "Global fatal landslide occurrence from 2004 to 2016", Nat. Hazards Earth Syst. Sci., 18, 2161–2181, 2018.
3. The Weather Channel (October 2018). Why India Needs to be Concerned About Landslides. https://weather.com/en-IN/india/science/news/2018-10-18-why-india-needs-to-be-concerned-about-landslides, last accessed April 7, 2019.
4. Chaturvedi, P., Srivastava, S., Kaur, P. B.: Landslide EarlyWarning System Development Using Statistical Analysis of Sensors' Data at Tangni Landslide, Uttarakhand, India. In: Sixth International Conference on Soft Computing for Problem Solving (2017), AISC, volume 547, pp. 259-270. Springer, Singapore (2017).
5. Parkash, S.: Historical Records of Socio-economically Significant Landslides in India. Journal of South Asia Disaster Studies 4(2), 177-204 (2011).
6. Du, J., Yin, K., Lacasse, S.: Displacement prediction in colluvial landslides, three Gorges reservoir, China. Landslides, vol. 10(2), pp. 203-218 (2013).

7. Mali, N., Chaturvedi, P., Dutt, V., Kala, V. U.: Training of Sensors for Early Warning System of Rainfall-Induced Landslides. In Recent Advances in Geo-Environmental Engineering, Geomechanics and Geotechnics, and Geohazards, pp. 449-452. Springer, Cham (2019).
8. Platt, J. (April 1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/, last accessed April 7, 2019.
9. Ehrmann, M., Ellison, M., Valla, N.: Regime-dependent impulse response functions in a Markov-switching vector autoregression model. Economics Letters, vol. 78(3), pp. 295-299 (2003).
10. Breiman, L.:Random forest, Machine Learning, vol. 45(1), pp. 5-32 (2001).
11. Breiman, L.: Bagging predictors. Machine Learning. vol. 24(2), pp. 123–140 (1996).
12. Ting, K.M., Witten, I.H.: Stacked generalization: when does it work? Hamilton, New Zealand: University of Waikato, Working paper 97/03, (1997).
13. Thornton, C., Hutter, F., Hoos, H. H., Leyton-Brown, K. . Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. pp. 847-855, (2013).
14. Zhou, C., Yin, K., Cao, Y., Ahmed, B.: Application of time series analysis and PSO–SVM model in predicting the Bazimen landslide in the Three Gorges Reservoir, China. Engineering geology, vol. 204, pp. 108-120, (2016).
15. Liu, Z., Shao, J., Xu, W., Chen, H., Shi, C..: Comparison on landslide nonlinear displacement analysis and prediction with computational intelligence approaches. Landslides, vol. 11(5), pp. 889-896 (2014).
16. Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Science, v.5(6), p853-862, 2005.
17. Zhu, X., Xu, Q., Tang, M., Nie, W., Ma, S.: Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: A case study in Sichuan Province, China. Engineering geology, vol. 218, pp. 213-222, 2017.
18. Zhu, C. H., Hu, G. D.: Time series prediction of landslide displacement using SVM model: application to Baishuihe landslide in Three Gorges reservoir area, China. In Applied Mechanics and Materials, vol. 239, pp. 1413-1420, (2013).
19. Pu, F., Ma, J., Zeng, D., Xu, X., Chen, N.: Early warning of abrupt displacement change at the Yemaomian landslide of the Three Gorge Region, China. Natural Hazards Review, vol. 16(4), pp. 04015004 (2015).
20. Li, H., Xu, Q., He, Y., Deng, J..: Prediction of landslide displacement with an ensemble-based extreme learning machine and copula models. Landslides, vol. 15(10), pp. 2047-2059 (2018).
21. Krkač, M., Špoljarić, D., Bernat, S., Arbanas, S. M.: Method for prediction of landslide movements based on random forests. Landslides, vol. 14(3), pp. 947-960, (2017).
22. Pham, B. T., Bui, D. T., Prakash, I.: Bagging based Support Vector Machines for spatial prediction of landslides. Environmental earth sciences, vol. 77(4), pp. 146, (2018).
23. India News (2013, August 13). Landslides near Badrinath in Uttarakhand. https://www.indiatvnews.com/news/india/landslides-near-badrinath- in-uttarakhand-26296.html, last accessed April 7, (2019).
24. Dorugade, A. V.: New ridge parameters for ridge regression. Journal of the Association of Arab Universities for Basic and Applied Sciences, vol. 15(1), pp. 94-99, (2014).

# Unemployment and Poverty as Disordered Social Observables in the Shannon Entropy Theory

Huber Nieto-Chaupis

Universidad Autonoma del Perú
Programa de Ingenierí a de Sistemas
hubernietochaupis@gmail.com

**Abstract.** In this paper we use the well-known Shannon entropy to simulate the identification of areas in Lima city where unemployment and poverty are largely correlated each other. Under the assumption that these social problems are seen as a disordered complex system, the logarithm modeling applies. Our approach is well adjusted to the official data with a discrepancy of order of 7% that supports the fact that the social anomalies are actually entropic systems with a high number of freedom degrees.

## 1 Introduction

Commonly, poverty [1] is understood as the lack of incoming of cash in people whatever their condition. Traditionally, the lack of employment has been seen as a key factor to the departure of a stable economy to one that would present instabilities.

Thus, the term poverty is a common social phenomenon that is still alive in developing countries [2] that are passing through successive phases as to improve their economy indicators.

In large cities, adult people are expected to apply various strategies for searching a kind of adaptability as to be accepted in such societies.

For example, in Latin American societies most of them are rule out by tradition in most cases, so that probably frank rules are not seen in a first instance in those immigrants that perceive their reality as to find both (i) housing and (ii) employment that constitute a double obstacle in the very beginning.

Clearly, the lack of employment, housing, and economy wellness might derive to episodes of poverty that would affect the healthy behavior of the societies [3].

Therefore, unemployment and poverty are a relevant social binomial that cannot be disentangled each other [4]. In particular, the permanent growth of population is perceived as a cause that might land to (i) Social instabilities [5]

(ii) Reconfiguration of the Social-Economy rules

(iii) Possible anomalous jumps of social states

Once these two issues are running together, then it gives rise to the apparition of social states which well defined in families, and these are for example:

(iv) Searching for best quality of life [6]

(v) Security and oscillating emotional Quietness

(vi) Education as windows and prospective issues

(vii) Personal and family progress

All these items listed above (i-vii) can be translated in terms of social equilibrium by which a determined society is committed to carry out a set of rules by expecting a result fully favorable to habitants but with a careful attention to vulnerable population [7].

In this paper, we report an application of the well-known Shannon entropy to identify areas that enclose both poverty and unemployment.

The applied mathematical model might to serve as a robust methodology to identify accurately the geographic areas that would be far away of the expected outcomes of the coherent implementation of an economy model to guarantee an acceptable wellness of the population [8].

We focus our attention to Lima city, with an approximated population of 10M enclosing all types of social layers. Although Lima city has experienced an interesting sustainable economic development as pointed out by the World Bank, as commonly occurs in large Latin American cities, airs of poverty are still perceived.

Turning out to the side of those decentralized theories that anticipate a wellness and healthy social economy from the symmetric dispersion of processes and economic activities in a country but isotropically, geographically speaking [9].

Of particular interest represent those vulnerable social layers that are in a clear disadvantage as to acquire resources for a solid development [10].

In this manner, we expect that the Shannon model [11] is able to measure the social equilibrium that in part is a must for a solid social behavior in such societies oriented to develop their most relevant edges [12].

In second section we present a brief approach to the Shannon model emphasizing the respective probability functions to be used as a model to an accurate identification of the binomial poverty and unemployment.

In third section, the application of the Shannon entropy is done with the assistance of official data. Finally, the conclusion of this paper is presented.

## 2 MODELING THE UNEMPLOYMENT AND POVERTY

A robust way to make predictions in social variables becomes the definition of quantities that are required to build a complete model using the criterion of the Shannon entropy. Consider that the total population of Lima city L in according to official data being the most recent as declared in [13] with the number of districts ND. Then the district population is given by the simple fraction:

$$d = \frac{L}{N_D} \tag{1}$$

When the estimations are not accurate enough, we can implement an intrinsic error in the manner,

$$d = \frac{L + \Delta L}{N_D + \Delta N_D} \tag{2}$$

And the rate of unemployment is given by

$$u = \frac{d}{L} U \tag{3}$$

with U the total number of employment people in Lima city. We remark that the unemployment is not a fix number, by contrary it evolves in time. Again, (3) can be statistically improved when the errors are fully incorporated, thus

$$d = \frac{L + \Delta L}{N_D + \Delta N_D} \times \frac{U + \Delta U}{L - \Delta L} \tag{4}$$

Aside the poverty population is then given by

$$p = \frac{\ell}{L} \tag{5}$$

where $\ell$ is an approximate number or estimated for all those that are directly belonging to the lowest social layers in terms of incomes per month. Now we define the relation between poverty and unemployment given by

$$p \times u = \frac{udU}{L^2} \tag{6}$$

From equation (4) we can argument that poverty is proportional to unemployment, however the fact that is inverse to L2 while the population do increase then poverty can be minimal, based on the fact that this growth might be advantageous as to create new working places. Based entirely in (6), poverty can be written as

$$p = \frac{udU}{(L + Q)^2} \tag{7}$$

with $Q$ is the total population that are under the transition of being employed to be unemployed in Lima city, with previous equations we can test a probability distribution function and it reads

$$p = \frac{d(u + \eta L)U}{1 + (L + Q)^2} \tag{8}$$

with $\eta$ a parameter that measures the fraction of the total population that becomes unemployed in a period of time.

### 2.1 THE SHANNON ENTROPY

With Eq. (8) in hands, we can propose a universal Shannon entropy such as

$$\mathbf{E} = -s\mathrm{Log}[p(L, \eta)] = -s\mathrm{Log}\frac{d(u + \eta L)U}{1 + (L + Q)^2}$$

$$= -s\text{Log}[d(u+\eta L)U] + s\text{Log}[1+(\text{L}+\text{Q})^2]$$

$$= -s\text{Log}d - s\text{Log}U - s\text{Log}(u+\eta L) + 2s\text{Log}(L+Q) \tag{9}$$

And $s$ that denotes a probability. Equation (9) can be accepted as a density of probability, so that the full probability is then

$$P(\eta, Q) = \int p(L,\eta,Q)dL = \int \frac{d(u+\eta L)U}{1+(L+Q)^2}dL \tag{10}$$

that can be evaluated in a closed-form and resulting in

$$P(\eta, Q) = Ud[u-\eta qQ]\text{Atan}(qQ+L) + \frac{\eta}{2}\text{Log}(1+(qQ+L)^2) \tag{11}$$

**INTERPRETATION OF EQ.11**

Once integrated (9) the interpretation that has been adopted is as follows:

- The quantity E measures the equilibrium

- Equilibrium in the present context has a meaning: the conjunction of unemployment and poverty in simultaneous.

- Thus, the term probability is then perceived as the chance that a certain number of people are as unemployment and in a poverty situation.

In addition Eq.(11) the would denote the probability that a human group under a transition of being displaced to a sector of unemployment. In the same time, poverty is also a fact that would characterize the sample. Eq. (11) is thus fully dependent on Q. In order to manage the flux of population that are under the transition. In this manner the independent variable Q has been multiplied by the integer q in order to difference the growth of population that can vary in time. In Fig.1, up to 4 distributions of probability are shown. While all of them are showing a dip, this can be understood as the boundary of two phase transitions.

## 3 APPLICATION AND RESULTS

The signs of Eq. (11) can be slighted changed in order to test the amplitude of robustness of the Shannon model. In fact, the purpose here is to exploit the flexibility of Eq.11 in the form of

$$P(\eta, Q) = Ud[u+\eta qQ]\text{Atan}(qQ-L) + \frac{\eta}{2}\text{Log}(1-(qQ+L)^2) \tag{12}$$

where clearly the signs have been changed, and the new morphologies are shown in Fig.2. It should be noted the apparition of peaks in spaced values of Q. It is interesting that the peaks fall down with the increasing of people under a transition to surpass the line of poverty. On the other hand, because the probability denotes the conjunction of two circumstances: the people is in a territory of poverty, and the episode of unemployment, the decreasing of the

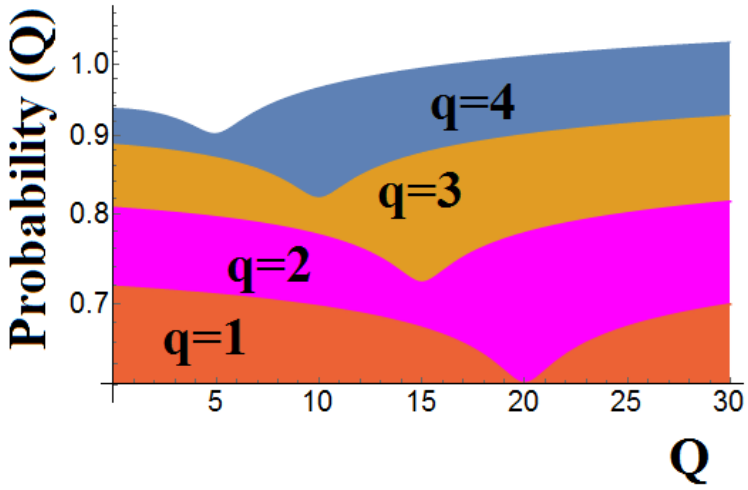



**Fig. 1.** The probability Eq(11) as function of Q for various values of q.

probability is perceived as the apparition of a hidden variable that breaks down the equilibrium that some extent targets to be minimize the risk of people of being unemployment and pass to the poverty state. Under this view we can split Eq(11) to be interpreted as the sum of two probabilities. Thus, in a first instance, the parameter  plays a role against to the probability of apparition of simultaneous event: poverty and unemployment:

$$P_P(\eta, Q) = Ud[(u + \eta qQ)\text{Atan}(qQ - L)]\text{Poverty} \tag{13}$$

$$P_P(\eta, Q) = \frac{\eta}{2} Ud[\text{Log}(1 - (qQ + L)^2)]\text{Unemployment} \tag{14}$$

## MATCHING MODEL TO DATA

In Fig. 3, the official map of urban growth in Lima city is shown. The numbers correspond to the ones as plotted in Fig. 2. The coincidence as to poverty is of order of 73%  7%, whereas unemployment persists in those new areas that enlarges the edges of Lima city, the concurrence with poverty is in the order of 60% derived from Fig.1, and Fig. 2. The coherent implementation of the  parameter, might suggests that actions aimed to restrict the demographic exploitation over the edges, can counteract the progress of unemployment and poverty. For example: to apply well-designed strategies to minimize large periods of unemployment in qualified people, technicians, among others.

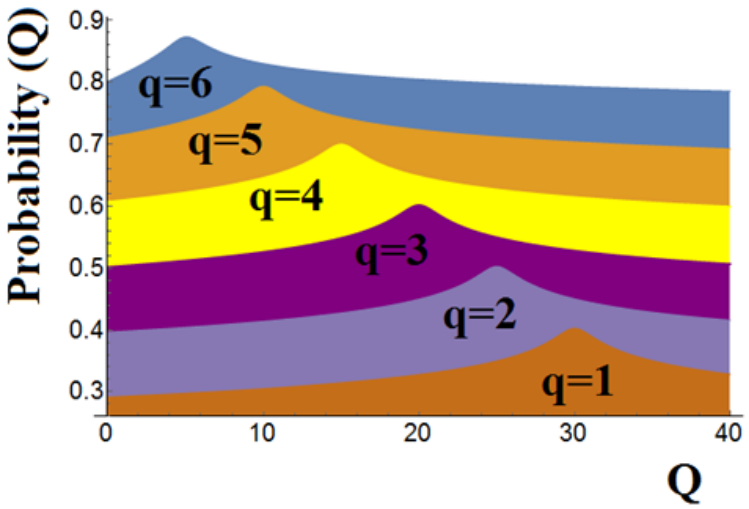



**Fig. 2.** The modified curves of probability as seen in Eq(12) .

## 4 CONCLUSION

In this paper, we have applied the well-known model of Shannon to board the social problem of poverty and unemployment in Lima city under the assumption that these two phenomena are running together. As it is expressed by official data, the predictions given by the Shannon entropy are superimposed to those areas that corresponding to the recent created Peri-urban areas demonstrating to some extent, the robustness of model to be applied to a social problem. The error of model has been of order of 7%, that is derived from the various quantities as shown from equations (1) to (11). In a future work, the relation between unemployment and criminality shall be boarded with the Shannon theory.

## References

1. Mxolisi Makinana: The problem of poverty, unemployment and social exclusion African Security Review, Volume 18, 2009 - Issue 2.
2. Georges Dionne, The effects of unemployment benefits on U.S. unemployment rates: A comment in Review of World Economics (1980).
3. Andria Smythe, Labor Market Conditions and Racial/Ethnic Differences in College Enrollment, Journal of Economics, Race, and Policy (2019).
4. Rainer Eppel, Helmut Mahringer, Getting a lot out of a little bit of work? The effects of marginal employment during unemployment, in Empirica (2019).
5. Koen Caminada, Kees Goudswaard, Chen Wang, Jinxian Wang, Income Inequality and Fiscal Redistribution in 31 Countries after the Crisis, in Comparative Economic Studies (2019).

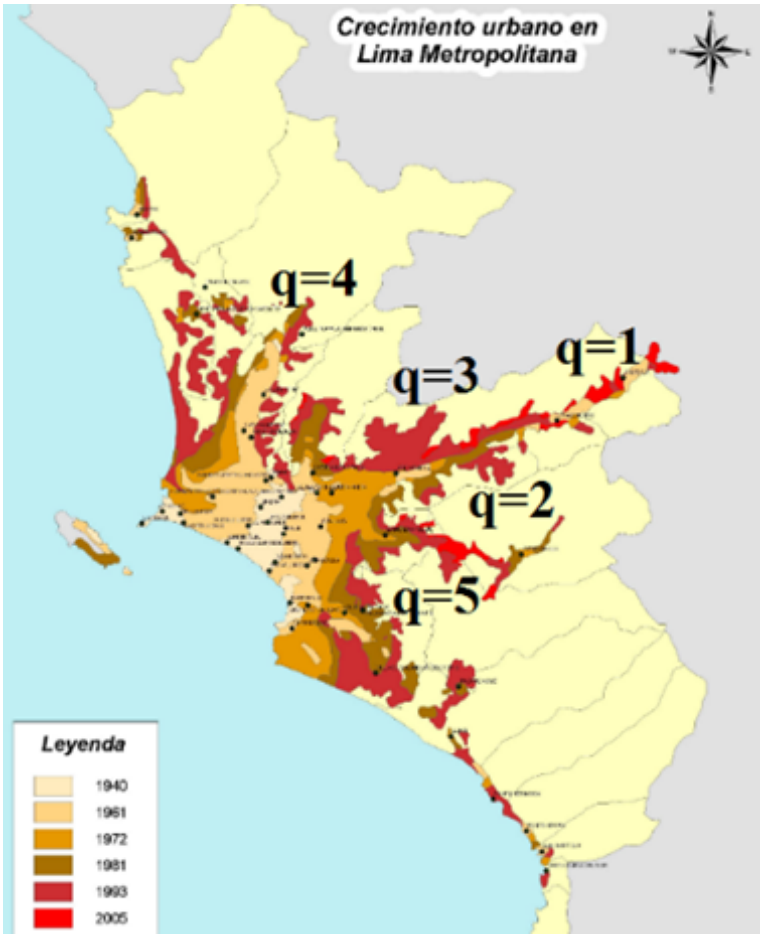



**Fig. 3.** Identification of poverty and unemployment in according to Eq(11) .

6. Sefa Awaworyi Churchill, Russell Smyth, Transport poverty and subjective wellbeing Transportation Research Part A: Policy and Practice, Volume 124, June 2019, Pages 40-54.
7. Cara L. Frankenfeld, Timothy F. Leslie, County-level socioeconomic factors and residential racial, Hispanic, poverty, and unemployment segregation associated with drug overdose deaths in the United States, 20132017 Annals of Epidemiology, In press, corrected proof, Available online 30 April 2019.
8. Sabina Scarpellini, M. Alexia Sanz Hernndez, Jos M. Moneva, Pilar Portillo-Tarragona, Mara Esther Lpez Rodrguez, Measurement of spatial socioeconomic impact of energy poverty. Energy Policy, Volume 124, January 2019, Pages 320-331.
9. Anita Haataja, Unemployment, employment and povert European Societies, Volume 1, 1999 - Issue 2.

10. Zhihong Qian and Tai-Chee Wong, The Rising Urban Poverty:A dilemma of market reforms in China Journal of Contemporary China, Volume 9, 2000 - Issue 23.
11. Huber Nieto-Chaupis, Identification of the Social Duality: Street Criminality and High Vehicle Traffic in Lima City by Using Artificial Intelligence Through the Fisher-Snedecor Statistics and Shannons Entropy, IEEE International Smart Cities Conference (ISC2), Year: 2018.
12. David W. Williams, Poverty and unemployment traps and trappings. The Journal of Social Welfare Law, Volume 8, 1986 - Issue 2.
13. https://www.mef.gob.pe/es/mapas-de-pobreza.

# Spatial integration of agricultural markets in the EU: Complex Network analysis of non-linear price relationships in hog markets.

Christos J. Emmanouilides

College of Engineering and Technology, American University of the Middle East, Kuwait
christos.emmans@aum.edu.kw

Alexej Proskynitopoulos

Department of Operations, Kellogg School of Management, Northwestern University, USA
alexej.proskynitopoulos@kellogg.northwestern.edu

**Abstract.** We present work in progress on the spatial price causality structure between the hog markets of 24 European countries using weekly time series data from 2007 to 2018 and non-linear Granger causality. The EU hog market is studied as a dynamic complex network of linkages between prices in member states. We investigate the temporal development of the spatial network of price relationships, and through the dynamics of its major structural characteristics we draw insights about the horizontal agricultural market integration process in the EU. Of particular interest is the evolution of the degree of market interconnectedness, the strength and reciprocity of price relationships, the development of influential markets (hubs) and of market clusters with strongly interacting components.

**Keywords:** Price relationships, Non-linear causality, Complex networks

## 1 Introduction

Spatial price relationships are commonly studied in economics to provide empirical insights about the integration of geographically separated markets. In efficient, well integrated markets, price dependencies tend to be strong, reciprocal, diffuse and more homogeneous. On the other hand, in more segregated markets, price relationships may be clustered and exhibit a high degree of heterogeneity. Since its foundation, a major goal of EU's economic policy has been the establishment of a frictionless, more homogeneous common market of commodities and services.

Several authors have studied empirically the horizontal integration of national EU agricultural markets; e.g. Serra et al. (2006), Emmanouilides and Fousekis (2012, 2015) employed tests for long-run price convergence (Law of One Price), while Emmanouilides et al. (2014) and Grigoriadis et al. (2016) used copulas to study price

adfa, p. 1, 2011.

co-movements. All these works have considered small subsets of national markets. Studies of long-run price convergence did not assess the causal structure of price dependence between markets. On the other hand, the copula-based dependence measures employed in the latter works do not reveal any information about the origin of price shocks that give rise to the observed price dependencies and their dynamics. As such, they treat each market in a pair as equi-important in determining the relationship.

To provide a more thorough look into the integration of EU primary commodity markets we include in our study 24 out of the 28 EU member states, excluding three countries having very small size (Malta, Cyprus, and Luxemburg) and Croatia that joined EU very recently (in 2013). We gain insights into the dynamics of price relationships by employing a non-linear Granger causality framework and analyze the whole EU hog market as a complex network of bipartite price linkages. These causal linkages are directional and generally asymmetric, in contrast to copula-based measures that are agnostic about price shocks' origins and directional asymmetries. Causal networks have been used recently to study linkages between financial markets (e.g. Vyrost et al. 2015; Baumohl et al. 2018) but not, to the best of our knowledge, to agricultural markets. Also, our network construction employs a non-linear framework for testing non-causality with clearly better power than the standard linear approach adopted yet. Below we describe briefly our work, together with some of our findings; Section 2 discusses the data and methods, Section 3 the empirical analysis and results, Section 4 offers some conclusions.

## 2 Data and Methods

### 2.1 Data

The data are complete series of weekly wholesale prices for pig animals (euro/100Kg) in 24 EU member states from 1/1/2007 to 29/10/2018, obtained by the European Commission. The series are positively correlated, mostly to a high degree (Pearson correlation coefficients range from 0.28 to 0.98, with a mean of 0.74), indicating that price changes are transmitted between market pairs. As is common empirical practice in studies of market integration, we analyze logarithmic price returns $r_{i,t}=d\ln p_{i,t}$ that de-trend the series from deterministic and stochastic components. $p_{i,t}$ denotes price at country $i$, $i=1,\ldots 24$, in week $t$, $t=1,\ldots 618$.

### 2.2 Filtering

Inference on causality can be sensitive to autocorrelation and ARCH effects that are typically present in price returns series; Autocorrelation might spuriously result in seemingly significant Granger causality between markets and may distort the direction of causality (e.g. Vyrost et al., 2015). Neglected non-stationarities, such as ARCH dependence, any associated volatility clustering or other structural changes, may be manifested as spurious non-linearities in the series of returns (e.g. Hsieh,

1991; Lee et al., 1993; Hiemstra and Jones, 1994; Anagnostidis and Emmanouilides, 2015), and consequently may bias inference.
To deal with these potential problems we filtered each series with an ARMA($m,n$)-GARCH($p,q$) model, using several alternative error distributions that allow for a variety of shape and skewness specifications. With the BIC criterion we selected parsimonious models with orders $m,n,p,q \in \{1,2,3,4,5\}$. Optimal models with any AR and ARCH effects removed from the residuals were retained[1]. Several conditional error distributions were tested and selected on grounds of parameter significance and parsimony, again via BIC.

### 2.3 Non-linear Granger causality networks

Denote with $s_{i,t}$ the standardized innovations of the ARMA-GARCH filtered price returns of market $i$. Causal price linkages between two markets $i$ and $j$ are then established through testing for Granger non-causality in conditional means in the following form

$$H_0: E\left(s_{j,t+1} \mid s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) = E\left(s_{j,t+1} \mid s_{j,t}^{L_j}\right), \quad H_1: E\left(s_{j,t+1} \mid s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) \neq E\left(s_{j,t+1} \mid s_{j,t}^{L_j}\right) \tag{1}$$

where $L_i$, $L_j$ indicate finite lags of the series of the two markets, respectively. Then, two equations for the conditional expectations are involved in testing non-causality, one for each hypothesis in eq. (1)

$$H_0: E\left(s_{j,t+1} \mid s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) = f_j\left(s_{j,t}^{L_j}\right), \quad H_1: E\left(s_{j,t+1} \mid s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) = f_{ji}\left(s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) \tag{2}$$

$f_j(.)$ and $f_{ji}(.)$ can be arbitrary, smooth functions of their arguments. In linear non-causality testing they assume the standard additive linear form. Péguin-Feissolle and Teräsvirta (1999) suggested a linear form including a potentially large number of cross-lag interaction terms as Taylor approximations of $f_j(.)$ and $f_{ji}(.)$.

Here we opt for a more flexible non-linear specification of functions $f_j(.)$ and $f_{ji}(.)$ introduced by Hastie and Tibshirani (1990) and further developed by others (e.g. Wood 2017) in the context of generalized additive models (GAMs). Under this specification, and assuming a gaussian link function relating the conditional mean with the lagged series, eq. (2) become

$$E\left(s_{j,t+1} \mid s_{j,t}^{L_j}\right) = a_{0j} + \sum_{r=1}^{L_j} f_r(s_{j,t-r}) + u_{j,t}, \quad u_{j,t} \sim iidN(0, {\sigma_{u,j}}^2) \tag{3a}$$

$$E\left(s_{j,t+1} \mid s_{j,t}^{L_j}, s_{i,t}^{L_i}\right) = a_{0ji} + \sum_{p=1}^{L_j} f_p(s_{j,t-p}) + \sum_{q=1}^{L_i} f_q(s_{i,t-q}) + \eta_{j,t}, \quad \eta_{j,t} \sim iidN(0, {\sigma_{\eta,j}}^2) \tag{3b}$$

[1] Residuals were tested for the presence of AR and ARCH effects with the Ljung-Box test and the ARCH test of Engle (1982).

Functions $\{f_p, f_q, f_r\}$ are usually specified as non-parametric smooth functions of a single lagged variable. Typical choices are local scatter smoothers (loess), smoothing splines or, as more recently developed, smooth expansions of basis functions chosen from a range of alternative families. Expansion coefficients are estimated together with a set of penalty parameters that regulate over-fitting using a penalized maximum likelihood iterative estimation method such as IRLS with the smoothing parameters determined at each iteration step via cross-validation. The estimation algorithm minimizes the penalized deviance

$$D(\mathbf{a}) + \sum_{m\in\{p,q\},l=i,j} \lambda_m \int f_m''(s_{l,t-m})^2 ds_{l,t-m} = D(\mathbf{a}) + \sum_{m\in\{p,q\},l=i,j} \lambda_m \mathbf{a}^T \mathbf{S}_m \mathbf{a} \tag{4}$$

where set $\{p,q\}$ indicates the full set of lags ($p=1,...,L_j$ and $q=1,...,L_i$), $\mathbf{a}$ is the vector of coefficients to be estimated, $D$ is the deviance, $\lambda_m$ are the penalties and $\mathbf{S}_m$ is a matrix of known parameters calculated by the basis functions and the penalties (for details see Wood et al. 2016, Wood 2017). Selection of optimal lags $L_i, L_j \in \{1,2,...,10\}$ is performed through the use of some information criterion, such a BIC. If the computational cost is too high, penalties can be set to a fixed value, but to the possible cost of not fully explaining non-linearity in dependence. However, to safeguard against this possibility, tests for neglected non-linearity (e.g. the BDS test of Brock et al., 1987) can be applied on the residuals of eq. (3) and accordingly re-adjust the degree of smoothing. Alternatively, the simpler and faster "backfitting" estimation method of Hastie and Tibshirani (1986, 1990) may be preferable.

In the context of GAM, testing for Granger non-causality can be performed with a generalized likelihood ratio test on the estimated models (3); Denote with $L(\hat{\mathbf{a}}_{H_0})$ and $L(\hat{\mathbf{a}}_{H_1})$ the likelihoods of the models (3a) and (3b), respectively, and with $\hat{\mathbf{a}}_{H_0}$, $\hat{\mathbf{a}}_{H_1}$ the corresponding sets of estimated parameters. Then, under the null hypothesis of non-causality and the usual regularity conditions the log-likelihood ratio follows asymptotically an approximate chi-square distribution, $2(\log L(\hat{\mathbf{a}}_{H_1}) - \log L(\hat{\mathbf{a}}_{H_0})) \sim \chi^2_\nu$, with $\nu = df_{H_1} - df_{H_0}$.

Once the Granger causality test is performed for a pair of markets, the Granger Causality Index (GCI), which is based on the Granger-Wald test (e.g. Hlavăckova-Schindler et al. 2007, Geweke 1982), and quantifies the strength of causal influence market $i$ exerts on market $j$, is computed as

$$GCI_{i\to j} = \left(1 - \hat{\sigma}^2_{n,j} \Big/ \hat{\sigma}^2_{u,j}\right) \tag{5}$$

A significant test result indicates the presence of a *directional* link $\{i\to j\}$ between the two markets with a *weight* $GCI_{i\to j}$. Price relationships can be bi-directional, if $\{j\to i\}$ is statistically significant, or not (otherwise), and generally asymmetric as it is expected that $GCI_{i\to j} \neq GCI_{j\to i}$.

The causal network at any time $t$ is defined as a graph $\mathbf{G}_t=(\mathbf{V},\mathbf{E}_t)$, consisting of a set $\mathbf{V}$ of vertices (nodes/markets) and a set $\mathbf{E}$ of directed weighted edges (links). Set $\mathbf{E}$ contains all directional weighted links $\{i\to j\}$ between markets $(i, j)$ for which the causality test gives a significant result.

### 2.4 Network measures

We consider two kinds of measures of network characteristics: measures that characterize (a) the connectivity of individual nodes, and (b) the cohesiveness of the global network.

**Individual node connectivity**

For a directed weighted network, the *in-strength* (or in-degree) of a node *i*, $d^{in}(i)$, is the sum of the weights of all incoming links to the node. In our context, it represents the total causal influence exerted to market *i* from all markets with a statistically significant causal influence to it. Correspondingly, the *out-strength* (or out-degree) of a node *i*, $d^{out}(i)$, is the sum of the weights of all links originating from the node. It quantifies the overall magnitude of a market's causal influence on the whole market system, and as such it may be viewed as a measure of a market's importance in driving other markets' price dynamics.

Another aspect of individual node connectivity refers to the 'importance' of a node with respect to other nodes in the network. A commonly used measure is *closeness centrality*, defined using the shortest path (sequence of edges) or geodesic distance d(*i,j*) between nodes (*i,j*). Closeness centrality quantifies how 'close' a node to the other nodes in the network is. It is defined as the inverse of the total distance of the node from the other nodes.

$$c_{CL}(i) = 1 \Big/ \sum_{j \in \mathbf{V}} d(i,j) \tag{6}$$

It can be readily normalized to range in [0,1] by multiplying with $|\mathbf{V}|-1$. In our context, a high value of closeness centrality would indicate a market that has a high degree of causal one-to-one relationships (i.e. is 'close') with each of several other markets in the system. Markets with high closeness centrality are more connected than markets with low closeness centrality.

**Global network cohesiveness**

Four measures of global network structure are considered; network density, average strength, average shortest path length, and reciprocity. Network *density* is simply the frequency of realized edges (causal links) relative to the total number of possible edges. For a directed graph it is calculated as

$$D = |\mathbf{E}| \Big/ \big(|\mathbf{V}|-1\big)|\mathbf{V}| \tag{7}$$

where |.| indicates set cardinality. The higher its value, the more densely interconnected the market system is. The *average strength*, i.e. the mean total strength (both "in" and "out") of all network nodes reflects the average strength of price linkages at the system level. The higher its value the strongest on average the price relationships between the markets are. The *average shortest path length* is the mean of the shortest path lengths between all market pairs, calculated as

$$\bar{l} = \sum_{i,j \in \mathbf{V}} d(i,j) \Big/ \big( |\mathbf{V}| - 1 \big) |\mathbf{V}| \tag{8}$$

This is a measure of system's price transmission efficiency; the smaller its value, the faster is the diffusion of price shocks in the system. *Reciprocity* is a measure of bi-directionality in causal price relationships. We calculate it as the relative ratio of the number of bi-directional edges over the total number of edges in the directed graph. Higher values correspond to a higher degree of mutual interactions between markets, indicating a higher efficiency in the flows of price shocks and a higher level of market integration.

### 2.5 Temporal evolution

To investigate the dynamics of the network of price relationships we estimated causal networks and network measures for 359 consecutive rolling windows of 5-years width (5×52=260 observations per window) in order to maintain sufficient sample sizes for the causality tests. Other plausible width options were also explored without noticing important qualitative differences in the results.

To empirically test for possible significant structural changes in the series of estimated global network measures we employed generalized M-fluctuation tests (e.g. Zeileis & Hornik, 2007). Sequences of such breaks, if present, may indicate the onset of different stages/regimes in the market integration process, characterized by distinct network market structures.

## 3 Empirical analysis and results

First, logarithmic price returns were tested for unit roots on the unconditional mean with standard ADF and KPSS tests, along with the spectral wavelet test of Nason (2013) that is shown to have good size properties for heavy tailed series as price returns. In all cases the tests did not provide evidence against weak stationarity.

Then, we applied an ARMA-GARCH filter to all series of returns. Models were estimated by maximizing the joint log-likelihood function of the system of equations involved. In most cases a skewed t-Student specification was adequate for the conditional error distribution. Ljung-Box and Engle's ARCH tests indicated absence of any residual AR or ARCH effects.

In the next step, we used the filtered series to (a) estimate for each rolling window and for each market pair the GAM models in eq. (3), (b) conduct the Granger non-causality tests, (c) construct the networks, and (d) calculate the network measures. As we perform a large number (24×23=552) of simultaneous tests to construct a single rolling window network of statistically significant causal links, we apply a Benjamini-Hochberg (1995) adjustment to the *p*-values from the likelihood ratio tests by controlling the false discovery rate (FDR) for our chosen significance level (we use $a=0.05$).

### 3.1 Network measures of individual node connectivity

Statistics for the rolling windows estimates of the individual market connectivity measures are shown in Table 1. Values of closeness are normalized. To summarize coarsely the temporal evolution of connectivity measures for each market we estimated a linear trend and calculated the coefficient of determination $R^2$. Insignificant trends ($\alpha=0.05$, HAC corrected) are underlined. Overall, the most influential market appears to be Germany, followed by Austria, Netherlands, Belgium and Poland with average out-strength values exceeding 2.00. Most linear trends are positive; Poland,

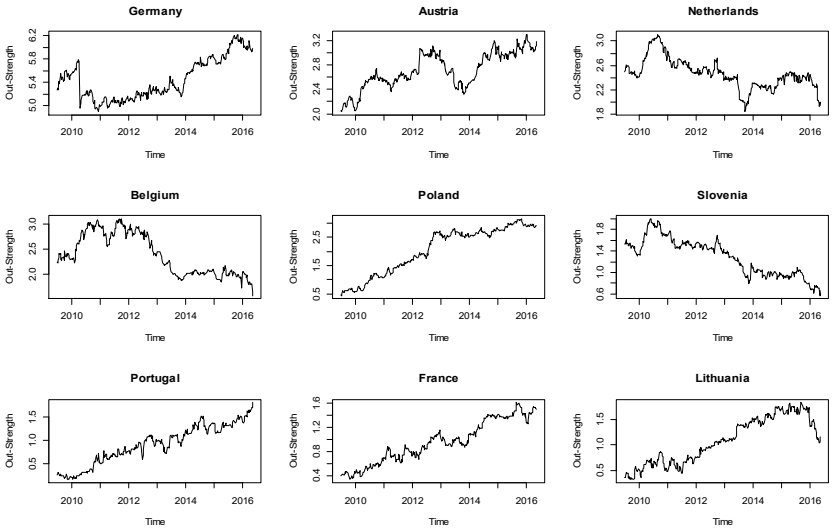
**Table 1.** Measures of individual node connectivity.

| Market | *Out-Strength* | | | | *In-Strength* | | | | *Closeness* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Trend | $R^2$ | Mean | SD | Trend | $R^2$ | Mean | SD | Trend | $R^2$ |
| BE | 2.38 | 0.41 | -0.15 | 0.55 | 1.07 | 0.25 | 0.11 | 0.76 | 0.83 | 0.05 | -0.02 | 0.43 |
| CZ | 0.74 | 0.16 | 0.06 | 0.56 | 2.24 | 0.15 | 0.01 | 0.02 | 0.65 | 0.05 | 0.02 | 0.60 |
| DK | 0.77 | 0.10 | 0.01 | 0.02 | 2.02 | 0.33 | 0.15 | 0.79 | 0.63 | 0.03 | 0.00 | 0.00 |
| DE | 5.45 | 0.35 | 0.12 | 0.50 | 0.85 | 0.16 | -0.04 | 0.22 | 0.84 | 0.04 | -0.01 | 0.50 |
| EE | 0.16 | 0.08 | -0.01 | 0.08 | 2.25 | 0.40 | 0.03 | 0.03 | 0.50 | 0.07 | 0.01 | 0.15 |
| GR | 0.30 | 0.19 | -0.05 | 0.27 | 0.83 | 0.37 | 0.14 | 0.57 | 0.56 | 0.05 | 0.00 | 0.02 |
| ES | 0.90 | 0.22 | 0.08 | 0.52 | 1.06 | 0.56 | 0.26 | 0.88 | 0.65 | 0.05 | 0.01 | 0.07 |
| FR | 0.96 | 0.35 | 0.17 | 0.94 | 0.89 | 0.20 | 0.07 | 0.49 | 0.70 | 0.08 | 0.04 | 0.91 |
| IE | 0.39 | 0.34 | -0.14 | 0.70 | 0.91 | 0.31 | 0.14 | 0.75 | 0.55 | 0.13 | -0.05 | 0.60 |
| IT | 0.30 | 0.14 | 0.03 | 0.21 | 0.36 | 0.15 | 0.05 | 0.36 | 0.55 | 0.04 | 0.01 | 0.42 |
| LV | 0.51 | 0.13 | 0.03 | 0.21 | 1.55 | 0.28 | -0.04 | 0.09 | 0.61 | 0.06 | 0.03 | 0.78 |
| LT | 1.08 | 0.44 | 0.20 | 0.85 | 1.33 | 0.15 | -0.03 | 0.17 | 0.70 | 0.05 | 0.02 | 0.72 |
| HU | 1.30 | 0.30 | 0.14 | 0.84 | 1.45 | 0.24 | 0.03 | 0.05 | 0.73 | 0.08 | 0.03 | 0.84 |
| NL | 2.47 | 0.24 | -0.08 | 0.42 | 0.92 | 0.10 | 0.04 | 0.49 | 0.87 | 0.04 | -0.01 | 0.31 |
| AT | 2.71 | 0.30 | 0.11 | 0.60 | 0.97 | 0.16 | -0.04 | 0.27 | 0.82 | 0.03 | 0.00 | 0.00 |
| PL | 2.04 | 0.82 | 0.39 | 0.91 | 0.86 | 0.35 | -0.13 | 0.55 | 0.80 | 0.08 | 0.02 | 0.33 |
| PT | 0.91 | 0.41 | 0.20 | 0.93 | 1.09 | 0.30 | 0.12 | 0.59 | 0.64 | 0.08 | 0.04 | 0.84 |
| SI | 1.29 | 0.34 | -0.15 | 0.79 | 1.37 | 0.16 | 0.05 | 0.41 | 0.70 | 0.05 | -0.01 | 0.32 |
| SK | 0.68 | 0.10 | 0.00 | 0.00 | 2.42 | 0.22 | 0.03 | 0.07 | 0.64 | 0.03 | 0.01 | 0.12 |
| FI | 0.28 | 0.21 | 0.08 | 0.60 | 0.34 | 0.20 | 0.08 | 0.55 | 0.46 | 0.09 | 0.02 | 0.17 |
| SE | 0.58 | 0.15 | -0.01 | 0.01 | 0.29 | 0.14 | 0.05 | 0.60 | 0.57 | 0.04 | 0.00 | 0.04 |
| UK | 0.24 | 0.17 | -0.04 | 0.20 | 0.31 | 0.12 | -0.03 | 0.24 | 0.52 | 0.06 | -0.02 | 0.22 |
| BG | 0.38 | 0.14 | -0.05 | 0.62 | 1.16 | 0.14 | -0.01 | 0.03 | 0.55 | 0.09 | -0.04 | 0.77 |
| RO | 0.55 | 0.22 | 0.09 | 0.72 | 0.82 | 0.17 | 0.03 | 0.10 | 0.62 | 0.07 | 0.03 | 0.74 |

Portugal, Lithuania, and France have the highest average annual (linear) growth rate, ranging from 0.17 to 0.39; while Slovenia and Belgium had a marked decline in out-strength in the period of study. Figure 1 shows the out-strength evolution for a subset of markets. Slovakia, Estonia, Czech Republic and Denmark appear to have the highest average in-strength values (2.02 or more), indicating that price shocks in these markets were rather driven externally. As might be expected, high out-strength markets tend to have small in-strength and vice versa, indicating a grouping into markets

with high power (or "hubs", Germany, Austria, Netherland, Poland, Belgium) driving price dynamics of markets with lower power (Estonia, Slovakia, Czech Republic, Latvia, Denmark), while the remaining markets appear to interact less strongly as they have rather low average values of both in- and out-strength. In-strength trends tend to be mostly positive, but smaller in magnitude than the out-strength trends. It is worth noting the evolution of interaction strengths in some markets; Over the observation period, Belgium and Ireland appear to have lost power (negative out- and positive in-strength trends), while Denmark and Spain show an increasing exposure to external price shocks (both have dominant negative out-strength trends). On the other hand, the power of Germany and Poland has increased over time (the latter experienced a strong positive out-strength trend with a sizeable negative in-strength trend).

**Fig. 1.** Out-strength temporal evolution for selected markets (5-years rolling windows).




The most connected markets, those with the highest closeness centrality, include markets with high values of interaction strengths and appear to be spatially located in central Europe and to be contiguous; Netherlands, Germany, Belgium, Austria, Poland, Hungary, France, Slovenia, but also Lithuania (average values from 0.7 to 0.87). Markets spatially located to the periphery of Europe have smaller closeness centralities (below 0.60, e.g. Finland, Estonia, UK, Bulgaria, Italy, Ireland, Greece, and Sweden). Linear time trends appear mixed in sign and small in magnitude.

### 3.2 Measures of global network cohesiveness and their evolution

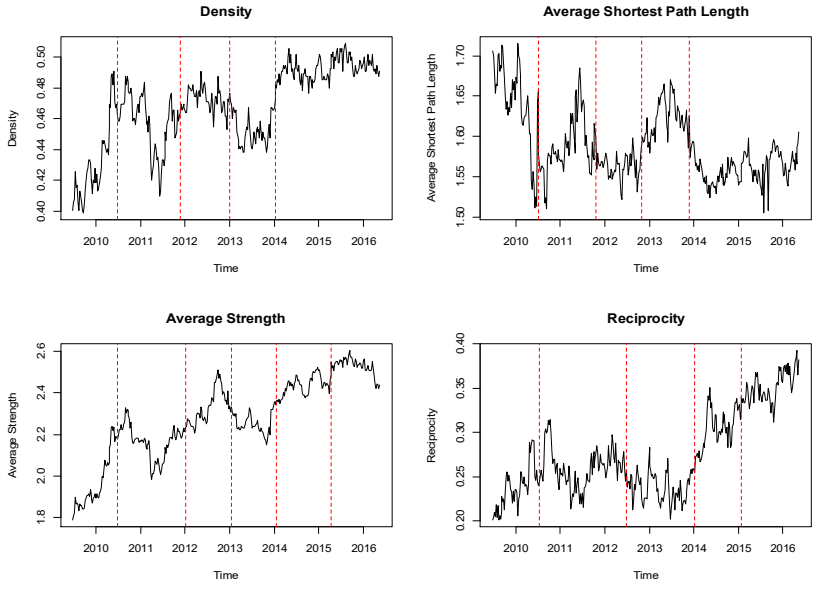
Summaries about the estimated distributions of the network cohesiveness measures over all 5-years rolling windows are given in Table 2. Again, as a rough indicator for their temporal evolution we estimated a linear trend and the corresponding $R^2$. All

trends are significant. HAC consistent generalized M-fluctuation tests detected the presence of breaks in all four series. The time series of the rolling windows estimates are plotted in Figure 2. The estimated break points are shown with dashed lines.

**Table 2.** Measures of network cohesiveness.

| **Statistics** | **Measures** | | | |
|---|---|---|---|---|
| | *Density* | *Average strength* | *Average shortest path length* | *Reciprocity* |
| Min | 0.40 | 1.79 | 1.50 | 0.20 |
| Mean | 0.47 | 2.28 | 1.59 | 0.28 |
| Median | 0.47 | 2.28 | 1.58 | 0.26 |
| Max | 0.51 | 2.61 | 1.72 | 0.39 |
| Std. Dev. | 0.03 | 0.20 | 0.04 | 0.05 |
| Skewness | -0.69 | -0.51 | 0.84 | 0.61 |
| Trend | 0.01 | 0.09 | -0.01 | 0.02 |
| $R^2$ | 0.54 | 0.78 | 0.22 | 0.56 |

**Fig. 2.** Temporal evolution of estimated network cohesiveness measures.




The estimated values of the cohesiveness measures and their temporal evolution indicate that the total market interconnectedness increases over time as the density (proportion of connected market pairs over the total) increases from around 0.40 to a plateau near 0.50. At the same time the interaction strengths between markets also

increase on average, by about 40% from 1.80 to 2.6. The average shortest path length, as might be expected, exhibits reversed time trends relative to the density and average strength, and its value shows an overall decrease (from about 1.68 to a plateau around 1.55), consistent with a shortening of the distance between markets and a more efficient, faster system-wise spread of price shocks. Reciprocity, the proportion of price links that are mutual (two-way causal), is fluctuating initially (mid 2009 to mid 2013) from 0.20 to 0.30, with a local average of about 0.25, and then increases rapidly to reach 0.40 at its maximum. Overall, the evidence points towards higher levels of market integration.

The number of identified breaks is five for average strength and four for the other measures. The first break occurs almost simultaneously for all measures in June/July 2010. It coincides with the leveling of a rapid increase phase for network density, average strength and reciprocity that occur together with a rapid decline of average shortest path length. The second, third and fourth breaks for density, average strength and average shortest path length occur very close the one with the other around the start of 2012, 2013 and 2014, respectively. The third break for reciprocity occurs also around the start of 2014, and the last one in the beginning of 2015, shortly before the last estimated break for average strength. It seems that there is a considerable degree of consistency in the appearance of breaks in the four network cohesiveness measures. These empirical findings provide evidence that the integration process in the EU hog market during the study period is rather a staged process, characterized by regimes within which the network of price relationships has distinct structural characteristics.

## 4 Concluding comments

We presented in this paper a part of an on-going research project on the study of price linkages between spatially separated primary commodity markets in Europe. Our analysis focused on some structural aspects of temporally evolving complex networks of causal relationships in wholesale hog prices. The networks were constructed with the use of GAM-based nonlinear models for testing and quantifying the strength and directionality of causal price relationships in the Granger sense.

The application of network analysis methods provided insights about key characteristics of the complex system of price interactions in the common EU market; Measures of individual connectedness were used to identify groups of markets with high power which have been leading the price transmission process during the study period. Temporal analysis of these measures provided also insights on the changing role of individual markets in the price transmission process. Our data provide evidence not only for a large degree of heterogeneity in market power between countries, but also for the existence of segregation into high and low power groups of markets that are strongly connected to each other. The existence of such groups is an inefficiency of the market system.

The analysis of system-level measures of cohesiveness shed some light into the aggregate market integration process; Results are suggestive of temporal increase in (a) system inter-connectedness, (b) overall strength of price interactions, and (c) preva-

lence of bi-directional price relationships. Also, the length of price transmission paths connecting markets together had been decreasing over time. However slow these changes may be, they all point to a growing degree of market integration in the EU hog market.

## References


1. Anagnostidis, P. and Emmanouilides, C.J. (2015). Nonlinearity in high-frequency stock returns: Evidence from the Athens Stock Exchange. Physica A: Statistical Mechanics and its Applications, 421, 473-487.
2. Baumöhl E., Kočenda, E., Lyócsa S, Výrost, T. (2018). Networks of volatility spillovers among stock markets. Physica A 490 (2018) 1555–1574.
3. Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B 57(1), 289–300.
4. Brock, William A., W. Davis Dechert, and Jose A. Scheinkman, (1987). A Test for Independence Based on the Correlation Dimension, SSRI Working Paper #8702, University of Wisconsin-Madison.
5. Emmanouilides, C.J. and Fousekis, P. (2012). Testing for the LOP under Non-linearity: An Application to Four Major EU Pork Markets, Agricultural Economics, 43(6), 715-723.
6. Emmanouilides, C.J., Fousekis, P., Grigoriadis, V. (2014). Price Dependence in the Olive Oil Markets of the Mediterranean, Spanish Journal of Agricultural Research, 12 (1), 3-14.
7. Emmanouilides, C. J. and Fousekis, P. (2015). Assessing the Validity of the LOP in the EU Broiler Markets", Agribusiness: An International Journal, 31(1), pp. 33-46.
8. Grigoriadis, V, Emmanouilides C.J. and Fousekis, P. (2016), "The Integration of Pigmeat Markets in the EU. Evidence from a Regular Mixed Vine Copula", Review of Agricultural and Applied Economics, 19 (1), 3-12
9. Geweke, J. (1982) Measurement of Linear Dependence and Feedback Between Multiple Time Series, Journal of the American Statistical Association, Vol. 77, No. 378, pp. 304-313.
10. Hastie, T. and Tibshirani, R.J. (1986). Generalized Additive Models. Statistical Science, Vol. 1, No 3, 297-310.
11. Hastie, T. and Tibshirani, R.J. (1990). Generalized Additive Models. Monographs on Statistics and Applied Probability 43, CRC Press.
12. Hiemstra C., Jones, J.D (1994) Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. The Journal of Finance, VOL. XLIX, NO. 5, 1639-1664.
13. Hlavăckova-Schindler, K., Paluš,M., Vejmelka, M., Bhattacharya, D. (2007). Causality detection based on information-theoretic approaches in time series analysis, Physics Reports 441, 1 – 46.
14. Hsieh, D. (1991). Chaos and nonlinear dynamics: Application to financial markets, Journal of Finance 46, 1839-1877.
15. Lee, T.H, White, H. and Clive W.J. Granger (1993). Testing for neglected nonlinearity in time series models A comparison of neural network methods and alternative tests. Journal of Econometrics, 269-290.
16. Nason, G.P. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. J. R. Statist. Soc. B, 75, 879-904.

17. Péguin-Feissolle, A. and Teräsvirta, T. (1999). A general framework for testing the Granger noncausality hypothesis, SSE/EFI Working Paper Series in Economics and Finance 343, Stockholm School of Economics.
18. Serra, T., Gil, J. and Goodwin, B. (2006). Local polynomial fitting and spatial price relationship: price transmission in EU pork markets, European Review of Agricultural Economics 33, 415–436.
19. Výrost, T., Lyócsa S, Baumöhl E. (2015). Granger causality stock market networks: Temporal proximity and preferential attachment, Physica A 427 (2015) 262–276.
20. Wood, S.N., N. Pya and B. Saefken (2016). Smoothing parameter and model selection for general smooth models (with discussion). Journal of the American Statistical Association 111, 1548-1575.
21. Wood, S.N. (2017). Generalized Additive Models: An Introduction with R. Chapman-Hall.
22. Zeileis A., Hornik K. (2007). Generalized M-Fluctuation Tests for Parameter Instability, Statistica Neerlandica, 61, 488–508.

# Comparative Study of Models for Forecasting Nigerian Stock Exchange Market Capitalization.

Basiru Yusuf and Nura Isah
Jigawa State Polytechnic Dutse (Nigeria)

## Abstract

This paper proposes two forecasting models for the Nigerian Stock Exchange Market Capitalization using the Autoregressive Integrated Moving Average (ARIMA) process and an Autoregressive Distributed Lag (ARDL) process. A better model was selected by comparing the forecast evaluation for the estimated models using pseudo-out of sample forecasting procedure over 2013q4 to 2016q3. The statistical loss functions ⟦MAE⟧_(t ) ⟦,RMSE⟧_t and ⟦MAPE⟧_(t ) for the t forecast horizon (t=1, 2, . . . , 12) are used to compare the forecast performance of the two estimated models. The results show that ARIMA model outperform ARDL model in three to four quarters forecast horizon. On the other hand, ARDL model outperform ARIMA in one to two quarters, five to seven quarters as well as nine to twelve quarters forecast horizon. Therefore, in forecasting Nigerian Stock Exchange Market Capitalization in both short and long horizons, it can be concluded that ARDL is better model to be used.

# Models predicting corporate financial distress and industry specifics

Dagmar Camska[1]

[1] Czech Technical University in Prague, MIAS School of Business, Kolejni 2637/2a, 16000 Prague 6, Czech Republic
dagmar.camska@cvut.cz

**Abstract.** This paper is focused on tools predicting corporate financial situation. There have been constructed plenty of models whose aim is to predict possible corporate default or distress. These models will be examined. Traditionally analyses would be focused on the explanatory power or models' accuracy. The aim of this paper is different. Although the models can be mainly used generally there are many specifics which affect results and gained conclusions. The specific highlighted in this paper is an industry branch. Companies operate in different industry areas which influence their performance and overall financial results and ratios and therefore it has an impact on the models' result. The paper works with three industry branches: Manufacture of fabricated metal products, except machinery and equipment (CZ-NACE 25), Manufacture of machinery and equipment (CZ-NACE 28) and Construction (CZ-NACE F). The results will be based on three data sample, specifically financial healthy companies 2012, insolvent companies 2012 and companies 2017. The results of different models predicting financial distress will be computed and compared. The main tools of descriptive statistics will be applied. It should prove or disapprove if industry specifics influence the models significantly.

**Keywords:** Corporate Financial Health, Possibilities of Prediction, Czech Republic.

## 1 Importance of Prediction

Models predicting corporate financial distress or bankruptcy models became on one hand popular for practical use and on the other hand serious research issue for academics in late 1960s. The beginnings are connected with works of Altman [2] and Beaver [5]. The prediction models provide a quick inexpensive answer for their users about financial situation of an analyzed company. The user can be in different position to the company, as in a role of a supplier, customer, financial institution, government etc. It is necessary to monitor the partner's financial situation in business relations because a financial unhealthy or unstable partner can threaten your own business. The prediction models offer their functioning to avoid this unfavorable situation.

There open many research questions connected to models predicting financial distress or default. First there have been constructed plenty of models whose aim is fulfilling the aforementioned discussed purpose. There can be found many researches concentrating on the explanatory power of the used models or new approaches with higher accuracy in comparison to the previous prediction models. Only in the area of the Czech Republic the following papers can be mentioned [19], [15], [13], [6] or [23]. Second question discussed especially in the time of economic recessions is the influence of the overall economic conditions on the results provided by the prediction models (specifically in [4], [16] or [18]). Third question stays partially hidden. It is a specific of an industry sector. An advantage of the prediction models is their general use, at least at the area of manufacturing. It can be raised a question if the industry branch influences the results of the prediction models. Companies can belong to the different industry branches and therefore they have not similar achieved performance and they do not give the similar values of the financial ratios which create bases of the prediction models. This paper focuses on the impact of the industry branch. The conducted analysis will prove or disapprove if industry specifics influence the models significantly.

## 2 Models according to Literature Review

Since 1960's there have been created many models predicting financial distress. Some have high accuracy but others should not be used at all because they do not provide relevant information for business decision making. It is the reason why the models are repeatedly tested in the case of their accuracy. Papers usually focus on one model or maximally 5 are tested. An exception is a research provided by Čámská [6] or [7] which tested almost 4 dozens of the prediction models. The results show that although many models have high explanatory power many models should be also excluded because they reach high level of errors. This paper is based on the models which were classified in the category with high accuracy. The further analysis will use financial data of Czech companies and therefore the emphasis is on Czech models predicting financial distress and the models created in economies with comparable conditions.

The Czech Republic is represented by IN01 [22], IN05 [21] and Balance Analysis System by Rudolf Doucha [9]. These approaches were created on roots of the models coming from the developed economies as Altman Z-Score [3], Bonita Index (in the German original Bonitätsanalyse [24], Kralicek [17] or Taffler [1]. They are still popular in the Czech Republic. On the other hand, it must be noted that these international models originated in countries with different history, development and level of economic and therefore it is still highly debatable if they should be used in the conditions of the Czech Republic. Their accuracy in the case of the Czech enterprises was proved by [6].

The Czech Republic belonged to the transition economies in 1990's and therefore it opens a possibility to test the models which originated in the countries with comparable historical, political and economic development. Therefore it will be introduced the models from the transition countries as Poland, Hungary and Baltic States (especially

Latvia and Lithuania). The introduced Polish models are Prusak, PAN-E, PAN-F, PAN-G, D2, D3, (all previous discussed in [14). The Hungarian model is called according to its authors Hajdu & Virág [10]. The Baltic approaches are represented by Šorins & Voronova [11], Merkevicius [20] and R model [8].

Due to paper page range the models' formulas are not mentioned in this paper. The relevant literature has been provided and therefore the formulas can be found in the case of the reader's interest.

## 3 Research Idea and Data

This chapter is mainly dedicated to the paper's idea and used data. The first subchapter focuses on the solved research question and methods which will be applied. The second subchapter defines the data sample and source of the data. The size of the data sample is discussed as well.

### 3.1 Paper's idea and used methods

The paper's idea is based on that the models predicting financial distress are used generally and they usually do not reflect any industry specifics. The companies belonging to the different industry branches can achieve different performance which influences the values of financial indicators and therefore also the prediction models. It remains a question if this influence is significant or not. There can be tradeoffs between the financial indicators. The value of one ratio is worse but the other one is better for that industry branch and at the end the results are comparable among the industry sectors.

The bankruptcy models or models predicting financial distress introduced in the previous part will be applied to the corporate data. The main descriptive statistics will be computed from the results for the individual companies. These statistics will be compared among the industrial sectors. It will prove or disprove if there are significant differences among the industry branches in the case of final values of the models predicting financial distress.

### 3.2 Data Sample

The paper's idea is based on the differences among industry sectors therefore the data sample has to contain companies belonging to the different industry branches. There were chosen three industry branches, specifically Manufacture of fabricated metal products, except machinery and equipment (CZ-NACE 25), Manufacture of machinery and equipment (CZ-NACE 28) and Construction (CZ-NACE F). These industries provided the largest data sample of insolvent companies caused by the last global economic crisis and they formed the basis of the previous research [6] or [7]. On the first look, the industry branches look similar because their purpose is to produce a final physical good for a customer. It would not be expected that they significantly differ in financial characteristics in comparison with wholesale trade and services.

The data sample itself consists of three subsamples. The first two subsamples are connected with the previous research because they describe companies which became insolvent in 2012 according to the Czech Insolvency Act and companies which were classified as creating economic value added in 2012 (according to [12]). The third subsample is the largest because it contains general companies (without any restrictions) belonging to the particular industry branch and it describes the financial situation of these companies in 2017. The year 2017 is the last for which there are available financial statements. The year 2018 will be still published during this year (2019). The table describes the sample distribution between the particular industry branches.

**Table 1.** Size of the analyzed sample

| Industry branch | Insolvent 2012 | Healthy 2012 | General 2017 |
|---|---|---|---|
| CZ-NACE 25 | 36 | 383 | 1525 |
| CZ-NACE 28 | 10 | 33 | 789 |
| CZ-NACE F | 38 | 229 | 3564 |

The data was obtained from the corporate database Albertina. Although there are more companies belonging to the each studied industry branch in the Czech Republic the sample size is final. It is caused by two limitations. First the companies do not publish their annual statements although it is an obligation. Second some data was incomplete or they contained some items equal to zero and therefore it was not possible to compute all ratios of the prediction models. Such companies were omitted.

It must be noted that the analysis will not work with the time comparison. Although there is data available for 2012 and 2017 these samples are incomparable. The sample 2012 is strictly polarized because it contains on one hand the insolvent companies and on the other hand the companies with the highest level of performance because they created positive economic value added. The sample 2017 is not polarized and it can be called that it contains general companies. It also explains why this sample is much larger.

## 4 Empirical Evidence

Final scores for the models introduced in chapter 2 were calculated for each company from the data sample. These individual values have been summed up using basic descriptive statistics. Table 2 provides an illustration on an example of Altman model. This kind of the table could be provided for each model. Altman formula has been chosen because of its worldwide popularity and extension of use. Table 2 displays the main descriptive statistics for healthy companies 2012 divided into three subsamples - Manufacture of machinery and equipment (CZ-NACE 28), Construction (CZ-NACE F) and Manufacture of fabricated metal products, except machinery and equipment (CZ-NACE 25).

**Table 2.** Altman Z-Score and its descriptive statistics for healthy companies 2012

| Healthy companies 2012 | Manufacture of machinery and equipment | Construction | Manufacture of fabricated metal products |
|---|---|---|---|
| Mean | 3.08 | 3.97 | 4.55 |
| Median | 2.63 | 3.36 | 3.76 |
| Minimum | 1.16 | 0.26 | 1.12 |
| Maximum | 6.76 | 22.89 | 64.67 |
| 1st quartile | 2.31 | 2.39 | 2.75 |
| 3rd quartile | 3.67 | 4.83 | 5.22 |
| St. deviation | 1.27 | 2.63 | 4.41 |
| Trim mean | 3.02 | 3.76 | 4.17 |

According to the results included in Table 2 there are visibly differences among Altman Zcore values in different industry sectors. Manufacture of machinery and equipment reached the lowest values and Manufacture of fabricated metal products reached the highest value. The difference is 1.47 point which can be almost identified with the range of the grey zone of the model [3]. The difference is significant. It proves that the companies belonging to the Manufacture of machinery and equipment had lower values of the prediction model and therefore they looked as less healthy in comparison with the other two industry sectors. The gained result confirms that industry specifics can play an important role in the case of the financial situation prediction.

Table 2 included only one model and one of three parts of the data sample therefore it is necessary to visualize the obtained data in more complex way. A helpful tool would be figure which displays results for the analyzed models predicting financial distress. The models are ordered in the figures as 1 – Altman, 2 – IN01, 3- IN05, 4 – Doucha, 5 – Bonita, 6 – Prusak 1, 7 – Prusak 2, 8 – PAN-E, 9 – PAN-F, 10 – PAN_G, 11 – D2, 12 – D3, 13 - Hajdu & Virág, 14 – Šorins & Voronova, 15 – Merkevicus, 16 - Rmodel. Kralicek Quick Test has been excluded because its metrics works differently in comparison to other models. Higher value means healthier company for majority of the models. Kralicek Quick Test is an exception because the desired result is the minimization of the final model score. Taffler model has been excluded from figures as well because its final values exceed at least two times the maximal values of the other models and therefore it would distort the display and differences among models would not be visible.

There are 3 figures, each one for one data subpart of the analysed data sample (insolvent companies 2012, healthy companies 2012 and general companies 2017). Blue line shows the industry sector Manufacture of machinery and equipment (CZ-NACE 28), red line represent the industry sector Construction (CZ-NACE F) and finally green one is for Manufacture of fabricated metal products, except machinery and equipment (CZ-NACE 25). The lines represent final scores of the prediction models because they display trim mean for a particular group of the companies. Trim mean separates outliers.

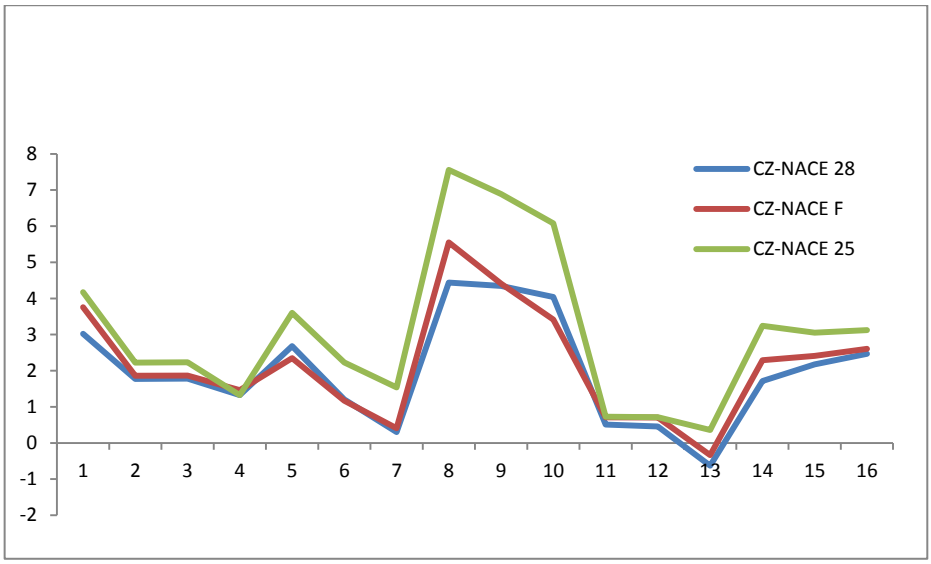



**Fig. 1.** Prediction models and their final values for healthy companies 2012

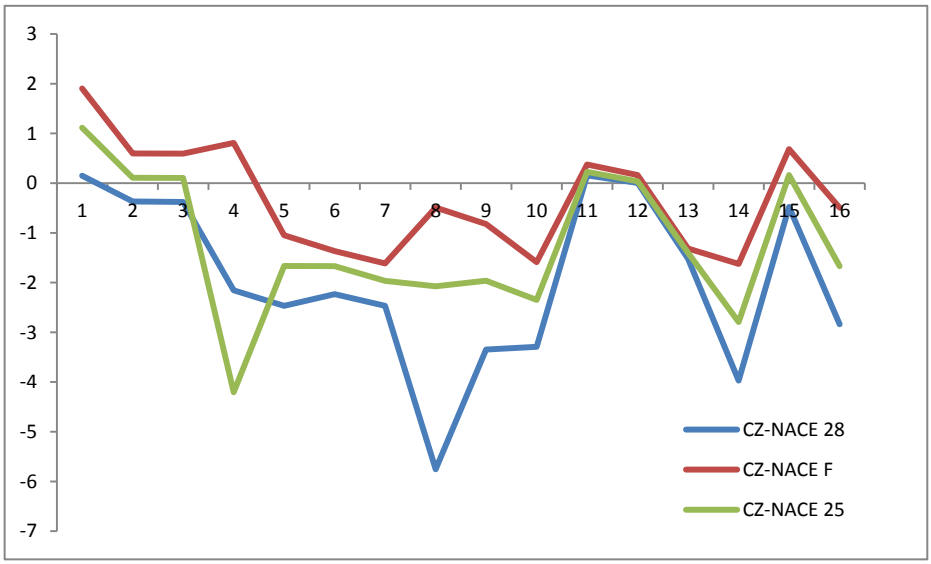



**Fig. 2.** Prediction models and their final values for insolvent companies 2012

Figure 1 confirms conclusions from Table 2 that there are significant differences among the industry sectors in the case of the healthy companies. CZ-NACE 25 reaches the highest scores, then CZ-NACE F and on the opposite site there is CZ-NACE 28. It must be noted that not all the models provide the same results because the final

position of CZ-NACE F and CZ-NACE 28 is not the same. It means that the final ranking cannot be generalized.

Figure 2 consists of results for the insolvent companies from the year 2012. Significant differences can be observed again in this case. The worst is the machinery industry (CZ-NACE 28) and the best is the construction industry (CZ-NACE F). This conclusion is not valid for all verified models because curves intersect or they are so close that there are almost any differences among industries. Divergent conclusions are caused by the indicators included in the individual models predicting financial distress. The models prefer different financial ratios and therefore there are tradeoffs between the values of the separated indicators which cause differences in the final scores.

It should be emphasized that the year 2012 cannot be classifies as stable at all. The Czech economy still dealt with consequences of a last global economic crisis in that time. The third part of the data sample contains the financial data from the stable time period of the year 2017. The Czech Republic was fully compensated with the economic crisis of previous years and it grew economically. The results of this period are introduced by Figure 3.

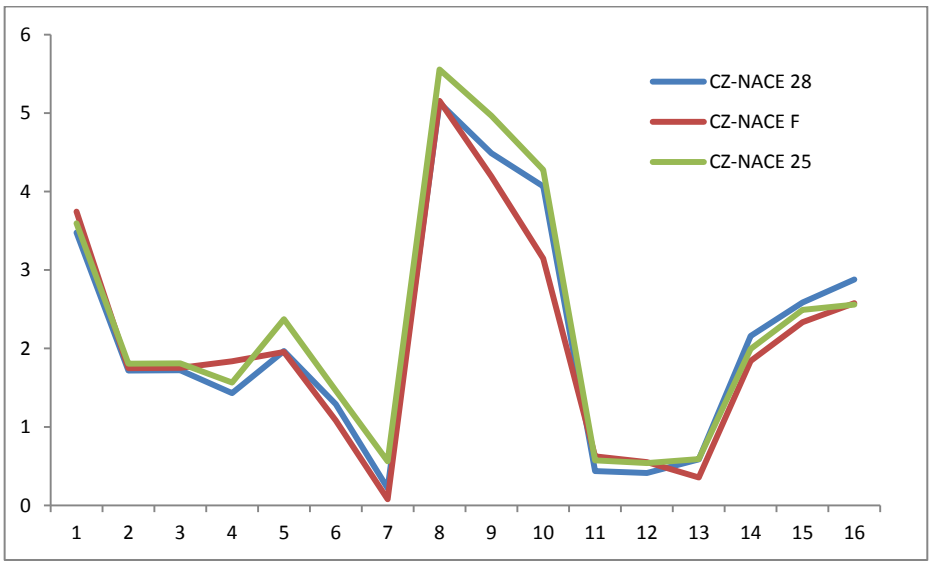



**Fig. 3.** Prediction models and their final values for general sample 2017

Although figure 3 uses the same approach as the previous figures displaying the year 2012 there cannot be observed the significant differences among the chosen industry branches. The results for the individual prediction models are comparable. The companies reach similar value of the final score without a respect to the industry branch. It cannot be pointed out which industry branch has the highest and which one the lowest values. Figure 3 does not the confirm conclusions obtained from figure 1 and figure 2. Although the companies belong to the different industry branches and they

have hardly the same sale, property and financial structure the final score of the models predicting financial distress does not differ significantly. In this case the research idea cannot be confirmed. Reasons causing these results will be discussed in a conclusion part.

Kralicek Quick Test and Taffler model have been left aside. The reasons were mentioned. Both models predicting financial distress confirm the gained conclusions because there are not significant differences in the time of stability and on other hand in 2012 it is possible to observe differences among industry branches. In the case of Kralicek Test the differences are slighter because the models is not working on the continuous basis but on the discrete one (values of ratios belong to the categories which are valued) and it limits the differences by nature.

## Conclusion

This paper was focused on the models predicting financial distress which were constructed in the past but they are still popular and highly used for predictions in the corporate practices. The analysis worked with almost 20 models. Untraditionally the paper did not follow the testing of models' accuracy and explanatory power but it focused on the specifics among industry sectors. On one hand the analyzed prediction tools are widespread because of their general usage on the other hand there are specifics of the industries which influences sale, property, capital structure etc. These specifics could influence the models' results significantly and therefore the final scores of the prediction models were tested. The results are based on three industry branches, specifically Manufacture of fabricated metal products, except machinery and equipment (CZ-NACE 25), Manufacture of machinery and equipment (CZ-NACE 28) and Construction (CZ-NACE F).

Final scores of the analyzed prediction models show that there were significant differences in the two subparts of the data sample – the insolvent companies and the healthy companies in 2012. The worst results achieved CZ-NACE 28 for both subsamples. CZ-NACE F was the best among the insolvent companies and CZ-NACE 25 among the healthy companies. The third subpart – the general companies 2017 – did not support these findings because there were comparable results without the respect to the industry sector. It is not possible to prove if the industry specifics influence the models significantly. Although the industry sector has an impact on the corporate financial statements, balance sheet and profit and loss account, it does not influence the final values of the models predicting financial distress significantly. The models are based on several financial ratios. As the results show there are some tradeoffs between the indicators which enter the final value of the prediction model. Differences among the individual indicators could be analyzed further in the following research work. There have to be discrepancies in the case of leverage, profit margin or net working capital among the industry branches.

It must be noted that the basic research idea was confirmed for the subsamples from 2012 which is connected with the consequences of the last global economic crisis. It can be concluded that in the time of overall economic stability there are not

significant differences in the case of the industry branches but the instability is bringing differences. It can be caused by several reasons. First the industries react differently and it influences the financial statements and then the financial ratios included in the prediction models. Second the industries were differently influenced by the economic crisis. Some were more affected than the others and this is visible on the gained results. For the confirmation it would be necessary to analyze the individual financial ratios and use also other kinds of information including industry statistics as a production slump, price movements or changed payment conditions in that industry.

## Acknowledgement

The paper is one of the outputs of the research project “Financial characteristics of enterprise in bankruptcy” registered at Grant Agency of Academic Alliance under the registration No. GAAA 10/2018.

## References

1. Agarwal, V., Taffler, R.J.: Twenty-five years of the Taffler z-score model: does it really have predictive ability? Accounting and Business Research 37(4), 285-300 (2007).
2. Altman, E.I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23(4), 589-609 (1968).
3. Altman, E.I.: Corporate Financial Distress and Bankruptcy. 2nd edn. John Wiley & Sons, New York (1993).
4. Altman, E,I.: Predicting corporate distress in a turbulent economic and regulatory environment. Rassegna Economica 68(2), 483-524 (2004).
5. Beaver, W.: Financial Ratios as Predictors of Failure. Journal of Accounting Research 4(3), 71-111 (1966).
6. Čámská, D.: Accuracy of Models Predicting Corporate Bankruptcy in a Selected Industry Branch. Ekonomický časopis 64(4), 353-366 (2016).
7. Čámská, D.: Models Predicting Financial Distress and their Accuracy in the Case of Construction Industry in the Czech Republic. In: Pastuszková, E., Crhová, Z., Vychytilová, J., Vytrhlíková, B., Knápková, A. (eds.) 7th International Scientific Conference Finance and Performance of Firms in Science, Education and Practice, Tomas Bata University in Zlin , pp. 178-190. Tomas Bata University in Zlin, Zlín (2015).
8. Davidova, G. Quantity method of bankruptcy risk evaluation. Journal of Risk Management 3, 13 – 20 (1999).
9. Doucha, R: Finanční analýza podniku: praktické aplikace. 1st edn. Vox Consult, Prague (1996).
10. Hajdu, O., Virág, M.: Hungarian Model for Predicting Financial Bankruptcy. Society and Economy in Central and Eastern Europe 23(1-2), 28-46 (2001).
11. Jansone, I., Nespors, V., Voronova, I.: Finanšu un ekonomisko risku ietekme uz Latvijas partikas mazumtirdzniecibas nozares attistibu. Scientific Journal of Riga Technical University Economics and Business Economy: Theory and Practice 20, 59-64 (2010).
12. Jordan, B.D., Westerfield, R., Ross, S.A.: Corporate finance essentials. McGraw-Hill, New York (2011).

13. Karas, M., Režňáková, M.: Bankruptcy Prediction Model of Industrial Enterprises in the Czech Republic. INTERNATIONAL JOURNAL of MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES 7(5), 519-531 (2013).
14. Kisielinska, J., WASZKOWSKI, A.: Polskie modele do prognozowania bankructwa przedsiębiorstw i ich weryfikacja. EKONOMIKA i ORGANIZACJA GOSPODARKI ŻYWNOŚCIOWEJ 82, 17-31 (2010).
15. Klečka, J., Scholleová, H.: Bankruptcy models enunciation for Czech glass making firms. Economics and management 15, 954-959 (2010).
16. Korol, T., Korodi, A.: Predicting Bankruptcy with the Use of Macroeconomic Variables. Economic computation and economic cybernetics studies and research 44(1), 201-219 (2010).
17. Kralicek, P., ERTRAGS- UND VERMÖGENSANALYSE (QUICKTEST), http://www.kralicek.at/pdf/qr_druck.pdf, last accassed 2019/05/15.
18. Liou, D-K, Smith, M: Macroeconomic Variables and Financial Distress. Journal of Accounting, Business & Management 14, 17-31 (2007).
19. Machek, O.: Long-term Predictive Ability of Bankruptcy Models in the Czech Republic: Evidence from 2007-2012. Central European Business Review 3(2), 14-17 (2014).
20. Merkevicius, E., Garšva, G., Girdzijauskas, S.: A Hybrid SOM-Altman Model for Bankruptcy Prediction. In: Alexandrov et al. (ed.). ICCS 2006, 364 – 371 (2006).
21. Neumaierová, I., Neumaier, I.: Index IN05. In: Evropské finanční systémy, Masarykova univerzita, pp. 143-148. Masarykova univerzita, Brno (2005).
22. Neumaierová, I., Neumaier, I.: Výkonnost a tržní hodnota firmy. 1st edn. Grada, Prague (2002).
23. Pitrová, K.: Possibilities of the Altman Zeta Model Application to Czech Firms. E&M Ekonomika a Management14(3), 66-76 (2011).
24. Wöber, A., Siebenlist, O.: Sanierungsberatung für Mittel- und Kleinbetriebe, Erfolgreiches Consolting in der Unternehmenkrise. Erich Schmidt Verlag GmbH, Berlin (2009).

# Analyzing Extreme Financial Risks: A Score-driven Approach

Fernanda Fuentes[1], Rodrigo Herrera[2], and Adam Clements[3]

[1]*Facultad de Ingeniería, Universidad de Talca, Curicó. Chile*
[2]*Facultad de Economía y Negocios, Universidad de Talca, Talca, Chile*
[3]*School of Economics and Finance, Queensland University of Technology, Brisbane, Australia*

**Abstract**

This paper develops a new class of dynamic extreme value models, driven by the score of the conditional distribution with respect to both the duration between extreme events and the magnitude of these events. This data-driven framework is a feasible method for capturing their time-varying arrival intensity, and magnitude. It is also shown how exogenous variables such as realized measures of volatility can easily be incorporated. An empirical analysis based on a set of major equity indices shows that both the arrival intensity and the size of extreme events vary greatly during times of market turmoil. The proposed framework performs well relative to a competing approach in terms of forecasting extreme tail risk measures.

*JEL classification:* C11; C58; G17; Q47; Q02

*Keywords:* Score-driven models, Time-varying parameters, Extreme value theory, Value at Risk, Expected Shortfall, Realized Volatility.

# 1 Introduction

Since the Global Financial Crisis and more recent sovereign debt crisis in several European countries, numerous new econometric approaches have been developed to quantify tail risk dynamics (see for instance Gandhi & Lustig 2015, Agarwal et al. 2017, Andersen et al. 2017, Lucas et al. 2017). The first typically consists of estimating a conditional volatility model, followed by modeling the dynamics of these events through the residuals standardized obtained in this first step (e.g. McNeil & Frey 2000, Garcia & Tsafack 2011, Bee et al. 2016, Koliai 2016, Sahamkhadam et al. 2018). The main shortcoming of this approach is that volatility measures based on only past returns cannot accurately capture tail risk during financial market turmoil (see Longin 2000, Bali 2000, Hong et al. 2007). Only models that contain a leverage term, irrespective of whether they are models based on daily returns (e.g., SV or GARCH) or realized measures of volatility (e.g., HAR or HEAVY) have been successful in explain the dynamic of extreme events, see Liu & Tawn (2013) and Trapin (2017). Hence, a measure of risk different to the volatility, harnessing only the likelihood of extreme downward market movements, is required.

In order to overcome this limitation, several models based on marked point processes have been proposed, which take into account the timing and the magnitude of large losses occurring over a high threshold acts as a proxy for volatility at extreme levels (e.g. Chavez-Demoulin & McGill 2012, Chavez-Demoulin et al. 2014, Gresnigt et al. 2016, Herrera & Clements 2018). The key point of this approach is that, in order to understand the dynamics of tail risk in financial markets, extreme events can be split into two stochastic components. One associated with their irregularly spaced occurrence through time, and another associated with the severity of these events.

The main contribution in this paper is to provide a novel observation-driven framework for modeling the dynamic behavior of both components. In particular, a new class of dynamic marked point processes for extreme events is proposed. Under this approach the stochastic process of the duration between extreme events (inter-exceedance times) form a path-dependent point process, whereas the associated magnitudes are the marks which depend on the arrival times and the history of the stochastic process of the extreme events.

The models proposed here are denoted as Score-driven Peaks Over Threshold (SPOT) models and embed the most important elements of a time-varying extreme value model. The main

difference is that the dynamics of the parameters are functions of the observations through the score function of the predictive density at the time of each extreme event; see Creal et al. (2013) and Harvey (2013). Score driven models have become popular in recent years in many economic and financial applications (see Calvori et al. 2017, Gorgi et al. 2018, Massacci 2016, Bernardi & Catania 2019), mainly because these are the only models with time-varying parameters whose updating equations will always reduce the local Kullback-Leibler divergence between the true conditional density and the model-implied conditional density (Blasques et al. 2015).

Additionally, an extension denoted here as the realized SPOT (rSPOT) model which includes realized volatility measures (e.g., simple realized variance, realized semi variance, jumps, negative jumps) is proposed. The use of realized volatility measures has been rapidly gaining popularity as an alternative to standard parametric volatility models, and they have been also utilized in the context of financial extreme risk in recent years (Bee et al. 2016, Bee et al. 2018, Fodor et al. 2013, Yeh & Chen 2014).

The main question of interest is to determine to what degree can the dynamic behaviour of extreme events (and hence tail risk) be explained by the two stochastic components, their occurrence times and/or their magnitudes. In addressing this issue, the question of whether there any gains from using realized measurements to produce more accurate estimates of extreme market risk will also be addressed.

The benefits of the proposed framework are illustrated in an empirical analysis of tail risk for a set of major world stock indices from 2000 to 2018. The main results can be summarized as follows. The estimation results confirm that using the information on the occurrence times and the magnitude of extreme events by means of the score-driven approach is helpful for describing tail risk dynamics. In particular, by decomposing the tail risk dynamics, financial crisis are almost simultaneously reflected in the dynamic of both stochastic processes; the arrival times of extreme events and their associated magnitudes. In fact, both the Subprime crisis and the European debt crisis can be clearly distinguished. In addition, the dynamic parameter that best describes the behavior of the inter-exceedance times is the shape parameter of the conditional hazard function, exhibiting heavy-tailed behavior. The dynamics of the magnitudes is best described by a time varying scale parameter of the conditional probability distribution function. Further, it is observed that incorporating realized measures of volatility into the SPOT framework leads to gains in terms

of goodness of fit.

Finally, the performance of the SPOT framework is examined under two measures of risk, Expected Shortfall (ES) and Value at Risk (VaR), considering two different backtesting periods. The performance of the SPOT models are compared with the one-factor GAS model of Patton et al. (2019). The results of the out-of-sample forecasting exercise show that the SPOT model outperforms the one-factor GAS model for the one-year VaR backtesting period, while it exhibits similar results during the longer two-year backtesting period, for most of the markets considered. In terms of a joint test of ES and VaR, the SPOT models outperforms the one-factor GAS model.

# References


Agarwal, V., Ruenzi, S. & Weigert, F. (2017), 'Tail risk in hedge funds: A unique view from portfolio holdings', *Journal of Financial Economics* **125**(3), 610–636.

Andersen, T. G. & Bollerslev, T. (1998), 'Deutsche Mark–Dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies', *the Journal of Finance* **53**(1), 219–265.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**(2), 579–625.

Andersen, T. G., Bollerslev, T. & Meddahi, N. (2005), 'Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities', *Econometrica* **73**(1), 279–296.

Andersen, T. G., Bollerslev, T. & Meddahi, N. (2011), 'Realized volatility forecasting and market microstructure noise', *Journal of Econometrics* **160**(1), 220–234.

Andersen, T. G., Fusari, N. & Todorov, V. (2017), 'Short-term market risks implied by weekly options', *The Journal of Finance* **72**(3), 1335–1386.

Bali, T. G. (2000), 'Testing the empirical performance of stochastic volatility models of the short-term interest rate', *Journal of Financial and Quantitative Analysis* **35**(2), 191–215.

Bali, T. G. & Weinbaum, D. (2007), 'A conditional extreme value volatility estimator based on high-frequency returns', *Journal of Economic Dynamics and Control* **31**(2), 361–397.

Balkema, A. A. & De Haan, L. (1974), 'Residual life time at great age', *The Annals of Probability* **5 (2)**, 792 – 804.

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(2), 253–280.

Bee, M., Dupuis, D. J. & Trapin, L. (2016), 'Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective', *Journal of Empirical Finance* **36**, 86–99.

Bee, M., Dupuis, D. J. & Trapin, L. (2018), 'Realized extreme quantile: A joint model for conditional quantiles and measures of volatility with EVT refinements', *Journal of Applied Econometrics* **33**(3), 398–415.

Bernardi, M. & Catania, L. (2019), 'Switching generalized autoregressive score copula models with application to systemic risk', *Journal of Applied Econometrics* **34**(May), 43–65.

Blasques, F., Koopman, S. J. & Lucas, A. (2015), 'Information-theoretic optimality of observation-driven time series models for continuous responses', *Biometrika* **102**(2), 325–343.

Brownlees, C. T. & Gallo, G. M. (2009), 'Comparison of volatility measures: A risk management perspective', *Journal of Financial Econometrics* **8**(1), 29–56.

Calvori, F., Creal, D., Koopman, S. J. & Lucas, A. (2017), 'Testing for parameter instability across different modeling frameworks', *Journal of Financial Econometrics* **15**(2), 223–246.

Chavez-Demoulin, V., Embrechts, P. & Sardy, S. (2014), 'Extreme-quantile tracking for financial time series', *Journal of Econometrics* **181**(1), 44–52.

Chavez-Demoulin, V. & McGill, J. (2012), 'High-frequency financial data modeling using Hawkes processes', *Journal of Banking & Finance* **36**(12), 3415–3426.

Christoffersen, P. (1998), 'Evaluating interval forecasts', *International economic review* pp. 841–862.

Corsi, F. (2009), 'A simple approximate long-memory model of realized volatility', *Journal of Financial Econometrics* **7**(2), 174–196.

Corsi, F., Fusari, N. & La Vecchia, D. (2013), 'Realizing smiles: Options pricing with realized volatility', *Journal of Financial Economics* **107**(2), 284–304.

Creal, D., Koopman, S. J. & Lucas, A. (2013), 'Generalized autoregressive score models with applications', *Journal of Applied Econometrics* **28**(5), 777–795.

Davison, A. & Smith, R. (1990), 'Models for exceedances over high thresholds', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 393–442.

Engle, R. F. & Gallo, G. M. (2006), 'A multiple indicators model for volatility using intra-daily data', *Journal of Econometrics* **131**(1-2), 3–27.

Fissler, T., Ziegel, J. F. et al. (2016), 'Higher order elicitability and Osband's principle', *The Annals of Statistics* **44**(4), 1680–1707.

Fodor, A., Krieger, K., Mauck, N. & Stevenson, G. (2013), 'Predicting extreme returns and portfolio management implications', *Journal of Financial Research* **36**(4), 471–492.

Fuentes, F., Herrera, R. & Clements, A. (2018), 'Modeling extreme risks in commodities and commodity currencies', *Pacific-Basin Finance Journal* **51**, 108–120.

Gandhi, P. & Lustig, H. (2015), 'Size anomalies in US bank stock returns', *The Journal of Finance* **70**(2), 733–768.

Garcia, R. & Tsafack, G. (2011), 'Dependence structure and extreme comovements in international equity and bond markets', *Journal of Banking & Finance* **35**(8), 1954–1970.

Gneiting, T. (2011), 'Making and evaluating point forecasts', *Journal of the American Statistical Association* **106**(494), 746–762.

Gorgi, P., Hansen, P., Janus, P. & Koopman, S. (2018), 'Realized Wishart-GARCH: A score-driven multi-asset volatility model', *Journal of Financial Econometrics* **17**(1), 1–32.

Gresnigt, F., Kole, E. & Franses, P. H. (2016), 'Specification testing in Hawkes models', *Journal of Financial Econometrics* **15**(1), 139–171.

Hansen, P. R., Huang, Z. & Shek, H. H. (2012), 'Realized GARCH: A joint model for returns and realized measures of volatility', *Journal of Applied Econometrics* **27**(6), 877–906.

Harvey, A. C. (2013), *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, Vol. 52, Cambridge University Press.

Herrera, R. & Clements, A. (2018), 'Point process models for extreme returns: Harnessing implied volatility', *Journal of Banking & Finance* **88**, 161–175.

Hong, Y., Li, H. & Zhao, F. (2007), 'Can the random walk model be beaten in out-of-sample density forecasts? Evidence from intraday foreign exchange rates', *Journal of Econometrics* **141**(2), 736–776.

Koliai, L. (2016), 'Extreme risk modeling: An EVT–pair-copulas approach for financial stress tests', *Journal of Banking & Finance* **70**, 1–22.

Kupiec, P. H. (1995), 'Techniques for verifying the accuracy of risk measurement models', *The Journal of Derivates* **3**(2).

Liu, Y. & Tawn, J. A. (2013), 'Volatility model selection for extremes of financial time series', *Journal of Statistical Planning and Inference* **143**(3), 520–530.

Longin, F. M. (2000), 'From value at risk to stress testing: The extreme value approach', *Journal of Banking & Finance* **24**(7), 1097–1130.

Lucas, A., Schwaab, B. & Zhang, X. (2017), 'Modeling financial sector joint tail risk in the euro area', *Journal of Applied Econometrics* **32**(1), 171–191.

Massacci, D. (2016), 'Tail risk dynamics in stock returns: Links to the macroeconomy and global markets connectedness', *Management Science* **63**(9), 3072–3089.

McNeil, A. & Frey, R. (2000), 'Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach', *Journal of Empirical Finance* **7**, 271–300.

Noureldin, D., Shephard, N. & Sheppard, K. (2012), 'Multivariate high-frequency-based volatility (heavy) models', *Journal of Applied Econometrics* **27**(6), 907–933.

Patton, A. J., Ziegel, J. F. & Chen, R. (2019), 'Dynamic semiparametric models for expected shortfall (and value-at-risk)', *Journal of Econometrics* . (In Press).

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *The Annals of Statistics* pp. 119 – 131.

Sahamkhadam, M., Stephan, A. & Östermark, R. (2018), 'Portfolio optimization based on GARCH-EVT-Copula forecasting models', *International Journal of Forecasting* **34**(3), 497–506.

Smith, R. (1989), 'Extreme value analysis of environmetal time series: An application to trend detection in ground-level ozone', *Statistics Sience* **4**, 367–393.

Trapin, L. (2017), 'Can volatility models explain extreme events?', *Journal of Financial Econometrics* **16**(2), 297–315.

Yeh, J.-H. & Chen, L.-C. (2014), 'Stabilizing the market with short sale constraint? New evidence from price jump activities', *Finance Research Letters* **11**(3), 238–246.

# Freedman's Paradox: an Info-Metrics Perspective

Pedro Macedo

CIDMA – Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, 3810-193, Aveiro, Portugal
pmacedo@ua.pt
https://cidma.ua.pt/

**Abstract.** In linear regression models where there are no relationships between the dependent variable and each of the potential explanatory variables – a usual scenario in real-world problems – some of them can be identified as relevant by standard statistical procedures. This incorrect identification is usually known as Freedman's paradox. To avoid this disturbing effect in regression analysis, an info-metrics approach based on normalized entropy is discussed and illustrated in this work. The results suggest that normalized entropy is a powerful alternative to traditional statistical methodologies currently used by practitioners.

**Keywords:** Big Data, Regression, Variable Selection.

## 1 Introduction

Consider a linear regression model defined as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ denotes a $(N \times 1)$ vector of noisy observations, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown parameters to be estimated, $\boldsymbol{X}$ is a known $(N \times K)$ matrix of explanatory variables, and $\boldsymbol{e}$ is the $(N \times 1)$ vector of random disturbances, typically assumed to have a conditional expected value of zero and representing spherical disturbances.

Freedman [1, p. 152] states that "[...] in a world with a large number of unrelated variables and no clear a priori specifications, uncritical use of standard methods will lead to models that appear to have a lot of explanatory power. That is the main – and negative – message of the present note." Through simulation studies and asymptotic theory it is demonstrated some technical features of this misleading interpretation, including the behaviour of the t-test, the F-test and the coefficient of determination, $R^2$. Freedman [1] shows that in a regression model where does not exist relationships between independent/explanatory variables and the dependent variable, if there are many explanatory variables in the model, then the $R^2$ will be high and some explanatory variables can be easily considered relevant variables through common significance tests.

Info-Metrics is a research area at the intersection of statistics, computer science and decision theory, where the maximum entropy principle established by Jaynes [6, 7] plays a central role. Maximum entropy provides a simple tool to make the best prediction (i.e., the one that is the most strongly indicated) from the available information and it can be seen as an extension of the Bernoulli's principle of insufficient reason.

To highlight the importance of the maximum entropy principle, Soofi [11, p. 1244] provides an interesting statement, which possibly remains valid nowadays: "Jaynes introduced the maximum entropy principle of inference with which many statisticians have some familiarity but for which the statistics community as a whole has not yet developed sufficient appreciation."

To illustrate an info-metrics approach to avoid the above mentioned disturbing effect in regression analysis, the generalized maximum entropy (GME) and generalized cross entropy (GCE) estimators are briefly presented in Section 2, along with the definition of normalized entropy. Although there are other methodologies for variable selection, this paper is intended only to illustrate the use of info-metrics procedures. A comparison with other recent methodologies (e.g., lasso and its generalizations) is left for future work. The remainder of the paper is laid out as follows: in Section 3 the simulation studies are implemented; some conclusions and topics for future research are given in Section 4.

## 2 Info-Metrics: Estimators and Normalized Entropy

Golan, Judge and Miller [3, pp. 86-93] proposed a reformulation of the linear regression model in (1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{2}$$

where

$$\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{p} = \begin{bmatrix} \boldsymbol{z}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{z}_K' \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_K \end{bmatrix}, \tag{3}$$

with $\boldsymbol{Z}$ a $(K \times KM)$ matrix of support spaces and $\boldsymbol{p}$ a $(KM \times 1)$ vector of unknown probabilities to be estimated, and

$$\boldsymbol{e} = \boldsymbol{V}\boldsymbol{w} = \begin{bmatrix} \boldsymbol{v}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{v}_N' \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix}, \tag{4}$$

with $\boldsymbol{V}$ a $(N \times NJ)$ matrix of support spaces and $\boldsymbol{w}$ a $(NJ \times 1)$ vector of unknown probabilities to be estimated. In this reformulation, each $\beta_k$, $k = 1, 2, \ldots, K$, and each $e_n$, $n = 1, 2, \ldots, N$, are viewed as expected values of discrete random variables $\boldsymbol{z}_k$ and $\boldsymbol{v}_n$, respectively, with $M \geq 2$ and $J \geq 2$ possible outcomes, within

the lower and upper bounds of the corresponding support spaces. Additional details can be found in Golan [2], Chapter 13.

For the linear regression model expressed in (1), the generalized maximum entropy (GME) estimator is given by

$$\underset{\boldsymbol{p},\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\boldsymbol{p}' \ln \boldsymbol{p} - \boldsymbol{w}' \ln \boldsymbol{w} \right\}, \tag{5}$$

subject to the model constraints,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{6}$$

and the additivity constraints for $\boldsymbol{p}$ and $\boldsymbol{w}$, respectively,

$$\begin{aligned} \mathbf{1}_K &= (\boldsymbol{I}_K \otimes \mathbf{1}'_M)\boldsymbol{p}, \\ \mathbf{1}_N &= (\boldsymbol{I}_N \otimes \mathbf{1}'_J)\boldsymbol{w}, \end{aligned} \tag{7}$$

where $\otimes$ represents the Kronecker product. On the other hand, with the same restrictions, the generalized cross entropy (GCE) estimator is given by

$$\underset{\boldsymbol{p},\boldsymbol{w}}{\operatorname{argmin}} \left\{ \boldsymbol{p}' \ln \left( \frac{\boldsymbol{p}}{\boldsymbol{q}_1} \right) + \boldsymbol{w}' \ln \left( \frac{\boldsymbol{w}}{\boldsymbol{q}_2} \right) \right\}, \tag{8}$$

where $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are vectors with prior information concerning the parameters and the errors of the model, respectively.

The estimators generate the optimal probability vectors $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{w}}$ that can be used to form point estimates of the unknown parameters and the unknown errors, through the reparameterizations (3) and (4) defined previously. It is important to note that the GME estimator is a particular case of the GCE estimator, when the prior information is expressed as a uniform distribution (vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$). In view of the fact that ill-posed real-world problems seem to be the rule rather than the exception, these estimators have acquired special importance in the set of statistical techniques, by allowing statistical formulations free of restrictive and unnecessary assumptions.

Additionally, to measure the information content of the signal component in a particular model, Golan, Judge and Miller [3, p. 93, p. 165] defined normalized entropy as

$$S(\widehat{\boldsymbol{p}}) = \frac{-\widehat{\boldsymbol{p}}' \ln \widehat{\boldsymbol{p}}}{K \ln M} \tag{9}$$

in the GME estimator, and

$$S(\widehat{\boldsymbol{p}}) = \frac{-\widehat{\boldsymbol{p}}' \ln \widehat{\boldsymbol{p}}}{-\boldsymbol{q}'_1 \ln \boldsymbol{q}_1} \tag{10}$$

in the GCE estimator context. This measure lies between zero (no uncertainty) and one (perfect uncertainty). Concerning variable selection, it is interesting to note that if all the $\boldsymbol{z}_k$ in $\boldsymbol{Z}$ are defined uniformly and symmetrically around zero, then $S(\widehat{\boldsymbol{p}}_k) \approx 1$ implies $\beta_k \approx 0$, because $\widehat{\boldsymbol{p}}_k$ is uniformly distributed. Thus,

a variable corresponding to $S(\widehat{\boldsymbol{p}}_k) \approx 1$ has no information content and should be excluded from the model.

Some advantages of this procedure are presented by Golan, Judge and Miller [3, p. 176]: is simple to perform, even for a large number of variables (just one analysis of the sample is needed, which represent important computational advantages; it does not require the evaluation of $2^K$ models); allows the use of non-sample information (through the supports in GME or the vectors with prior information in GCE); is free of asymptotic requirements; involves a shrinkage rule that reduces mean squared error; allows to account for model misspecifications and model uncertainty; and it can be implemented for well- and ill-posed models.

Additional details on maximum entropy estimation, normalized entropy, simulation studies, properties and asymptotic theory can be found in Golan, Judge and Miller [3], Mittelhammer, Cardell and Marsh [9], and Golan [2].

## 3 Simulation Studies

The simulation studies conducted in this work follow the same structure of the ones performed by Freedman [1]. Different matrices are created with 100 rows and 51 columns. All the entries are independent observations generated from the standard normal distribution. To establish a multiple regression model, the first 50 columns are considered as the explanatory variables and the last column as the dependent variable. Given this construction, all the regression coefficients should be considered statistically not significant by the standard t-test. However, this won't be the case (as expected).

Freedman [1] performed two successive model estimations: in the first one are identified the number of coefficients that are statistically significant at the 25% (representing an exploratory analysis) and the 5% (representing a confirmatory analysis) levels; in the second one, only the variables whose coefficients are significant at the 25% level enter to the regression model and the number of coefficients that are statistically significant at the 25% and the 5% levels are identified, again. All the results are misleading, in particular on the second stage, where are identified between one and nine statistically significant coefficients in the models (depending on the simulation), at the 5% significance level.

To illustrate variable selection using normalized entropy, the GME and GCE estimators are performed with four different supports: $[-100, 100]$, $[-10, 10]$, $[-5, 5]$ and $[-2, 2]$ for all the parameters. The supports are defined as closed and bounded intervals in which each parameter is restricted to lie. Since there is empirical evidence that different supports provide different results in terms of variable selection, four supports (with five points) are tested in this work, reflecting different levels of prior information about the parameters.

For each error support is used the three-sigma rule, considering the standard deviation of the noisy observations (usual procedure in GME literature by using a sample scale parameter), with three points. The number of points in the supports is usually between three and seven, since there is likely no significant improvement in the estimation with more points in the supports.

Regarding the GCE estimator, and following Golan, Judge and Miller [3, p. 166], which state that "If we believe that potential extraneous variables with zero coefficients exist in the linear statistical model specifications, it would seem reasonable to shrink those close to zero more than others.", a vector with prior information is defined as $\boldsymbol{q}_1 = [0.1, 0.2, 0.4, 0.2, 0.1]$ for all the parameters, which will accomplish the idea of additional shrinkage. As mentioned by Golan, Judge and Miller [3], the priors take over as the solution when they are consistent with the data. This feature of the GCE estimator is revealed in the results.

### 3.1 Results

Due to space limitations, only two models are used here: 18 and 14 are the number of regression coefficients statistically significant at 25% level, in the first stage, which means that, in the second stage, the two models have only 18 and 14 variables. Table 1 presents the number of regression coefficients statistically significant, at different significance levels, in the first stage, for both models.

**Table 1.** Number of coefficients statistically significant (first stage).

| | Significance levels | | | | | |
|---|---|---|---|---|---|---|
| | 1% | 2% | 3% | 4% | 5% | 10% |
| Model 1 (50 variables) | 0 | 2 | 3 | 4 | 4 | 6 |
| Model 2 (50 variables) | 0 | 2 | 2 | 2 | 4 | 7 |

Considering the three usual significance levels, in the first stage with both models with 50 variables each, six and seven coefficients are considered statistically significant at 10% level, respectively, in Model 1 and Model 2. Additionally, four coefficients are considered statistically significant at 5% and none of them is considered statistically significant at 1% level, in both models.

Table 2 presents the number of regression coefficients statistically significant, at different significance levels, in the second stage. Is this second estimation, seven coefficients are considered statistically significant at 5% and two coefficients are considered statistically significant at 1% level, in both models. Additionally, 11 coefficients are considered statistically significant at the 10% level, in Model 1.

**Table 2.** Number of coefficients statistically significant (second stage).

| | Significance levels | | | | | |
|---|---|---|---|---|---|---|
| | 1% | 2% | 3% | 4% | 5% | 10% |
| Model 1 (18 variables) | 2 | 3 | 5 | 5 | 7 | 11 |
| Model 2 (14 variables) | 2 | 3 | 6 | 7 | 7 | 7 |

Since the models are pure noise, the results are disturbing because they suggest relationships that do not exist between explanatory variables and the dependent variable.

Table 3 presents the normalized entropy (truncated to four decimals) of the models, $S(\widehat{\boldsymbol{p}})$, considering different supports for GME and GCE estimators. It is interesting to see that all values are near one, indicating no information content of the signal in the models (in both stages).

**Table 3.** Normalized entropy for the models.

| | | | Supports | | | |
|---|---|---|---|---|---|---|
| | | | $[-100, 100]$ | $[-10, 10]$ | $[-5, 5]$ | $[-2, 2]$ |
| GME | Model 1 | 50 variables | 0.9999 | 0.9998 | 0.9994 | 0.9972 |
| | | 18 variables | 0.9999 | 0.9997 | 0.9991 | 0.9954 |
| | Model 2 | 50 variables | 0.9999 | 0.9998 | 0.9994 | 0.9972 |
| | | 14 variables | 0.9999 | 0.9997 | 0.9991 | 0.9951 |
| GCE | Model 1 | 50 variables | 0.9999 | 0.9998 | 0.9995 | 0.9982 |
| | | 18 variables | 0.9999 | 0.9998 | 0.9994 | 0.9969 |
| | Model 2 | 50 variables | 0.9999 | 0.9998 | 0.9995 | 0.9982 |
| | | 14 variables | 0.9999 | 0.9998 | 0.9993 | 0.9967 |

However, to improve the research, a more detailed analysis is developed and all the $S(\widehat{\boldsymbol{p}}_k)$ for each model are also obtained. The results are reported through boxplots, from Fig. 1 to Fig. 8.

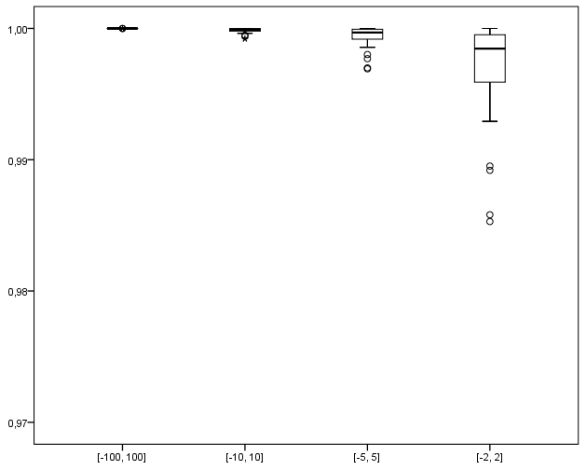



**Fig. 1.** $S(\widehat{\boldsymbol{p}}_k)$ with GME in Model 1 (50 variables).

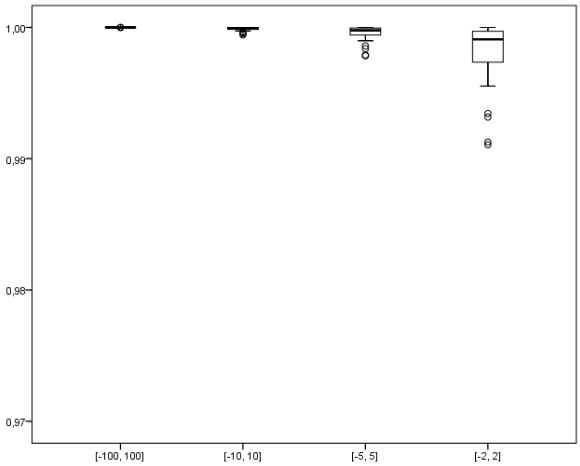



**Fig. 2.** $S(\widehat{\boldsymbol{p}}_k)$ with GCE in Model 1 (50 variables).

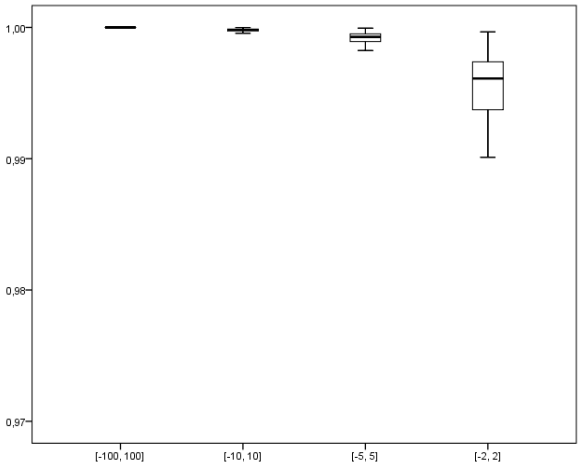



**Fig. 3.** $S(\widehat{\boldsymbol{p}}_k)$ with GME in Model 1 (18 variables).

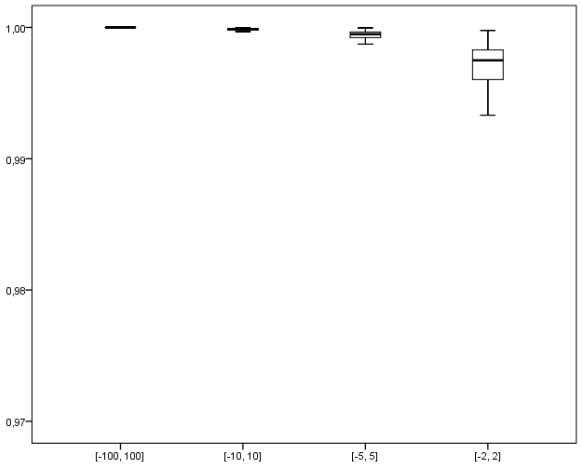



**Fig. 4.** $S(\widehat{\boldsymbol{p}}_k)$ with GCE in Model 1 (18 variables).

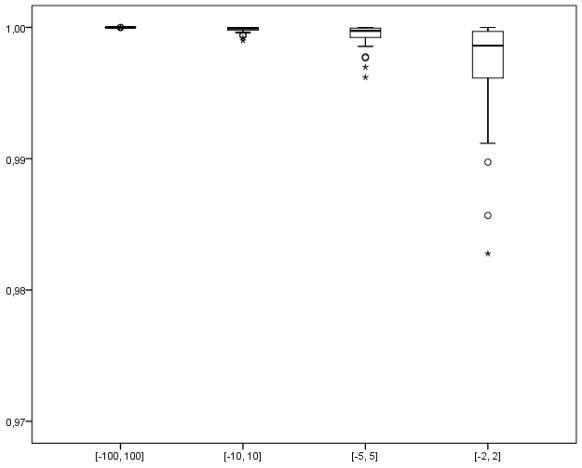



**Fig. 5.** $S(\widehat{\boldsymbol{p}}_k)$ with GME in Model 2 (50 variables).

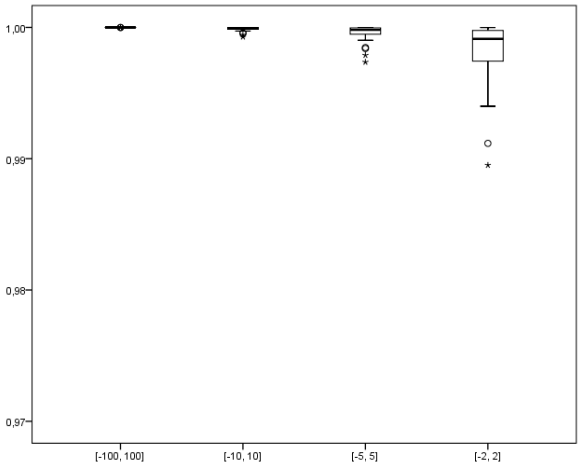



**Fig. 6.** $S(\widehat{\boldsymbol{p}}_k)$ with GCE in Model 2 (50 variables).

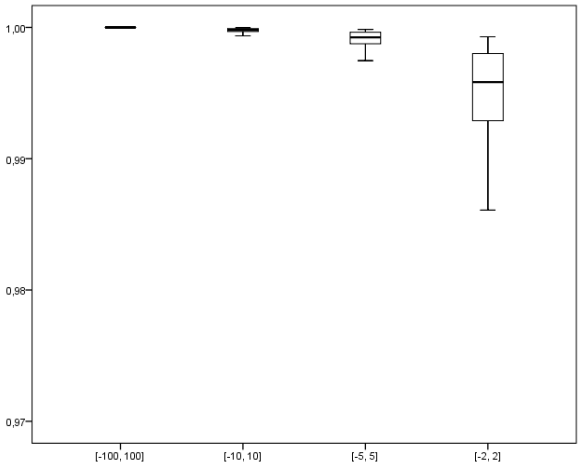



**Fig. 7.** $S(\widehat{\boldsymbol{p}}_k)$ with GME in Model 2 (14 variables).

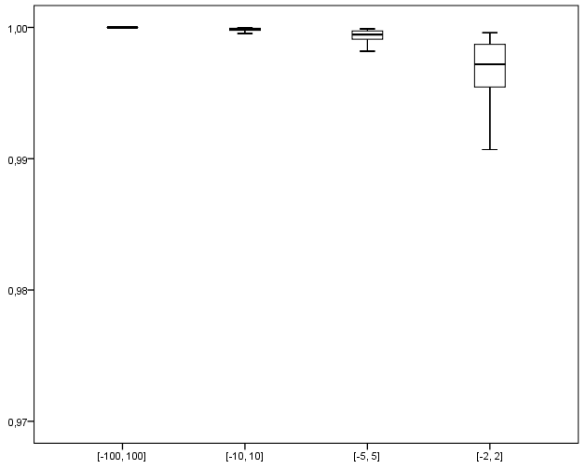



**Fig. 8.** $S(\widehat{\boldsymbol{p}}_k)$ with GCE in Model 2 (14 variables).

Although in some scenarios, especially in the ones with supports of lower amplitude, slightly lower normalized entropy values are obtained, the values are always very high. Note that the y-axis is defined just between 0.97 and 1.00, and normalized entropy values range between zero and one.

The good performance of the normalized entropy procedure in terms of variable selection, which is observed in Table 3, as well as in Fig. 1 to Fig. 8, is also achieved in other simulated models not reported here.

## 4 Concluding Remarks

The results in Table 3 suggest no information content of the signal in the models, regardless the supports considered or the maximum entropy estimator used. This is an interesting result because, under the conditions of the previous models, in the first stage, when $N \to \infty$ and $K \to \infty$, so that $K/N \to \rho$, where $0 < \rho < 1$, then the $R^2$, a standard procedure usually evaluated by practitioners, tends to $\rho$, and the ratio of the number of relevant variables by $N$ tends to $\alpha\rho$, where $\alpha$ represents the significance level considered; see Freedman [1].

Taking into account that a variable corresponding to $S(\widehat{\boldsymbol{p}}_k) \approx 1$ has no information content (it is considered irrelevant) and should be removed from the model, the analysis of Fig. 1 to Fig. 8 suggests the exclusion of all the variables, in both stages. Although in some scenarios, namely in the one with the support defined as $[-2, 2]$, lower normalized entropy values are obtained, all of them are greater than 0.98. Indeed, if the criterion of inclusion considered by Golan, Judge and Miller [3, p. 165] is applied, $S(\widehat{\boldsymbol{p}}_k) \leq 0.99$, a few variables are considered relevant when the support $[-2, 2]$ is used, although the number of incorrect inclusions is lower when the GCE estimator is applied, as expected

given the prior information considered. As mentioned previously, the priors take over as the solution when they are consistent with the data.

Naturally, without a formal rule to define a cutoff value, the identification of "relevant" variables (with "relevant" information content) can be considered difficult in the cases with normalized entropy values "near" one. Nevertheless, regarding this possible concern, is it really necessary a cutoff value? Is it not sufficient the evaluation of the information embodied in the normalized entropy? Possible answers to these questions should always take into account, although in a different perspective, the theoretical discussions provided by Wasserstein and Lazar [12], and Hurlbert, Levine and Utts [5], where some recommendations to statisticians are provided, namely to eliminate the choice of specific significance levels or to abolish the use of the terms "statistically significant", when p-values are interpreted in hypothesis testing. (It is important to note that s-values can be much more useful than p-values; e.g., Greenland [4]. Information measures based on Shannon's work [10] are very attractive in statistical inference.)

The results in this work suggest that the evaluation of normalized entropy is a promising approach to avoid the disturbing effect in regression analysis described by Freedman's paradox. Future research on the definition of the supports and in the amount of pressure around zero, established by the prior information vector for the GCE estimator, should be accomplished, along with the comparison with recent methodologies (e.g., lasso and its generalizations). As a final remark, a MATLAB code to compute normalized entropy using the GME estimator can be easily obtained from the code available in Macedo [8]; see Appendix.

**Acknowledgments.** The author was supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

## References

1. Freedman, D.A.: A note on screening regression equations. Amer. Statist. 37(2), 152–155 (1983)
2. Golan, A.: Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information. Oxford University Press, New York (2018)
3. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics - Robust Estimation with Limited Data. John Wiley & Sons, Chichester (1996)
4. Greenland, S.: Valid $P$-Values Behave Exactly as They Should: Some Misleading Criticisms of $P$-Values and Their Resolution With $S$-Values. Amer. Statist. 73(S1), 106–114 (2019)
5. Hurlbert, S.H., Levine, R.A., Utts, J.: Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. Amer. Statist. 73(1), 352–357 (2019)
6. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. 106(4), 620–630 (1957)
7. Jaynes, E.T.: Information theory and statistical mechanics. II. Phys. Rev. 108(2), 171–190 (1957)
8. Macedo, P.: Ridge regression and generalized maximum entropy: an improved version of the Ridge-GME parameter estimator. Comm. Statist. Simulation Comput. 46(5), 3527–3539 (2017)

9. Mittelhammer, R., Cardell, N.S., Marsh, T.L.: The data-constrained generalized maximum entropy estimator of the GLM: asymptotic theory and inference. Entropy 15, 1756–1775 (2013)
10. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. 27(3), 379–423 (1948)
11. Soofi, E.S.: Capturing the intangible concept of information. J. Amer. Statist. Assoc. 89(428), 1243–1254 (1994)
12. Wasserstein, R.L., Lazar, N.A.: The ASA's Statement on p-Values: Context, Process, and Purpose. Amer. Statist. 70(2), 129–133 (2016)

## Appendix: MATLAB code

To adapte the code available in Macedo [8], the first line of the original code can be replaced, for example, by

```
function [b3,nep,nepk]=nentropy(Y,X)
```

Suppose a model, for example, with $K = 6$ and consider all the supports in $\boldsymbol{Z}$, for example, as $[-10, -5, 0, 5, 10]$. Lines 38-71 are replaced by

```
intg=[-10,10;-10,10;-10,10;-10,10;-10,10;-10,10];
```

Lines 116-132 are replaced by

```
p=a(1:dp)';
b3=Z*p;
nep=(-p'*log(p))/(k*log(m));
nepk=zeros(k,1);
for i=1:k
    pos=(i-1)*m+1;
    nepk(i,1)=-p(pos:pos+m-1)'*log(p(pos:pos+m-1))/log(m);
end
```

All lines with comments and features related to the original code should be eliminated. Other changes can be made (e.g., the number of points in the supports).

**UNIVERSITY OF TARTU**

Faculty of Social Sciences
School of Economics and Business Administration

# Powers of Text

Diana Gabrielyan[1], Jaan Masso[2], Lenno Uuskula[3]
[1]University of Tartu, Tartu, Estonia

**Submitted: 22 May 2019**

[1] Diana Gabrielyan, PhD Student, School of Economics and Business Administration, University of Tartu, Tartu, Estonia. E-mail: diana.gabrielyan@ut.ee

[2] Jaan Masso, Senior Research Fellow, School of Economics and Business Administration, University of Tartu, Tartu, Estonia. E-mail: jaan.masso@ut.ee

[3] Lenno Uuskula, Senior economist, Research Division, Bank of Estonia, Tallinn, Estonia. E-mail. Lenno.Uuskyla@eestipank.ee

**Abstract**

In this paper we take advantage of technological advances and use high frequency multidimensional textual news data available in the internet and propose a new index of inflation expectations. We utilize the power of text mining and its ability to convert large collections of text from unstructured to structured form for in-depth quantitative and qualitative analysis of Guardian news data. Main contribution of his paper is to explore online news as novel data source to capture the inflation expectations in real time. We do so by building an index of inflation expectations and capture the intensity and uncertainty of expectations as well as the quantitative value. The preliminary results show that the new inflation index is correlated with the actual inflation dynamics. Moreover, the inflation news precedes actual inflation by a few months. To validate our results, we build a linear regression using our newly built indices and market-based inflation expectations and confirm that our methodology results in a model with good forecasting power.

**JEL Classification**. C53, E31, E47

**Keywords**: forecasting, inflation, machine learning, text mining, guardian

## 1 Introduction

Household surveys of inflation indicate show often that the perception of the current inflation and expectations about the future expectations are different from actual inflation values and differ strongly from the surveys of professional forecasters and implied inflation rates of financial markets, see for example Coibion et al (2018). Economic relationships such as the Phillips curve that are estimated with the actual statistics are not performing well, especially in last decades.

Potential reason for the difference is that households and firms get only very partial information while doing everyday shopping aggregating the information is very costly. Consumers build expectations through personal experiences and prior memory of inflation, which however can be inaccurate, irrational and diverse. Instead they rely on public media when thinking about overall price changes. Imperfect information affects the expectations formation negatively.

This paper measures inflation from public media using machine learning approach. There are a lot of news that cover prices and price developments. Frequency of the news and the tone of the text can drive inflation perceptions. Alternatively, households' answers about inflation is a very noisy measure of the actual inflation perception. Measuring directly the news about inflation could help to understand economic relationships such as the Phillips curve better.

To address the above-mentioned issues and drawbacks we propose a new index of inflation expectations based on online news data. We utilize the power of text mining and its ability to convert large collections of text from unstructured to structured form for in-depth quantitative and qualitative analysis of Guardian news data. We capture consumers' inflation expectations by building high dimensional indicator based on articles over the last 3 years. The need for such a real-time indicator is based on the lack of analogous indicators and the delay in the publishing of official statistics[4]. And because using textual data from the news is a recent phenomenon there are few research papers that use such rich data source for opinion mining and sentiment analysis. To our best knowledge, there is currently no literature available that extracts consumer inflation expectations from news data.

The preliminary results show that the new inflation index is correlated with the actual inflation dynamics. Moreover, the inflation news precedes actual inflation by a few months. Further research will use the newly built index in estimation of various economic relationships.

[4] Market-based expectations are available daily but include risk premia. Survey-based expectations are published monthly.

We use text pre-processing techniques to filter out the data and get the frequency count of each word. Then we use Latent Dirichlet Allocation method for topic extraction, as suggested by Blei, Ng and Jordan (2003) and then proceed with dictionary-based approaches for document sentiment analysis, as suggested by Thorsrud (2018). The index of inflation expectations is then built and is compared to official inflation statistics.

Machine learning methods are considered to be very promising avenue for future academic and applied research (Bank of England 2015). Extensive overview of literature and how

innovations in data mining can lead to high-quality inference about model parameters is provided in Belloni, Chernozhukov and Hansen (2014).

One modern strand of machine learning is text mining. Although still new in economics, yet there is already a number of literature available that use text mining in different forms to extensively extract information from text documents. Novel sources of data, such as social media (e.g. Twitter, Google) allow analysis and different kind of understanding of consumer behavior.

The idea is that, every search in Google is someone expressing interest in or demand for something (Brynjolfsson 2012). These searches also capture further expectations and can therefore be helpful in forecasting. For example, Tuhkuri (2016) uses actual search volumes on Google to construct a state-level panel data, along with the constructed Google Index to verify that Google searches, indeed, anticipate the US unemployment rate. Similarly, D'Amuri and Marcucci (2017) asses the performance of Google job-search intensity as a leading indicator for predicting the US monthly unemployment and find that the models based on Google indicators outperform the others and are able to forecast quarterly unemployment rate more accurately than Survey of Professional Forecasters. Yu et al (2018) use Google trends to build an online oil consumption forecasting model and confirm that it improves upon other models significantly for both directional and level predictions.

Another great source of textual data for economics are the daily business newspapers. Assuming, that that the given news website provides accurate and relevant description of the events happening in the economy, it is possible to understand the importance of given topic for economy and its future based on the intensity and the extent of how much it is discussed. This, in its turn, allows to understand consumers expectations. It is a known fact that, inflation expectations are of great importance for Central banks, as they reflect the monetary authority's commitment credibility to price stability (Berge 2007). In addition, they contribute to understanding the houseful consumptions, as well as saving and investment decisions (Cavallo et al. 2017).

One factor affecting households' and firms forming expectations is central bank's policy actions and communication reports. Hence, market players update their expectations as new information in released. This highlights the importance of high frequency real time data for accurate formations of expectations. The problem, however, is that the available survey- based inflation expectations have low frequency and the high-frequency market-based forecasts involve risk and may be uncertain. One way to overcome this issue of low- frequency-but-more-reliable and high-frequency-but-unstable data problem is to take advantage of technological advances and use high frequency multidimensional textual data available in the internet, which makes it possible to build an indicator that is not included in the official statistics. The idea is that this textual data may contain information about expectations that is not contained in the quantitative data and may capture "true" expectations.

There is a strand of literature on extracting public's perception from various textual data from news outlets and central bank communications. For example, Hendry and Madeley (2010) use Latent Semantic Analysis to extract information from Bank of Canada communication statements and analyze which time of information affects returns and volatility in short-term and long-term interest rate. El-Shagi and Jung (2015) find that the minutes of Bank of England's Monetary Policy Committee have contributed to markets expectation formations

on the future of monetary policy. The interest rate skew from these minutes has helped explain future changes of the bank's rate. Lucca and Trebbi (2009) measure the content of central bank communication about future interest rate decisions based on information from news sources and internet.

When estimating Treasury yield responses to the shocks, the find that communication is a more important factor for Treasury rates than the contemporaneous policy rate decisions. For example, Sapiro, Sudhof and Wilson (2018) use computational text analysis of economic and financial news articles to asses time series measures of economic sentiment that drive consumption. Nyman and his co-authors (2015) study the role of narratives and emotions in driving developments in the financial system by analyzing large amounts of unstructured financial markets-based text data. Onsumran et al (2015) develop a gold price volatility prediction model using text mining approach to analyze how news articles influence gold price volatility. Thorsrud (2018) constructs a perfectly accurate new business cycle index based on quarterly GDP growth as well as information from daily business newspaper that classifies the phases of the business cycle and provides meaningful insights on which type of news drive or reflect economic fluctuations.

Main contribution of this paper is to explore online news as novel data source to capture the inflation expectations in real time. We do so by building an index

of inflation expectations and capture the intensity and uncertainty of expectations as well as the quantitative value.

We do not invent a new methodology to extract data from the news website, nor do we propose new model for inflation forecasting. Instead we use existing methods and combine them with the novel source of information to prove that online news can provide a real-time and accurate indication of consumer's expectations on inflation.

The paper is organized as follows. Section 2 describes the data sources used to build the index. Section 3 discusses thoroughly the methodology and models that we employ to capture the inflation expectations. Section 4 provides results and compares them with the official statistic and Section 5 provides alternative applications and concludes.

## 2 Data

Paper uses two types of data: long sample of online newspaper corpus and data from official statistics. The official inflation expectations data was taken from the Office for National Statistics and was used as a proxy to examine how close is our constructed index of inflation to existing sources of expectations.

The choice of the news outlet is due relevance to our research in terms of content and readership, as well as the availability of open source data. As such, we chose Guardian news data for our analysis. Guardian is a British newspaper, that also publishes all its news online. In May 2013, it was the most popular UK newspaper website with 8.2 million unique visitors per month and in April 2011, it was the fifth most popular newspaper in the world[5]. Any news is public and readable by anyone by default. Overall, we collect 9857 documents and 6.27 million terms (out of which around 500K unique terms) from January 2016 to January 2019, which is enough to make our analysis.

We only fetch articles from the business section, since this is the most relevant section linking to economy in general. In addition, articles were also filtered based on subjectively chosen keywords, which in our opinion are relevant to inflation expectations topic. Namely, they are "price", "price increase", "expensive", "cheaper" etc. We, purposefully, did not include term "inflation" in the keyword list when fetching the data to make sure we grasp all other business-related topic from the Guardian news database, from which we can build the expectations.

[5] "Guardian.co.uk most read newspaper site in UK in March". www.journalism.co.uk. May 2013. "MailOnline overtakes Huffington Post to become world's no 2". MediaWeek. Haymarket, April 2011.

As mentioned, choice of the news source was also based on the readers' geolocation. UK is particularly interesting country to study the topic of inflation expectations, particularly because of 2016 Brexit vote, which lead to an immediate shift in expectations about UK's economic future, with market participants downgrading their expectations of UK economy.

## 3 Methodology

The whole process of building the inflation expectation index can be divided into data collection part, as described in section 2 and analysis of the data. This section describes the analysis of the data by means of text mining.

### 3.1 Pre-processing

Large amount of the textual data that we extract from the Guardian's database, makes statistical computation challenging. Therefore, some clean-up is needed. Like any text mining research, we start with pre-processing, which a set of activities performed on the corpus. This way, the unstructured is modified into structured form, the dimensionality of the data is reduced, noise is eliminated, and we get more understandable results. In this paper we use text mining's "bag of words"[6] approach, which means all words are analysed as a single token and their structure, grammar or order does not matter. We mostly follow suggestions for pre-processing by Bholat and co-authors (2015), at the same time adding more steps and more developed methods. Each of these techniques has its own pros and cons. For example, along with reducing dimensionality, these techniques might obscure meaning for some words or might count words that are written similarly but have different meanings as same word.

### 3.2 Topic Extraction

The pre-processing results in a data frame which consists of the words used in the text and their frequencies. These words consist a document-term matrix, where each row of the matrix is a unique term and each column is a unique document. To proceed to building the index, topics need to be extracted from the DTM. Topic modelling is the statistical approach for discovering topics from the collection of text document. In other words, it is the process of looking into a large collection of documents and identifying clusters of words based on similarity, patters and multitude. Since any document can be assigned to several topics at a time, the probability distribution across topics for each document is therefore needed. Blei D, Ng A, Jordan A (2003) were the first to suggest the use of Latent Dirichlet Allocation (LDA) for this purpose. LDA is a statistical

[6] In text mining, vector representations of text are called bag-of-words representations

model that identifies each document as a mixture of topics (related to multiple topics) and attributes each word to one of the document's topics, therefore, clustering words into topics. With LDA method it is possible to derive their probability distribution by assigning probabilities to each word and document. Assigning words and documents to multiple topics also has advantage of

In LDA each document is given a probability distribution and for each word in each document, a topic assignment is made. The joint distribution of topic mixture **θ,** a set of **N** words **w** is given by

$$p(\theta,z,w|\alpha,\beta) = p(\theta|\alpha)^{*}\prod_{n=1}^{N} p(z_n|\theta) * p(w_n|z_n,\beta)\ , \quad (1)$$

where parameters **α** and **β** are k-vectors with components greater than zero, with k being the dimensionality of Dirichlet distribution, that is the directionality of topic variable **z**. In addition, the topic distribution of each document is distributed as

$$\theta \sim \text{Dirichlet}(\alpha),$$

term distribution is modeled by

$$z_n \sim \text{Dirichlet}(\beta),$$

and

$$N \sim \text{Possion}(\xi).$$

LDA model's goal, is therefore to estimate θ and φ in order to estimate which words are important for which topic and which topics are important for a given document. For **α** and **β,** the higher they are, the more likely each document will contain a mixture of most topics instead of a single topic and the more likely each topic will contain a mixture of most of the words and not just single words. More technical and through specifications on the LDA model and topic modeling in general in Blei (2013) and Griffiths and Steyvers (2004). LDA results in a vector indicating the distribution of topics in each document and most popular/relevant words within them. This step in the analysis is now concluded.

### 3.3 Index and Sentiment Analysis

For each document within a day, 5 most popular words are identified and their frequency for the day is counted. This allows counting also the frequency of each topic for a given day.

At this step, our results of topic decompositions and distribution is used to build the new high frequency index that will capture the intensity of inflation expectations. The index is built for every day, that is, we build daily time series using Guardian's business articles for each day. To do so, we first sum together all articles for a given day into one document, grouping them into one plain text

for each date. Next, based on the first 10 most frequent words in each topic the article's daily frequency is calculated. In other words, the frequency is calculated for the given day as the raw count of frequencies with which the most common words in each topic appear in that day. For example, to understand the intensity of how many times the word "vote" has been used on June 22, 2016 (the day of the Brexit vote), we will summarize all the Guardian articles for that day as one big text document, then calculate the number of times the word "vote" appear in the text. Here the Brexit can be our topic and the "vote" is the term.

The news volume **I(t)** of given topic **z** is given by

$$I_z(\mathrm{t}) = \sum_{d \in I(t)} \sum_w N(d, w, z), \quad (2)$$

where $N(d, w, z)$ is the frequency with which the word **w** tagged with topic **z** appears in document **d**.

These time series $I_z(\mathrm{t})$ are measures volume measure, that is, they measure the intensity of given topic for given time period. This is the first version of index we build, which will serve as robustness index. This index is normalized and compared to official series in the results section. We also build a second type of index $\overline{I_z}(\mathrm{t})$, which also includes the sentiment of the topic, meaning an indication of whether the news is positive or negative. This is important, particularly, since our aim is to build the index of inflation expectations. There are number of ways to identify the emotion of each topic, such as rule-based or dictionary approaches, supervised or unsupervised machine learning. Thorsrud (2018) uses dictionary approach for this purpose and uses Harvard IV-4 Psychological Dictionary with the negative and positive words already listed in it. Others use Support vector machine (SVM), which is part of supervised learning methods. For this an initial dataset identified as positive text or negative text needs to be supplied and the method can letter classify the sample into one of these or into neutral tone, which will be decided in case the resulting probability of one document belonging to specific tone is below defined threshold after applying SVM method.

For the analysis in this paper, the Harvard IV-4 psychological dictionary is most convenient to use, since Guardian news are in English. We therefore follow Thorsrud(2018) and use the dictionary based approach to classify the articles emotion's. We construct three indexes, one for each emotion (positive, negative, neutral). Inflation expectations defined by sentiment index are then given by below formula:

$$S_z(\mathrm{t}) = \overset{pos}{\tilde{S}_z}(\mathrm{t}) - \overset{neg}{\tilde{S}_z}(\mathrm{t}), \quad (3)$$

Where, $\overset{pos}{\widetilde{S_z}}$(t) and $\overset{neg}{\widetilde{S_z}}$(t) are the count of positive words from the dictionary in all articles at day **t** for topic **z**. The inflation expectations is therefore the difference between positive and negative sentiments. To build the $\overset{pos}{\widetilde{S_z}}$(t), $\overset{neg}{\widetilde{S_z}}$(t)'s, for each day **t** and each topic **z**, we find the article that is best described by that topic. This is done by looking at the document-topic probabilities resulting from LDA, since besides estimating each topic as mixture of words, LDA also models each document as mixture of topics. Examining this per-document-per-probabilities, we can find the highest probable article for the given topic. The result of this is a topic – article mapping, which is then used to identify the tone of the given day for given topic. For each of these articles, for each day for each topic we count the number of positive and negative words provided by the Harvard IV-4 dictionary. The difference of these two statistics is then calculated, resulting in $S_z$(t) and the final adjusted inflation expectations are calculated using the below formula

$$\overline{I_z}(t) = I_z(t) * S_z(t) \quad (4)$$

As a last step we use 30-day and 60-day moving average filter to remove the high-frequency noise from the index series. As highlighted by Thorstrud (2018) this is a common practice in factor model studies and authors like Stock and Watson (2016) have applied this method in their studies.

## 4 Inflation Index

The results show that the LDA decomposition gives meaningful classification of topics of the Guardian news website. We manually pruned the topics based on the topic distributions and kept only those that we found to be most relevant for building the inflation expectations index. Particularly, we focused on topics related to inflation and oil & energy (hereinafter, Energy). Thus, out of 50 original topics, we filtered 40 topics and kept 10 topics to work with (4 for inflation topic, 7 for Energy), Table 1 below lists the chosen topics with their 10 most frequent words. As mentioned, LDA does not assign names to the topics, however seeing the most frequent words within the topics it is possible to understand the theme.

| Topic Number | Top Frequent Words | Primary Identified Topic |
|---|---|---|
| 17 | "expens, price, cost, increas, rise, higher, import, good, fall" | Inflation |
| 18 | "model, car, industri, product, sale, vehicl, diesel, electr" | Energy |

| | | |
|---|---|---|
| 24 | "fall, inflat, price, rise, rate, wage, cost, consum, live" | Inflation |
| 26 | "climat, energi, power, gas, industri, electr, renew, coal, wind" | Energy |
| 32 | "saudi, oil, price, opec, barrel, crude, product, cut, iran" | Energy |
| 36 | "economi, bank, rate, rise, expect, interest, inflat, rais, polici" | Inflation |
| 44 | "euro, ecb, draghi, bank, rate, central, polici, eurozon, inflat" | Inflation |
| 45 | "oil, china, market, chines, global, trade, price, world, economi" | Energy |
| 46 | "competit, energi, custom, price, supplier, cap, big, market, bill" | Energy |
| 48 | "energi, govern, project, power, nuclear, point, plan, build, edf" | Energy |

Table 1: Top inflation and energy topics from LDA results

We also ran LDA model by varying the number of topics from 10-50 and the results did not change significantly. Below graph visualizes the main topics with the Guardian articles from January 2016 to January 2019.

The high-frequency noise from the time series is removed using the 60-day moving average filter. For robustness, we also used 30-day and 80-day moving average filter alternative methods but found no significant change in the results. The official inflation data is the consumer price index CPIH, which is the measure of rate of UK consumer price inflation that includes owner occupiers' housing costs. Both CPIH and our newly built indices are standardized.

The time evolution of the news volume for most relevant topics related directly to inflation and energy/oil along with the official inflation series are given in below in figures 1 and 2. These indices series are translated into tone adjusted time series using methodology presented in section 3.3. This way we get frequency of mentioning each topic each day and the tone (positive or negative). At the same time, the indices also capture the intensity and uncertainty of inflation expectations. For robustness we analyzed the indices without tone adjustment. Their time evolution is presented in Appendix D.

Figure 1: CPIH Inflation at Monthly Rate vs newly built index from Inflation Topic 24

Figure 2: CPIH inflation at monthly rate vs newly built index from energy topic 26

The figures show interesting trend. Not only is the newly built index time evolution in line with the official inflation statistics, but a more thorough look to the figures also shows that the newly build index is on average ahead of the official inflation in terms of direction. Figure 3 demonstrates this finding. Our newly built index of inflation, if shifted two months forward (and scaled) classifies the phases of the UK inflation at monthly rate with a great accuracy. This means, that the Guardian news give the indication of inflation change direction up to 2 months before it actually happens. And since the consumers form their expectations based on the numerous factors, including news they read, then they also assume the direction of inflation change months before the change

happens. Similar results hold for energy topics and for other inflation-related topics. Moreover, for some of the indices 3 months ahead adjustment was needed to be in line with official inflation's trend.

Figure 3: CPIH inflation at monthly rate vs newly built index from energy topic 48

Further analysis of rolling correlation between the indices and official inflation shows that the series are highly correlated.

## 5 Applications

Above results are clear indication that the inflation expectations indices built based on Guardian newspaper articles are able to not only capture the UK inflation at monthly rate, but also predict the direction of inflation change up to 3 months in advance. To further validate our findings and confirm the robustness of our results, we built a linear regression using our newly built indices and market-based inflation expectations. For the latter, we used our daily indices and the UK 1-year inflation swaps for from January 2016 to January 2019[7].

For the analysis we took all 50 topics but performed principal component analysis to reduce the number of regressors. We obtained 8 principal

[7] ICE Benchmark Administration Limited (IBA), ICE Swap Rates, 11:00 A.M. (London Time), Based on British Pound, 1 Year Tenor [ICERATES1100GBP1Y], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/ICERATES1100GBP1Y , May 16, 2019.

components, hence out of initial 50 topics, we were left with 8 regressors, that describe our dependent variable, inflation expectations at 1-year swap rates, well. To verify the quality of our linear regression, the data was divided into training (in-sample) and test samples (out-of-sample) using 80/20 principle. Results of the forecasts based on the regression is given in Figure 7.

Figure 7: Results from forecasting

Above results along with those from series of out-of-sample forecasting exercises confirms that the methodology presented in this paper results in a model with good forecasting power and can therefore be competitive to existing models. This finding also provides insight on our questions raised in the introduction that news influences the consumer's inflation expectations, at the same time capturing them quite well.

The dataset used is novel, available in real-time and future work may validate the results with longer time series. In addition, our suggestion for future research is use other sources of textual data, exploiting the variety available in the internet (e.g. Twitter, Google, New York Times etc.) to forecast, track or analyze the dynamics of other macroeconomic variables. Indicators built based on methodologies presented in this paper may allow more effective macroeconomic policymaking for Central banks.

## References


**Alessi L, Ghysels E, Onorante L, Peach R and Potter S (2014),** "Central Bank Macroeconomic Forecasting During the Global Financial Crisis: The European Central Bank and Federal Reserve Bank of New York Experiences", *Journal of Business & Economic Statistics,* Vol. 32, No. 4, pages 483–500. Available at https://doi.org/10.1080/07350015.2014.959124

**Askitas N and Zimmermann K F (2009)**, "Google Econometrics and Unemployment Forecasting", *Applied Economics Quarterly*, Vol. 55, No. 2, pages 107–120. Available at https://doi.org/10.3790/aeq.55.2.107

**Bank of England (2015)**, "One Bank Research Agenda Discussion Paper". Available at http://www.bankofengland.co.uk/research/Documents/onebank/discussion.pdf

**Belloni A, Chernozhukov V and Hansen Ch (2014)**, "High-Dimensional Methods and Inference on Structural and Treatment Effects", *Journal of Economic Perspectives*, Vol. 28, No. 2, pages 1–23. Available at http://www.mit.edu/~vchern/papers/JEP.pdf

**Berge T J (2017)**. "Understanding survey-based inflation expectations", *Finance and Economics Discussion Series,* Vol. 46. Available at https://doi.org/10.17016/FEDS.2017.046

**Brynjolfsson E (2012)**, "Big Data: A revolution in decision-making improves productivity", *MIT Sloan Experts.*

**Bholat D, Hansen S, Santos P and Schonhardt-Bailey Ch (2015),** "Text mining for central banks", *Handbooks, Centre for Central Banking Studies, Bank of England*, No. 33. Available at https://www.bankofengland.co.uk/-/media/boe/files/ccbs/resources/text-mining-for-central-banks.pdf?la=en&hash=C49C23BF808B13FAD5361D0D2516DA12646120A6

**Blei D M, Ng A Y and Jordan M I (2003),** "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, Vol. 3, pages 993–1022. Available at http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

**Cavallo A, Guillermo C and Perez-Truglia R (2017)**, "Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments", *American Economic Journal: Macroeconomics,* Vol. 9, No. 3, pages 1-35. Available at
https://www.nber.org/papers/w20576.pdf

**Cavallo A (2013),** "Online and official price indexes: Measuring Argentina"s inflation," *Journal of Monetary Economics*, Vol. 60, No. 2, pages 152-165. Available at
http://www.thebillionpricesproject.com/wp-content/papers/Cavallo_Real-Consumption_AEAPP.pdf

**Chan J C C, Song J (2017),** "Measuring inflation expectations uncertainty using high-frequency data," *CAMA Working Papers*, Vol. 61. Available at https://cama.crawford.anu.edu.au/sites/default/files/publication/cama_crawford_anu_edu_au/2017-10/61_2017_chan_song.pdf

**D'Amuri F and Marcucci J (2017)***, "*The predictive power of Google searches in forecasting US unemployment", *International Journal of Forecasting,* Vol. 33, No. 4, pages 801-816, Available at
*https://www-sciencedirect-com.ezproxy.utlib.ut.ee/science/article/pii/S0169207017300389*

**Eckley P (2015),** "Measuring economic uncertainty using news-media textual data", *MPRA Paper*, No. 64874. Available at http://mpra.ub.uni-muenchen.de/64874/

**George E I and McCulloch R E (1993),** "Variable Selection Via Gibbs Sampling", *Journal of the American Statistical Associations*, Vol. 88, No. 423, pages 881-889. Available at https://www.jstor.org/stable/2290777

**Griffiths T L and Steyvers M (2004)**, "Finding scientific topics", *Proceedings of the National academy of Sciences of the United States of America 101* (Suppl. 1), pages 5228– 5235.
Available at
https://www.pnas.org/content/101/suppl_1/5228

**Hendry S and Madeley A (2010)**, "Text Mining and the Information Content of Bank of Canada Communications", *Bank of Canada Working Paper*, Vol. 31, Available at
https://www.banqueducanada.ca/wp-content/uploads/2010/11/wp10-31.pdf

**Jung A and El-Shagi M (2015),** "Has the publication of minutes helped markets to predict the monetary policy decisions of the Bank of England"s MPC?", *European Central Bank Working Paper Series*, Vol. 1808, Available at

https://www.ecb.europa.eu/pub/pdf/scpwps/ebwp1808.en.pdf?72df01fc9a00e07fb70a717505ca5785

**Lucca D O and Trebbi F (2009),** "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements", *NBER Working Papers,* Vol. 15367. Available at https://www.nber.org/papers/w15367.pdf

**Murphy K P (2012),** "Machine learning: a probabilistic perspective", *MIT press*.

**Nyman R, Gregory D, Kapadia S, Ormerod P, Tuckett D and Smith R (2015**), "News and narratives in financial systems: exploiting big data for systemic risk assessment", *mimeo*. Available at https://www.norges-bank.no/contentassets/49b4dce839a7410b9a7f66578da8cf74/papers/smith.pdf

**Onsumran Ch, Thammaboosadee S and Kiattisin S (2015),** "Gold Price Volatility Prediction by Text Mining in Economic Indicators News", *Journal of Advances in Information Technology*, Vol. 6, No. 4. Available at http://www.jait.us/uploadfile/2015/1027/20151027113126355.pdf

**Shapiro A, Sudhof M and Wilson D (2017)**, "Measuring News Sentiment", Federal *Reserve Bank of San Francisco Working Paper Series,* Vol. 1. Available at https://www.frbsf.org/economic-research/files/wp2017-01.pdf

**Thorsrud L A (2018),** "Words are the new numbers: A newsy coincident index of business cycles", *Forthcoming in Journal of Business & Economic Statistics*. Available at https://doi.org/10.1080/07350015.2018.1506344

**Tuhkuri J (2016),** "Forecasting Unemployment with Google Searches", *ETLA Working Papers*, No 35. Available at http://pub.etla.fi/ETLA-Working-Papers-35.pdf

**Yu L, Zhao Y, Tang L and Yang Z (2018),** "Online big data-driven oil consumption forecasting with Google trends", *International Journal of Forecasting*, In press. Available at https://doi.org/10.1016/j.ijforecast.2017.11.005

# Big Data: Does it really improve Forecasting techniques for Tourism Demand in Spain?

Author: Miguel Ángel Ruiz Reina [1]

Programa de Doctorado: Economía y Empresa

Facultad de Ciencias Económicas y Empresariales – Universidad de Málaga (UMA)

- This study applies innovative forecasting techniques using Big Data, short-term and long-term for hotel demand in Spain.
- SARIMA and ARDL with seasonality needs strong mathematical assumptions.
- Singular Spectrum Analysis models weak mathematical assumptions.
- Matrix U1 Theil is proposed a novel technique for the selection of forecasting models in Big Data.
- The models have been tested with time horizon $h = 1, 3, 6, 12$
- The results are contrasted and statistically significant.

Abstract:

In this study, innovative forecasting techniques and data source from Big Data are used for the study of Hotel Overnight Stays for Spain, from January 2012 to December 2018. The unstoppable development of the tourism sector, together with the application of Big Data technologies, allow to make efficient decisions by economic agents. In this paper, univariate forecasting methodologies such as SARIMA and SSA are used. The use of the data obtained from the Google Data Mining tools allows to obtain knowledge. The ARDL models with seasonality explain easily when economic agents will make their decisions. ECM allows make forecasting for short-term and long-term. This fact means that tourist offers and demands can be perfectly adjusted at every moment of the year. As a criterion for the selection of models, the innovative Matrix U1 Theil is proposed, this allows to quantify how much a model is better than another in terms of forecasting.

Acknowledgements: The author wishes to acknowledge the support given by the University of Malaga. Ph.D. Program in Economics and Business, effective from July 16, 2013. Especially to Professor Antonio Caparrós Ruiz from the Department of Statistics and Econometrics of the University of Málaga, for reviewing this work.

Keywords:

ARDL; ECM; Singular Spectrum Analysis; Seasonality; Matrix U1 Theil; Forecasting; Tourism Demand; Spain; Google; Big Data

1 Corresponding author.
E-mail address: ruizreina@uma.es
Full Postal address: Universidad de Málaga (UMA), Facultad de Ciencias Económicas y Empresariales. Departamento de Economía Aplicada (Econometría y Estadística). Calle El Ejido, 6, 29071 Málaga, Spain.

## 1. Introduction

The use of Big Data technologies has meant a change in many data driven environments. The analysis of the "Tourism Industry" (Juul, 2015) has been one of the main objectives sectors due to the high volume of data that is generated. Spain as one of the main tourist markets at an international level is involved (Guevara Manzo & Turner , 2018). The emergence of positive and negative externalities of the high contribution of the tourism market to Spanish GDP is known to all stakeholders (Young Chung, 2009; Pegg, Patterson, & Vila Gariddo, 2012).

This paper introduces a modern analysis to obtain knowledge about the Spanish tourism market. Specifically, data from official sources are used and analysed together with secondary sources of Big Data from "Google Trends[2]".

In the predictive analysis, univariate techniques are used, such as the Seasonal Autoregressive Integrated Moving Average (SARIMA) methodology and a relatively novel Singular Spectral Analysis (SSA) forecast methodology (Hassani, 2007; Sun & Li, 2017; Golyandina & korobeynikov, 2018). These two univariate methodologies are compared with the multivariate method of Autorregresive Distributed Lags with seasonal variables (ARDL + seasonality). This multivariate method uses as an explanatory variable for hotel overnight stays in Spain a search interest rate (generated by Google Trends) and seasonal dummies variables for monthly data.

One of the novelties of this paper is the declaration of interest of future tourists in Spain. The demand for tourism in Spain through their searches on the Internet could be calculated. In addition to predicting tourism demand, with the application of dynamic and seasonal modelling we will be able to identify when consumer interest occurs. This second contribution is a very relevant fact, since tourism agents will be able to make efficient decisions in the tourism market. Ultimately, a criterion for the selection of new models, such as Matrix U1 Theil, has been developed. This Matrix will allow us to define whether or not the contributions of Big Data technologies improve the predictive capacity of the univariate SARIMA or SSA models.

The remainder of this research is as follows: Section 2 provides a review of the existing literature on the forecasting of Tourism Demand; in Section 3 data analysis is initially carried out along with the methodological development and information criteria. In section 4 an empirical analysis is carried out. Section 5 shows the final conclusions and future lines of research for Data Scientists. Finally, there is a section for the bibliographical references used.

## 2. Literature review

Tourism demand is caused by multiple exogenous factors. The interest of the researchers is remarkable due to their economic implication worldwide. There are numerous studies on the demand and tourist offer (Li, Song, & Wit, 2005; Song & Li, 2008; Peng, Song, & Crouch, 2014; Xiaoying Jiao & Li Chen, 2018). Traditionally these studies have been influenced by the techniques of the moment. In our study we will carry out an analysis with novel techniques and will be compared with most used techniques, a contribution of this study is the use of Big Data (Silva, Hassani, Heravi, & Huang, 2019) tools and that Google summarizes it in an index. Assuming an innovation in terms of the type of data used and knowledge of temporary geolocation that are provided.

Regarding the methodology used, the most used univariate methodology can be the one proposed by Box-Jenkins (2008). The second technique used is the non-parametric SSA technique being considered in studies since 2007 (Hassani, 2007; Saayman & Botha, 2015; Hassani, Webster, Simiral Silva, & Heravi, 2015; Hassani, Silva, Antonakakis, & Filis, 2017). In the case of ADRL models or other types of models used with search databases from Google, it has been used for analysis of epidemics (Plat, 2015); political results (González, 2017); stock market (Pyo, 2017); finances (Tkacz, 2013); retail forecasting, Automotive sales, Home Sales and Travel (Choi & Varian, 2009; Camacho & Pacce, 2017; Drago, 2017); Macroeconomics (Li, Shang, Wang, & Ma, 2015; Götz & Knetsch, 2017; Drago, 2017);

[2] http://www.google.com/trends

Unemployment (Tuhkuri, 2015; McKellips, 2017; Tuhkuri, 2017; Chancellor & Counts, 2018); consumer behaviour (Goel, Hofman, Lahaie, Pennock, & Watts, 2010) among others.

In the tourism sector it has been used for tourist demand in Amsterdam (Rödel, 2017) or a collaborative economy in the Iberian Peninsula (Palos-Sanchez & Correia, 2018). Another source of data is the Chinese search engine Baidu[3], which has been conducted similar studies of tourism demand (Tang, Qiu, & Liu, 2018). Previously, predictive analysis of data from the Chinese search engine between correlations and tourist flows was carried out (Lu, Zhao, Wu, & Hang, 2007; Li, Qiu, & Chen, 2008; Ma, Sung, Huang, & Zhou, 2011).

As observe above, the tourist industry has had interest in the past, in the present and in the future, and it will continue to have it. Mainly because it is an industry signal of the evolution of the service economy. So, the modelling used is very diverse, one aspect to be taken into account has been the criteria of information on the selection of models. It has been observed in the literature review the use of: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE); Theil index (Theil, 1958; Theil, 1966; Bliemel, 1973; Ahlburg, 1984) ; Symmetric Mean Percentage Error (SMAPE) (Amstrong, 1985; Flores, 1986; Tofallis, 2015). Some authors developed the RMSE ratio (Hassani, Webster, Simiral Silva, & Heravi, 2015; Hassani, Silva, Antonakakis, & Filis, 2017; Hassani., Silva, Grupta, & Das, 2018; Silva, Hassani, Heravi, & Huang, 2019) and in this article we will develop the Matrix U1 Theil as a criterion for the selection of forecasting models.

To summarize the literature review, we can say that novel models have been used in Data Science and an improved Ganger-Causality test is developed for data with seasonality. Big Data tools have been used from one of the largest search engines worldwide, and that a decision matrix on predictive capacity has been developed for different time horizons.

## 3. Methodology and Data

In this section, the scheme (Figure 1) of the cycle between supply and demand in tourism has been developed. Specifically, in our paper, the objective is modelling and forecasting, however we will suppose ad-doc the data from the Datawarehouse (Kimball, 2004; Dedić & Stanier , 2016). In this sense, the data will come from official sources of the INE[4] and Google[5]. So, all of Extraction, Transformation and Loading - ETL (Vassiliadis, 2009; Dunning & Friedman, 2014) work will come from the data engineering of these entities. The main objective is to make efficiencies predictions based on knowledge to improve the user experiences of tourism demand and the offers of the stakeholders.

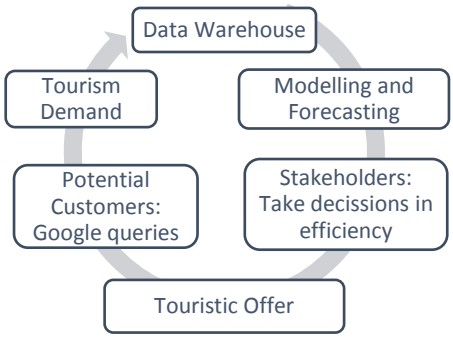



*Figure 1 Scheme of decisions in efficiency*

3 http://www.baidu.com
4 INE: Instituto Nacional de Estadística (Spain). The National Statistics Institute (Spain). www.ine.es
5 www.google.com

## 3.1. Modelling and Forecasting.

### 3.1.1. SARIMA model

The use of the methodology proposed by Box-Jenkins in the 70s (Box, Jenkins, & Reinsel, 2008) and its subsequent development is of great relevance in the scientific field. In this article, the TRAMO-SEATS tool will be used as support. TRAMO-SEATS (Gómez & Maravall, 1997) has shown better forecasts with seasonality for example than X-12 ARIMA (Vergori, 2010; 2012).

A generic scheme on modelling could be:

$$\phi_p(B^p)\Phi_P(B^P)\nabla^d\nabla_s^D Y_t{}^6 = \theta_q(B^q)\Theta_Q(B^Q)\varepsilon_t \qquad (1)$$

Where $p$ and $q$ represents nonseasonal-ARIMA order, and $d$ is the number of seasonal differences ($P, D, Q$ represent seasonal order of ARIMA) $B$ is the backshift operator and could be defined $Y_t(B^s) = Y_{t-s}$ . Due to the widespread use in the literature (see Literature review), this is an introductory section of this methodology. More information in Box, Jenkins, & Reinsel (2008).

### 3.1.2. Singular Spectrum Analysis (SSA)

In SSA analysis could be summarized (Golyandina, Korobeynikov, & Zhigljavsky, 2018):

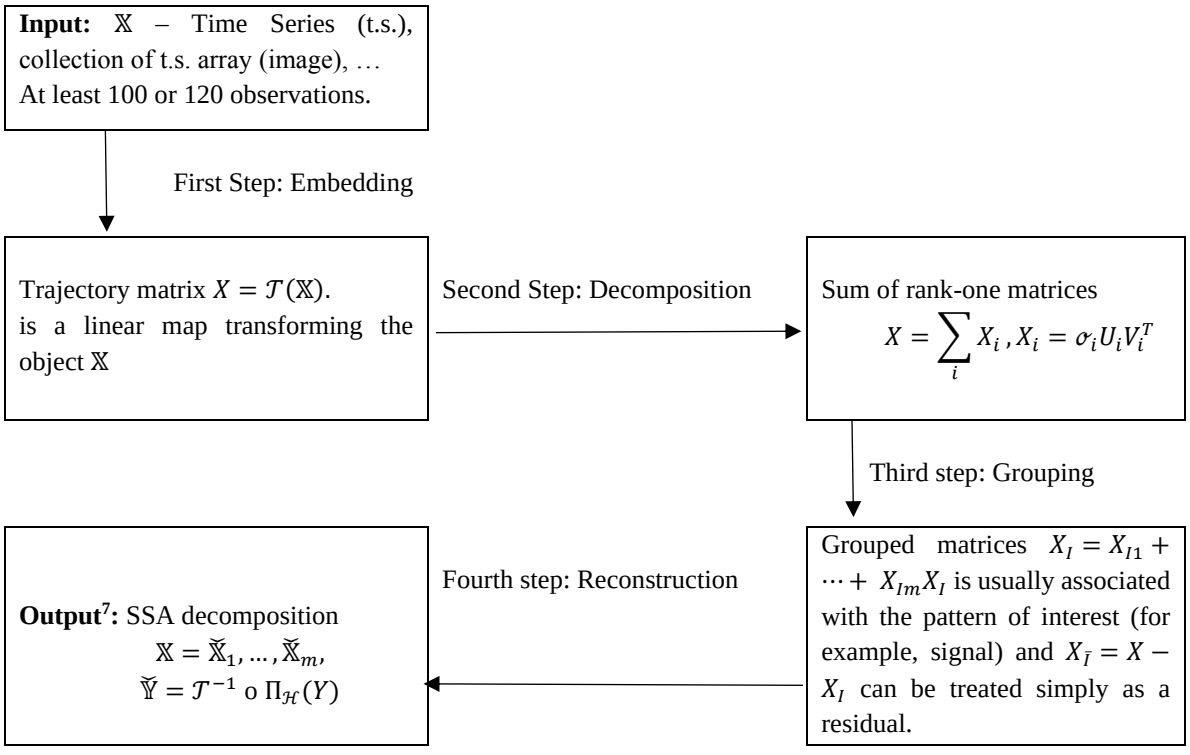



*Figure 2 SSA family: Generic Scheme*

7 If the grouping is elementary, then reconstructed objects $\tilde{\mathbb{X}}_k$ are called elementary components. For convenience of referencing. Step 1 and 2 of the generic SSA scheme are sometimes combined into the so-called "Decomposition stage" and Steps · and 4 are combined into "Reconstruction stage" (Hassani, 2007).

### 3.1.3. Autorregresive Distributive Lags (ARDL)

This subsection is divided into two blocks, on the one hand demonstrating, through the Granger-Causality test, the explanation that the dependent variable is explained by an explanatory variable and its lags in addition to the seasonal component.

### 3.1.4. Granger Causality and Seasonality testing

In the applied analysis of correlation does not imply causality, we develop the contrast proposed by Granger (1969) and discussed by Montero (2013).

The model considered by Granger is for two variables $(y_t, x_t)$. Due to the great influence of seasonality (Young Chung, 2009; Vergori, 2012) in the tourism sector, the following equation is proposed with HAC covariance method:

$$\ln(y_t) = b_0 \ln(x_t) + \sum_{j=1}^{m} b_j \ln(x_{t-j}) + \sum_{j=1}^{m} a_j \ln(y_{t-j}) + \sum_{i=2}^{12} \delta_i w_{t-i} + \varepsilon'_t \qquad \varepsilon' \sim White\ Noise \tag{2}$$

The decision of causality with seasonal effects (Testing linear restrictions for parameters of $x_{t-j}$ and $w_{t-j}$) is taken based on F-Fisher-Snedecor [8] ($T < 60$) or asymptotically ($T \geq 60$) as Chi-squared in otherwise (Buse, 1982)

$$F_{obs} = \frac{SSE_{restricted} - SSE_{unrestricted}}{SSE_{unrestricted}} * \frac{T-k}{q} \sim F_{q;T-k}$$

$$Chi - squared_{obs} = qF_{obs} \sim Chi - squared_q$$

### 3.1.5. ARDL and Error Correction Model (ECM)

The most general expression of a dynamic model named Autoregressive Distributive Lags - ARDL[9] $(m, n)$ with seasonal components is as follows (Hylleberg, Engle, Granger, & Yoo, 1990; Nkoro & Uko, 2016):

$$\gamma(L) \ln(y_t) = \mu + \delta(L) \ln(x_t) + \sum_{j=2}^{12} \alpha_j w_j + \varepsilon_t \qquad \varepsilon_t \sim White\ Noise \tag{3}$$

Where $w_j$ is a dummies deterministic seasonal component[10].

$$j \begin{cases} j = 1\ if\ month\ 1 \\ j = 2\ if\ month\ 2 \\ j = 3\ if\ month\ 3 \\ \vdots \\ j = 12\ if\ month\ 12 \end{cases} \quad w_j \begin{cases} w_1 = -1, for\ others\ w_j = 0 \\ w_1 = -1, w_2 = 1\ for\ others\ w_j = 0 \\ w_1 = -1, w_3 = 1\ for\ others\ w_j = 0 \\ \vdots \\ w_1 = -1, w_{12} = 1\ for\ others\ w_j = 0 \end{cases}$$

With the interest of evaluating the dynamic persistence of an effect on the exogenous variable at a certain moment, the Error Correction Model (ECM regression or ARDL Error Correction Regression). However, we must be cautious because the number of observations (Lehmann & Casella, 1998) does

8 T: observations included; K parameters estimated; q: numbers of restrictions.

9 $m$ is the number of exogenous variables $(y_t)$; $n$ is the number of endogenous variables $(x_t)$. ln is the Natural Logarithm. $(L)$ is the Lag operator. Stability conditions: if Inverted roots are $|\gamma(L)| < 1$.

10 In modelling, the variable $w_1$ is eliminated to avoid problems of perfect multicollinearity.

not exceed 8 years (Otero, 1989) in length and using a deterministic trend could be very restrictive (Harvey, 1997). The ECM[11] regression is as follows:

$$\Delta \ln(y_t) = \delta_0 \Delta \ln(x_t) + \sum_{j=1}^{n} \lambda_j \Delta \ln(x_{t-j}) + \sum_{j=1}^{m} \delta_j \Delta \ln(y_{t-j}) - \gamma(L)[y_{t-1} - \beta^{12} x_{t-1}] + \sum_{j=2}^{12} \alpha_j w_j + \varepsilon_t \quad (4)$$

In this model short term effect is represented by parameters of first variables differentiated, long term effect $\gamma(L)$ is represented by Correction Error term. According to Zivot (2000), if long term effect is not significant statically cointegration does not exist.

### 3.1.6. Forecasting Evaluation

Once the three predictive models are estimated, we will proceed to decision criteria. A criterion that could be considered usual in the literature as Root Mean Squared and another novel based on a decision matrix is proposed.

#### 3.1.6.1. Root Mean Square Error (RMSE)

Most used definition is as follows:

$$RMSE = \sqrt{\sum_{l=1}^{n} \left( \frac{y_{T+h} - \hat{y}_{T+h}}{h} \right)^2} \quad (5)$$

Where $y_{T+h}$ represents data reserved to compare with forecasting obtained $(\hat{y}_{T+h})$ through different methods proposed, both h step ahead from training sample.

The main disadvantage of RMSE criterion that it is a dimension measure, in this way the decision could be affected by the units of measurement of the variables.

#### 3.1.6.2. Theil´s measures

To improve the possible effect on the decision making affected by the units of measurement, we will work with the inequality index of Theil (Theil, 1958):

$$U_1 = \frac{\left[\frac{1}{n}\sum_{i=1}^{N}(y_{T+h} - \hat{y}_{T+h})^2\right]^{1/2}}{\left[\frac{1}{n}\sum_{i=1}^{N}(y_{T+h})^2\right]^{1/2} + \left[\frac{1}{n}\sum_{i=1}^{N}(\hat{y}_{T+h})^2\right]^{1/2}} \quad (6)$$

Ratio Theil´s (RT´s) is designed to comparisons between predicted variables

$$RT's_{y_{it},y_{jt}} = \frac{U_1^{y_{it}}}{U_1^{y_{jt}}}$$

11 Granger-Engle representation theorem and parameters are estimated in two stages (1987). Consistency and Efficiency of estimators are fulfilled.

12 $\beta = \frac{\delta(L)}{\gamma(L)}$: Total variation experimented by endogenous variable $(y_t)$ as consequence of a unitary change in exogenous variable $(x_t)$.

Matrix U1 Theil is defined as follows[13] :

$$Matrix\ RT's_{y_{it},y_{jt}} = \begin{pmatrix} RT's_{y_{1t},y_{1t}} & RT's_{y_{1t},y_{2t}} & \cdots & RT's_{y_{1t},y_{jt}} \\ RT's_{y_{2t},y_{1t}} & RT's_{y_{2t},y_{2t}} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ RT's_{y_{it},y_{1t}} & RT's_{y_{it},y_{2t}} & \cdots & RT's_{y_{jt},y_{jt}} \end{pmatrix}$$

The aim of the $Matrix\ RT's_{y_{it},y_{jt}}$ is to make comparisons between predicted variables and identify in which variable partially better results have been obtained.

$$U_1^{y_{it}}\ Preferred\ U_1^{y_{jt}}\ if\ RT's_{y_{it},y_{jt}} < 1$$

$$U_1^{y_{jt}}\ Preferred\ U_1^{y_{it}}\ if\ RT's_{y_{it},y_{jt}} > 1$$

$$U_1^{y_{jt}}\ Indiferent\ U_1^{y_{it}}\ if\ RT's_{y_{it},y_{jt}} = 1$$

## 3.2. Data

Data[14] of number of Hotel Overnight Stays in Spain (named "Spain") have been used provided by INE[15]. For number of tourists in Spain, overnight stays from first month of 2012 to December of 2018 were obtained. Researches for Tourism in Spain used this official source of data (Garín Muñoz, 2007; Martín Martín, Jiménez Aguilera, & Molina Moreno, 2014; Cisneros-Martinez & Fernández-Morales, 2015).

According to the data of average total monthly hotel occupancy obtained from Figure 1 in Spain were 25.978.054 in the period cited. The maximum number of hotel occupancy was recorded in August 2017 with 46.657.187 and the minimum 11.887.105 in January 2013.

To obtain data from Google, the Big Data tool called Google Trends has been used. This tool provides an "volume index word" among the countless correlated data generated in the network. In our model the data collection is ad-doc being developed by Google engineers the summary index. Specifically, we will use the keyword "visit Spain" as the explanatory variable. This word summarizes clusters of locations, images, videos, searches in text, and in general, all the words worldwide that are related to "visit Spain". This index fluctuates between 0 and 100, showing 0 the minimum interest and 100 the maximum interest of the keyword "visit Spain" (Google support, 2019). Previously Google Trends tools have been used to make forecasts (Choi & Varian, 2009; Lin & Chen, 2009; Lim, Alananzeh, & Hua, 2019).

The data series is not too long due to the improvements (Google has changed searching algorithms since 2004), considering the author criterion the best descriptor of the data generation process is from the beginning of 2012. Analysing the data obtained of interest for the word "visit Spain", for the date of May 2017 there was the greatest worldwide interest of the word just with three periods of advance to the maximum historical overnight stays in Spain. On the other hand, the lowest interest occurred in November and December of the year 2012. analysing the data obtained of interest for the word "visit Spain", for the date of May 2017 there was the greatest worldwide interest of the word just with three periods of advance to the maximum historical overnight stays in Spain. On the other hand, the lowest interest occurred in November and December of the year 2012. These two periods preceded the worst data of the historical sample selected of overnights stays in Spain.

13 The axioms of completeness, transitivity and rationality are completed (Villar, 1999)

14 No missing values in sampling.

15 INE: Instituto Nacional de Estadística (Spain). The National Statistics Institute (508/2001 of 11 May)

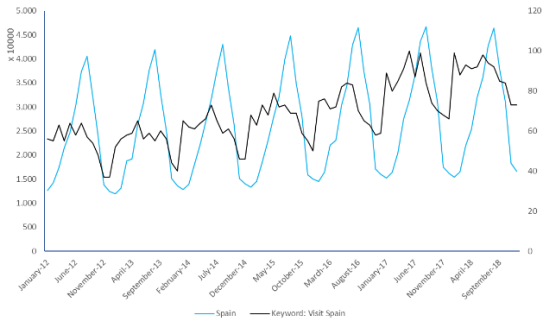



*Figure 3 number of hotel overnight stays and Keyword "visit Spain" (January 2012 to December 2018)*

With the observation of the maximum and minimum values of both series analysed, it is observed graphically that searches on the Internet are made with at least one period in advance to which the hotel reservation is made.

| | *Mean* | *Max.* | *Min* | *Standar Desv.* | *Skew.* | *Kurt.* | *Jarque- Bera (Prob.)* | *ADF (Prob.)* | *KPSS* |
|---|---|---|---|---|---|---|---|---|---|
| *Spain* | 25,78,054 | 46,657,187 | 11,887,105 | 10,394,564 | 0.38 | 1.88 | 6.34 (0.042) | -0.61 (0.86) | 0.16 |
| *"Visit Spain"* | 68.37 | 100 | 37 | 15.09 | 0.27 | 2.48 | 1.94 (0.37) | 1.50 (0.99) | 1.06 |

*Table 1 Stationary and Descriptive Analysis of Spanish hotel overnight stays and Keyword "visit Spain"*

From table 1 it is worth noting that the variable "Spain" does not fulfil the hypothesis of normality (Jarque-Bera) and is not stationary in terms of variance (KPSS). On the other hand, the index variable "visit Spain" fulfils the assumptions. From the result of the ADF test it is deduced that both series are integrated in order 1, so we can justify the use of Cointegration models expressed in terms of ECM.

## 4. Empirical results

In this paper of predictive techniques, we will focus expressly on the dynamic model with explanatory variables of Internet searches ("visit Spain") and seasonal factors. The univariate models SARIMA and SSA are used as a comparative tool for forecasting. Take into account that forecasting modelling has been carried out for short-term time horizons ($h = 1, 3, 6, 12$) as literature uses (Silva, Hassani, Heravi, & Huang, 2019). In the present article is considered training period January 2012-December 2017 and out-sample period January-December 2018.

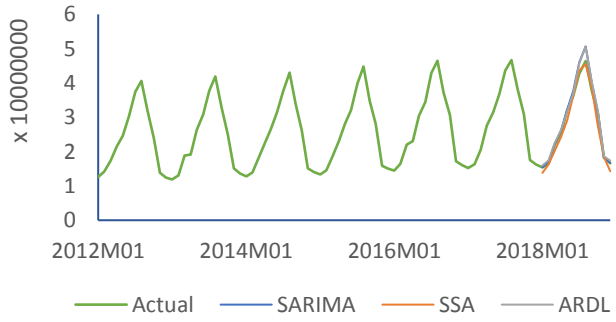



*Figure 4 Out-Sample forecast Hotel Occupancy in Spain h=12*

The results obtained through the Granger-Causality test including seasonal factors have determined that the number of hotel overnight stays in Spain could be explained by the number of searches

generated on the internet and by a systematic seasonality. The model ARDL with seasonality obtained[16]:

$$\ln(y_t) = 12.16 + 0.17\ln(x_t) + 0.22\ln(x_{t-1}) + 0.18\ln(y_{t-1}) + \sum_{j=2}^{12}\hat{\alpha}_j w_j + \hat{\varepsilon}_t$$

$$R^2 = 0.9932;\ \bar{R}^2 = 0.9915; Log.Likelihood = 140.20$$

To express the long-term dynamic effects, we can express the model in terms of ECM

$$\Delta\ln(y_t) = 12.16 + 0.17\,\Delta ln(x_t) - 0.81[y_{t-1} - 0.48x_{t-1}] + \sum_{j=2}^{12}\hat{\alpha}_j w_j + \hat{\varepsilon}_t$$

$$R^2 = 0.9802;\ \bar{R}^2 = 0.9760; Log.Likelihood = 140.20$$

In both model's seasonality parameters are as follows:

$$\sum_{j=2}^{12}\hat{\alpha}_j w_j = -0.38w_2 - 0.13w_3 - 0.02w_4 + 0.13w_5 + 0.23w_6 + 0.42w_7 + 0.49w_8 + 0.27w_9 + 0.09w_{10} - 0.37w_{11} - 0.31w_{12}$$

Once the results of the three forecasting models cited in the methodology have been obtained, the Matrix U1 Theil can be applied to quantify which model is better in predictive terms.

| *h=1* | *SARIMA* | *SSA* | *ARDL* |
|---|---|---|---|
| *SARIMA* | *1,0000* | *0,0082* | *0,0254* |
| *SSA* | *121,7259* | *1,0000* | *3,0915* |
| *ARDL* | *39,3743* | *0,3235* | *1,0000* |
| *h=3* | *SARIMA* | *SSA* | *ARDL* |
| *SARIMA* | *1,0000* | *0,1965* | *0,4262* |
| *SSA* | *5,0901* | *1,0000* | *2,1693* |
| *ARDL* | *2,3464* | *0,4610* | *1,0000* |
| *h=6* | *SARIMA* | *SSA* | *ARDL* |
| *SARIMA* | *1,0000* | *0,4225* | *1,0228* |
| *SSA* | *2,3667* | *1,0000* | *2,4205* |
| *ARDL* | *0,9777* | *0,4131* | *1,0000* |
| *h=12* | *SARIMA* | *SSA* | *ARDL* |
| *SARIMA* | *1,0000* | *0,9670* | *1,0016* |
| *SSA* | *1,0341* | *1,0000* | *1,0358* |
| *ARDL* | *0,9984* | *0,9654* | *1,0000* |

*Table 2 Matrix Theil´s*

For a period, *h=1* we must highlight the superiority of the SARIMA models over ARDL and SSA. As the models expand their forecasting time horizon until *h=12* months they are equalizing their forecasting capacity ($RT's_{SARIMA,ARDL} = 1.0016$) . One aspect to be taken into account is that the ARDL model with seasonality shows similar accuracy capacity to SARIMA, with the difference that the first implies a causality analysis by the searches produced on the internet. Also comment on the improvement in the predictive capacity of the SSA model as the time horizon of the prediction has been extended. However, the accuracy capacity of SSA against SARIMA and ARDL is lower. So, we could consider this technique as the least advantageous.

16 Model and lags selected under Akaike Info Criterion (1974). Model´s residuals obtained are empirically demonstrated as white noise $\hat{\varepsilon}_t \sim N(0, \sigma^2_{\hat{\varepsilon}_t})$. All parameters are significant with 95% of confidence except $\hat{\gamma}_1$, $\hat{\alpha}_4$, $\hat{\alpha}_{10}$ and have been used theorically to construct ECM model.

## 5. Conclusion

In this paper the importance of Forecasting modelling and historical analysis carried out in the literature review has been highlighted. In addition, this article has used more common techniques (SARIMA or ARDL) with a novel technique named SSA (Hassani & Mahmoudvand, 2018; Golyandina & korobeynikov, 2018).

The use of primary data source (INE) and secondary (Google) have allowed build knowledge based on the data. This last one is a novel aspect in the analysis, since the users show their interest through the search of information in Internet through web pages, social networks, images or videos. This supposes a relevant change of trend on the traditional surveys on consumer interests.

The contribution, in particular, can be divided into the following points:

1) A Granger-Causality test has been developed. Causality including seasonality, in the literature it was usual to perform the contrast between endogenous and exogenous variable. In our study we developed the contrast including seasonality.

2) A criterion of model's selection based on the predictive capacity of the models has been developed, since the Matrix U1 Theil allows to quantify the differences between models used.

3) Related to the previous point, we have deduced that the causal models slightly improve the results according to the time horizon is greater. We have found better results when horizon of forecasting is growing up.

4) In relation to the dynamic models with seasonality, we have empirically demonstrated that hotel demand decisions are made with at least a period of time in advance. At the same time, the current overnight stays depend on the current searches along with those that were made in the past of a period. Another important aspect, related to the lags is that the current overnight stays have a dilated behaviour over time expressed in the ECM model.

Answering the initial question of the paper: "Big Data: Does it really improve Forecasting techniques for Tourism Demand in Spain?", We can conclude that it does improve the forecast and also provides temporal and seasonal knowledge on the decision making of hotel demand in Spain.

The econometric interpretation of causality models is simpler than that of the two univariate models. This economic interpretation can facilitate an adjustment of the offer in terms of prices or even advertising to the agents interested in visiting Spain.

This article has been the basis of future research in which data from Big Data technologies are used to make efficiency decisions. An example may be the study by nationalities of origin for Spanish Tourism Demand. The theoretical framework being the same but expanding the data by countries of origin.

## References


Ahlburg, D. (1984, July/ September). Forecast evaluation and improvement using theil's decomposition. *Journal of Forecasting*, 345-351. Retrieved from https://doi.org/10.1002/for.3980030313

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on Automatic Control*, 716-723.

Amstrong, J. (1985). Long-range Forecasting: From Crystal Ball to Computer. Wiley.

Bliemel, F. W. (1973, November). Theil's Forecast Accuracy Coefficient: A Clarification. *Journal of Marketing Research, 10*(4), 444-446. Retrieved from https://www.jstor.org/stable/3149394?seq=1#page_scan_tab_contents

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis, Forecasting and Control.* United States of America: Wiley.

Box, G., Jenkins, G., & Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control* (Fourth edition ed.). John Wiley & Sons, Inc.

Buse, A. (1982, August). The Likelihood Ratio, Wald, and Langrange Multiplier Test: An Expository Note. *The American Statitician, 36*(3), 153-157. Retrieved from https://www.stat.washington.edu/jaw/COURSES/580s/581/HO/Buse.AmerStatist.82.pdf

Camacho, M., & Pacce, M. (2017). Forecasting travellers in Spain with Google's search volume indices. *Tourism Economics*. doi:DOI: 10.1177/1354816617737227

Chancellor, S., & Counts, S. (2018). Measuring Employment Demand Using Internet Search Data. *CHI '18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* Montreal.

Choi, H., & Varian, H. (2009). *Predicting the Present with Google Trends.* Google Inc. Retrieved from https://static.googleusercontent.com/media/www.google.com/es//googleblogs/pdfs/google_predicting_the_present.pdf

Cisneros-Martinez, J., & Fernández-Morales, A. (2015). Cultural tourism as tourits segment for reducing seasonality in a coastral area: the case study oif Andalusia. *Current Issues in Tourism, Vol. 18*(Num. 8), pags 765-784. doi:10.1080/13683500.2013.861810

Dedić, & Stanier . (2016). An Evaluation of the Challenges of Multilingualism in Data Warehouse Development. *18th International Conference on Enterprise Information Systems - ICEIS 2016*, (p. 196).

Drago, C. (2017). Forecasting the Measured Perceived Touristic Interest Using Autoregressive Neural Networks and Big Data: the Case of Florence. *AIQUAV.*

Dunning, T., & Friedman, E. (2014). *Time Series Databases: New Ways to Store and Access Data.* O'Reilly Media. Retrieved from http://shop.oreilly.com/product/0636920035435.do

Engle, R., & Granger, C. (1987). Co-Integration and error correction: representation, estimation and testing. *Econometrica*, pag. 251-276.

Flores, B. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega, 14*(2), 93-98. Retrieved from https://doi.org/10.1016/0305-0483(86)90013-7

Garín Muñoz, T. (2007). German demand for tourism in Spain. *Tourism Management*(28), 12-22.

Goel, S., Hofman, J., Lahaie, S., Pennock, D., & Watts, D. (2010, September). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*. doi:https://doi.org/10.1073/pnas.1005962107

Golyandina, N., & korobeynikov, A. (2018). *Singular Spectrum Analysis with R.* Springer. Retrieved from https://doi.org/10.1007/978-3-662-57380-8

Golyandina, N., Korobeynikov, A., & Zhigljavsky, A. (2018). *Singular Spectral Analysis with R.* Springer.

Gómez, V., & Maravall, A. (1997). *Programs TRAMO (Time Series Regression with ARIMA Noise, Missing Observations, and Outliers) and SEATS ((Signal Extraction in ARIMA Time Series). Instructions for the User.*

González, R. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. *Anthropology Today, 33*(3), 9-12. Retrieved from https://doi.org/10.1111/1467-8322.12348

*Google support*. (2019, May 16). Retrieved from https://support.google.com/trends/answer/6248105?hl=es&ref_topic=6248052

Granger, C. (1969). Investigating causal relations by econometric models and cross spectral methods. *37*(3), 424-438. doi:10.2307/1912791

Guevara Manzo, G., & Turner , R. (2018). *Travel & Tourism: Economic Impact Spain.* World Travel & Tourism Council (WTTC), London.

Harvey. (1997, January). Trends, Cycles and Autoregressions. *The Economic Journal, 107*(440), 192-201. Retrieved from https://academic.oup.com/ej/article-abstract/107/440/192/5144346

Hassani. (2007). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science*(5), 239-257.

Hassani, H., & Mahmoudvand, R. (2018). *Singular Spectrum Analysis.* Retrieved from https://doi.org/10.1057/978-1-137-40951-5

Hassani, H., Webster, A., Simiral Silva, E., & Heravi, S. (2015). Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. *Tourism Management*(46), 322-335.

Hassani, Silva, E. S., Antonakakis, N., & Filis, G. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research, 63*, 112-127.

Hassani., Silva, E., Grupta, R., & Das, S. (2018). Predicting global temperature anomaly: A definitive investigation using an ensemble of twelve competing forecasting models. *Physica A: Statistical Mechanics and its Applications, 509*, 121-139.

Hylleberg, S., Engle, R., Granger, C., & Yoo, B. (1990). Seasonal integration and cointegration. *Journal of Econometrics*(44), 215-238.

Juul, M. (2015). Tourism and The European Union: Recents Trends and Policy Developments. *EPRS | European Parliamentary Research Service*. doi:10.2861/310682

Kimball, R. (2004). *The Data Warehouse ETL Toolkit.* Willey.

Lehmann, E., & Casella, G. (1998). *Theory of Point Estimation* (Second edition ed.). New York Berlin Heidelberg: Springer. doi:ISBN 0-387-98502-6

Li, C., Song, H., & Wit, S. (2005). Recent Developments in Econometric Modeling and Forecasting. *Journal of Travel Research, 44*(1).

Li, S., Qiu, R., & Chen, L. (2008). Cyberspace attention of tourist attractions based on Baidu index: temporal distribution and precursor effect. *Geography and Geo-Information Science, 24*(6), 102-107.

Li, X., Shang, W., Wang, S., & Ma, J. (2015, March). A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications, 14*(2), 112-125. doi:10.1016/j.elerap.2015.01.001

Lim, C., Alananzeh, O., & Hua, K. (2019, January). Perceptions of Risk and Outbound Tourism Travel Intentions among Young Working Malaysians. *Human and Social Sciences, 46*(1), 365-379. Retrieved from https://www.researchgate.net/publication/330741416_Perceptions_of_Risk_and_Outbound_Tourism_Travel_Intentions_among_Young_Working_Malaysians

Lin, L., & Chen, Y. (2009). study on the influence of purchase intentions on repurchase decisions: the moderating effects of reference groups and perceived risks. *Tourism Review, 64*(3), 28-48. doi:doi:10.118/16605370910988818

Lu, Z., Zhao, Y., Wu, S., & Hang, B. (2007). The time distribution and guide analysis of visiting behavior of tourism website user. *Acta Geographica Sinica*, 621-630.

Ma, L., Sung, G., Huang, Y., & Zhou, R. (2011). A correlative analysis on the relationship between domestic tourists and network attention. *Economic Geography, 31*(4), 680-685.

Martín Martín, J., Jiménez Aguilera, J., & Molina Moreno, V. (2014). mpacts of Seasonality on Environmental Sustainability in the Tourism Sector Based on Destination Type: An Application to Spain'S Andalusia Region. *Tourism Economics, 20*(1), 123-142.

McKellips, F. (2017). *Nowcasting the Unemployment Rate in Canada Using Google Trends Data.* Retrieved from http://ifsd.ca/web/default/files/Presentations/Reports/17012%20-%20Nowcasting%20Unemployment%20Rate%20with%20Google%20Trends%20-%20Final.pdf

Montero, R. (2013). Test de Causalidad. *Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España*.

Nkoro, E., & Uko, K. (2016). Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation. *Journal of Statistical and Econometric Methods*, 63-91.

Otero. (1989). *Modelos Econométricos y Series Temporales*. Editorial AC.

Palos-Sanchez, P., & Correia, M. (2018). The Collaborative Economy Based Analysis of Demand: Study of Airbnb Case in Spain and Portugal. *Journal of Theoretical and Applied Electronic Commerce Research, 13*(3), 85-98. doi:10.4067/S0718-18762018000300105

Pegg, S., Patterson, I., & Vila Gariddo, P. (2012). The impact of seasonality on tourism and hospitality operations in the alpine. *International Journal of Hospitality Management*(31), 659-666.

Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism. *Tourism Management*, 181-183. doi:http://dx.doi.org/10.1016/j.tourman.2014.04.005

Plat, A. (2015). *Data Science and Ebola.* Inaugural Lecture, Universiteit Leidenon . Retrieved from https://www.researchgate.net/publication/274744049_Data_Science_and_Ebola

Pyo, D.-J. (2017). Can Big Data Help Predict Financial Market Dynamics?: Evidence from the Korean Stock Market. *journal East Asian Economic Review*, 147-165. doi:http://dx.doi.org/10.11644/KIEP.EAER.2017.21.2.327

Rödel, E. (2017). *Forecasting tourism demand in Amsterdam with Google Trends.* Master Business Administration.

Saayman, A., & Botha, I. (2015, January). Non-linear models for tourism demand forecasting. *Tourism Economics, 23*(3), 1-28.

Silva, E., Hassani, H., Heravi, S., & Huang, X. (2019). Forecasting tourism demand with denoised neural networks. *Annals of Tourism Research*, 134-154. Retrieved from https://doi.org/10.1016/j.annals.2018.11.006

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting- A review of Recent research. *Tourism Management*. doi:10.1016/j.tourman.2007.07.016

Sun, M., & Li, X. (2017). Window length selection of singular spectrum analysis and application to precipitation time series. *application to precipitation time series, 17*(2), 306-317.

Tang, H., Qiu, Y., & Liu, J. (2018, June). Comparison of Periodic Behavior of Consumer Online Searches for Restaurants in the U.S. and China Based on Search Engine Data. *IEEE Access*. doi:10.1109/ACCESS.2018.2832196

Theil, H. (1958). *Economic Forecasts and Policy.*

Theil, H. (1966). *Applied Economic Forecasting*.

Tkacz, G. (2013). 387Predicting Recessions in Real-Time: Mining Google Trends and Electronic Payments Data for Clues. *Financial Services*.

Tofallis, C. (2015). A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation. *Journal of the Operational Research Society, 66*(8), 1352-1362.

Tuhkuri, J. (2015). *Big Data: Do Google Searches Predict Unemployment?* Master's Thesis, University of Helsinki.

Tuhkuri, J. (2017). *Big Data: Google Searches Predict Unemployment in Finland.*

Vassiliadis, P. (2009). A Survey of Extract–Transform–Load Technology. *International Journal of Data Warehousing & Mining, 5*(July-September), 1-27.

Vergori. (2010). La stagionalità della domanda di servizi turistici: un'analisi econometrica. *Economia dei Servizi, 1*, 29-50.

Vergori. (2012). Forecasting tourism demand: the role of seasonality. *Tourism Economics, 18*(5), 915-930. doi:10.5367/te.2012.0153

Villar, A. (1999). *Lecciones de Microeconomía.* Barcelona: Antoni Bosch, editor, S.A.

Xiaoying Jiao, E., & Li Chen, J. (2018). Tourism forecasting: A review of methodological developments over the last decade. *Tourism Economics, XX*(X), 1-24. doi:DOI: 10.1177/1354816618812588

Young Chung, J. (2009). SEASONALITY IN TOURISM: A REVIEW. *e-Review of Tourism Research (eRTR), 7*(5), 82-96.

Zivot, E. (2000, June). The Power of Single Equation Tests for Cointegration When the Cointegrating Vector Is Prespecified. *Econometric Theory, 16*(3), 407-439. Retrieved from https://www.jstor.org/stable/3533230?seq=1#page_scan_tab_contents

# Estimation of parameters and reconstruction of hidden variables for a semiconductor laser from intensity time series

Mikhail Prokhorov[1], Ilya Sysoev[2], Vladimir Khorev[2], and Vladimir Ponomarenko[1,2]

[1]Saratov Branch of the Institute of Radio Engineering and Electronics of Russian Academy of Sciences, Zelyonaya Street, 38, Saratov 410019, Russia, mdprokhorov@yandex.ru
[2]Saratov State University, Astrakhanskaya Street, 83, Saratov 410012, Russia

The problem of reconstructing mathematical models of dynamical systems from time series has a long history. As the dimension of the system increases, the reconstruction of its equations becomes more difficult. For example, for systems with time-delayed feedback, having an infinite-dimensional phase space, it is necessary to develop special methods of reconstruction from time series [1, 2]. The problem of reconstruction becomes more complicated if a time-delay system has hidden variables that are inaccessible for observation. At the same time, such a task is of practical interest, since the reconstruction of a mathematical model can be used as a method of indirect measurement of variables unavailable for observation.

In this paper, we propose a method for the reconstruction of a time-delayed feedback system, in which only one of the three dynamical variables is observable, and the other two variables are hidden, including a hidden variable with a time delay. The method is applied to the reconstruction of the system of Lang-Kobayashi equations, which describes the dynamics of a single-mode semiconductor laser with time-delayed feedback. Such lasers exhibit a wide variety of oscillation regimes, depending on the choice of parameters, and can be used for constructing secure communication systems. The confidentiality of chaotic communication systems based on lasers is mainly due to the difficulty for an eavesdropper to recover the transmitter parameters from a transmitted chaotic signal.

The model equations of the laser under study are as follows:

$$\begin{aligned} \dot{\rho}(t) &= F(t)\rho(t) + \eta\rho(t-\tau)\cos\left(\phi(t) - \phi(t-\tau) + \Omega\tau\right), \\ \rho(t)\dot{\phi}(t) &= \alpha F(t)\rho(t) - \eta\rho(t-\tau)\sin\left(\phi(t) - \phi(t-\tau) + \Omega\tau\right), \\ T\dot{F}(t) &= P - F(t) - \left(1 + 2F(t)\right)\rho^2(t), \end{aligned} \tag{1}$$

where $\rho(t)$ and $\phi(t)$ are the modulus and phase, respectively, of the complex electric field $E(t) = \rho(t)\exp(i\phi(t))$, $F(t)$ is the excess carrier number, the dot indicates differentiation over the time $t$, which is measured in the photon lifetime $\tau_p$, $T = \tau_s / \tau_p$ is the ratio of the carrier lifetime $\tau_s$ to the photon lifetime, $P$ is the dimensionless pumping current above threshold, $\tau$ is the ratio of the external cavity round-trip time and the photon lifetime, $\eta$ is the strength of the feedback, $\alpha$ is the linewidth enhancement factor, and $\Omega$ is the dimensionless angular frequency of the solitary laser.

We consider a typical situation for a physical experiment, when only the time series of laser intensity $I(t) = \rho^2(t)$ is available, and the phase $\phi(t)$ of the electric field and the excess carrier number $F(t)$ are hidden variables. The control parameters $P$ and $\eta$, which can be easily varied in the laser, are assumed to be unknown. Our goal was to reconstruct the parameters $P$ and $\eta$ and hidden variables $\phi(t)$ and $F(t)$ from the intensity time series $I(t)$. It should be noted that the delay time $\tau$ characterizing the optical feedback is also an important laser parameter. To recover $\tau$ from intensity time series, the autocorrelation function or delayed mutual information is usually used, which often give an overestimation of $\tau$ [3]. For a more accurate estimation of $\tau$, one can use a method based on a statistical analysis of extrema in the intensity time series [4] or a method based on the nearest neighbor analysis [5].

The main idea of the proposed method is to include the starting guesses for hidden variables in the number of unknown parameters of the model and assign only a small number of starting guesses for $\phi(t)$ on the delay time interval. Since the variable $\phi(t)$ describes the phase of the electric field, it is convenient to use the points of a harmonic function as the starting guesses for $\phi(t)$. To obtain the remaining initial conditions for the hidden variable $\phi(t)$, we use the interpolation of its trajectory with a cubic spline.

As the objective function, we use the sum of squares of the distances between the points of the model time series and observed time series of the laser intensity. To minimize the objective function, we find its gradient and the Hessian matrix, introduce the vector of corrections to the starting guesses, and again calculate the objective function. This procedure is repeated until the changes in the starting guesses are small enough.

It is shown that the proposed method allows one to reconstruct the time series of hidden variables and the unknown parameters of the Lang-Kobayashi equations using the scalar time series of chaotic and periodic oscillations of laser intensity. The dependence of the quality of the laser system reconstruction on the accuracy of the assignment of starting guesses for unknown parameters and hidden variables is investigated. It is shown that for periodic regimes, the region of starting guesses, which provides high quality of reconstruction, is greater than for chaotic regimes.

This work was supported by the Russian Foundation for Basic Research, Grant No. 19-02-00071.

## References

1. Prokhorov, M.D., Ponomarenko, V.I., Karavaev, A.S., Bezruchko, B.P. Physica D 203, 209 (2005).
2. Prokhorov, M.D., Ponomarenko, V.I. Phys. Rev. E 80, 066206 (2009).
3. Rontani, D., Locquet, A., Sciamanna, M., Citrin, D.S., Ortin, S. IEEE J. Quantum Electron. 45, 879 (2009).
4. Ponomarenko, V.I., Prokhorov, M.D., Koryukin, I.V. Tech. Phys. Lett. 31, 939 (2005).
5. Khorev, V.S., Prokhorov, M.D., Ponomarenko, V.I. Tech. Phys. Lett. 42, 146 (2016).

# WILL THE SPANISH REGIONS CONVERGE IN THE NEAR FUTURE?

Sofia Tirado Sarti[a], Rafael Flores de Frutos[b] and Manuel León Navarro[c]

## Abstract

This paper analyses interregional effects on capital stock in Spain in order to get results that allow us to evaluate the efficiency of public investment policies and provide new insights into the analysis of regional disparities.

To this aim, the Vector Autoregressive (VAR) methodology is used to estimate the dynamic effects of capital on output and employment and evaluate the relevance and magnitude of regional spillover effects.

The empirical results for Spanish economy suggest that there are two groups of regions. GDP and Employment of small Spanish regions seem to be I(1) variables, while their corresponding Capital Stocks seem to be I(2) variables. This contrasts with the statistical properties of the same variables for the Spanish big regions, all being I(2).This result implies that, for the smaller regions, only permanent changes in the rate of growth of Capital Stock can produce permanent effects on the levels of GDP and/or Labor. The same occurs when the investment takes place outside the smaller regions. A small region cannot beneficiate, in the long run, from investments, neither inside nor outside the region itself. These statistical properties suggest that a kind of circular cumulative effects, á la Myrdal, take place which could difficult the convergence (in terms of GDP) of small regions.

We conclude that to achieve real convergence in Spanish regions, a long sequence of big pushes in investment is required for the smaller regions to change their growth path and to become a development engine.

Key Words: Growth, Convergence, Integration orders, VECM models

[a]PhD in Economics and Professor, Department in Economics, CES Cardenal Cisneros, C/General Díaz Porlier, 58, 28006, Madrid.Spain. Tel:+34913096120, mailing address: stirados@universidadcisneros.es
[b]Full Professor,Department in Economics, CES Cardenal Cisneros, C/General DíazPorlier, 58, 28006, Madrid.Spain. Tel:+34913096120,mailing address: rfloresf@universidadcisneros.es
[c] PhD in Economics and Professor, Department in Economics, CES Cardenal Cisneros, C/General Díaz Porlier, 58, 28006, Madrid.Spain. Tel:+34913096120, mailing address: mleon@universidadcisneros.es

# Estimation of Vector Long Memory Processes

Hao Wu and Peiris Shelton
The University of Sydney (Australia)

## Abstract

Time series modelling has been shown to be an effective tool for analyzing data from macroeconomics and finance. There seems to be an increased interest to extend univariate models to multivariate case in various application domains. This paper provides an overview of the most important development in parametric multivariate long memory time series modelling, with estimation mainly based on a Gaussian likelihood. It discusses the model specification and estimation methodology for vector autoregressive fractionally integrated moving average (VARFIMA) model and vector Gegenbauer ARMA (VGARMA) model, in both methodology and empirical applications. It standardizes the state space representation of multivariate time series models and presents the simulation results of quasi-maximum likelihood estimator via Kalman filter. The performance of different estimation methods in both time domain and frequency domain are also compared. Finally, the likely directions of future research are identified.

# A robust method for estimating the number of factors in an approximate factor model

Higor Cotta[1,2], Valdério Reisen[1,2], and Pascal Bondon[2]

[1]NuMEs - DEST/PGGEA - Federal University of Espírito Santo - Brazil
[2]L2S - CentraleSupélec - France
higor.cotta@l2s.centralesupelec.fr
valderio.reisen@ufes.br
pascal.bondon@l2s.centralesupelec.fr

**Abstract.** This paper considers the approximate factor model for high-dimensional time series with additive outliers. We propose a robustification procedure of the information criteria proposed by [1]. The robust estimator of the number of factors is obtained by replacing the standard covariance matrix with $M$-covariance matrix. Simulations are carried out under the scenarios of multivariate time series with and without additive outliers to assess the impact of additive outliers on the standard information criteria and to analyze the finite sample size performance of the proposed robust estimator of the number of factors.

**Keywords:** Factor Analysis, High Dimension, Multivariate Time Series, Outliers, Robust Autocovariance Function

## 1 Introduction

Nowadays, thanks to the improvements in computer power and data storage capacity, data scientist have now the possibility to work and study high dimensional data sets. As time passes by more data is generated, the dimension increases and so does the number of parameters to be estimated of many statistical models. Therefore, new techniques that accommodate high dimensional data sets are needed. In this context, the factor analysis (FA) is, undoubtedly, one of the most used techniques employed by the analyst for summarizing information while reducing the dimension of a large amount of data.

The FA model assumes that the common factors are latent (not observed) and in order to the model be identifiable some assumptions about the underlying factor structure are required. This fact leads to the development of various factor models. In this direction, a common assumption is that the covariance matrix of the idiosyncratic component is diagonal and this is the starting point of the widely used orthogonal factor model. New factor models are created insofar this basic assumption is relaxed. The approximate factor model in the sense of [2] allows some correlation among the idiosyncratic component.

In this context, one possible approach to estimate the factor model is to assume normality and to use the maximum likelihood estimation or Kalman filter approaches. However, the assumption of normality may be too strong when working with applied data. Another drawback is the number of parameters to be estimated using the Kalman filter approach increases as the more variables are considered.

The approach considered here implements the framework of [1], which employs the principal component analysis (PCA) technique to estimate the latent factors. Nevertheless, the PCA tool is the most popular estimation method due to its performance and ease of use. As pointed out by many authors, the PCA method is sensitive to the occurrence of outliers among the collected data [3]. For example, [4] showed that the number of factors is influenced by the presence of additive outliers and proposed the use of a robust autocovariance function estimator to mitigate the effect of additive outliers.

In addition, to be of high dimension due to a large number of variables measured at air pollution monitoring stations scattered over different regions, it is well-known that air pollution data may

also present high peaks which may seem as outlying observations. In this scenario, the usual solution is to remove observations that are suspicious to be outliers, but doing so, one is tacitly implying the outliers to be errors which are not the case of most of the high peaks of air pollution time series as they may cause serious harms to the human health and environment.

Nonetheless, these high-level observations can be seen as aberrant values from a statistical point of view. In this direction, these outlying observation can directly affect the statistical properties of the standard estimates such as the sample mean and sample covariance which will affect any sub-sequential method, for example, the FA method making use of the standard PCA technique. In this scenario, the usual solution is to remove observations that are suspicious to be outliers but doing so, one is implicitly assuming the outliers to be errors. Thus, in this paper, robust estimators are proposed for tackling this common issue.

Therefore, this paper considers multivariate time series with additive outliers using the FA technique for dimension reduction where the number of factors is estimated using the criteria of [1]. In this context, it is here proposed and studied a robust version of the estimators given in [1].

The paper is organized as follows: besides the introduction, Section 2 introduces the model and the estimation procedure here considered. Section 3 discuss the impact of additive outliers on the factor model and presents a robust methodology in order to mitigate the effect of the outlying observations. Some Monte Carlo experiments are presented in Section 4. The application and the concluding remarks are in Sections **??** and 5, respectively.

## 2 Model and estimation

Let $N$, $N \in \mathbb{N}$, denotes the number of variables and $T$, $T \in \mathbb{N}$, the sample size. For, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, the observation $X_{it}$ is said to a have factor structure if it can be written as

$$X_{it} = \lambda_i' F_t + \epsilon_{it} = C_{it} + \epsilon_{it}, \tag{1}$$

where $F_t$ is a vector of common factors, $\lambda_i$ is a vector of factor loadings associated with $F_t$, and $\epsilon_{it}$ is the idiosyncratic component of $X_{it}$. $C_{it}$ called the common component of $X_{it}$.

The model is latent, i.e, the factors and their corresponding loadings, and the idiosyncratic component are not observable. If the factors were to be observable, the model could be easily estimated, for example, by multiple linear regression. Note that the $X_{it}$ has a contemporaneous relationship with $F_t$, thus (1) is referred as the static factor model. This is in contrast with the dynamic factor model, in which $X_{it}$ does not have a contemporaneous relationship with the factors. Dynamic factor models are studied by [5] [6] and [7], among others and are beyond the scope of this paper. In this paper, a robust estimator of the number of factors $r$, $r \in \mathbb{N}$ is proposed.

Let $F_t^0$ and $\lambda_i^0$ denote the true common factors and their corresponding factor loadings. Thus, (1) can be written as an $N$-dimensional time series with $T$ observations. At a time $t = 1, \ldots, T$,

$$X_t = \Lambda^0 F_t^0 + \epsilon_t, \tag{2}$$

where $X_t = (X_{1t}, \ldots, X_{Nt})'$, $\Lambda^0 = (\lambda_1, \ldots, \lambda_N)'$ and $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$.

(1) can also be written as a $T-$dimensional vector of random variables. For a given $i$,

$$X_i = F^0 \lambda_i^0 + \epsilon_i, \tag{3}$$

where $X_i = (X_{i1}, \ldots, X_{iT})'$, $F_t^0 = (F_1^0, \ldots, F_T^0)'$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$.

Finally, in matrix form,

$$X = F^0 \Lambda^{0\prime} + \epsilon, \tag{4}$$

where $X = (X_1, \ldots, X_N)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ are $T \times N$ matrices.

Model in (4) has the covariance structure of the static factor model is given by

$$\Sigma_X = \Lambda^0 \Sigma_F \Lambda^{0\prime} + \Sigma_\epsilon, \tag{5}$$

where $\Sigma_X$, $\Sigma_F$ and $\Sigma_\epsilon$ are $N \times N$ covariance matrices of $X$, $F^0$ and $\epsilon$, respectively.

Let $\mathrm{Tr}(A)$ and $||A|| = (\mathrm{Tr}(A'A))^{1/2}$ denote the trace and the norm of a matrix $A$, respectively. According to [1], in order for the factor model be identifiable the following assumptions are made:

(**A1**) $\mathbb{E}(||F_t^0||^4) < \infty$ and $T^{-1}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} \to \Sigma_F$ as $T \to \infty$ for some positive definite matrix $\Sigma_F$.

(**A2**) $||\lambda_i|| \leq \bar{\lambda} < \infty$, for some positive $\bar{\lambda}$ and $||\Lambda^{0\prime}\Lambda^0/N - D|| \to 0$ as $N \to \infty$ for some $r \times r$ positive definite matrix $D$.

(**A3**) There exists a positive constant $M < \infty$ such that for all $N$ and $T$,

1. $\mathbb{E}(\epsilon_{it}) = 0$, $\mathbb{E}(|\epsilon_{it}|^8) \leq M$;
2. $\mathbb{E}(\epsilon_s'\epsilon_t/N) = \mathbb{E}(N^{-1}\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}) = \gamma_N(s,t)$,$|\gamma_N(s,s)| \leq M$ for all $s$, and $T^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$;
3. $\mathbb{E}(\epsilon_{it}\epsilon_{jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some $\tau_{ij}$ and for all $t$, and $N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M$;
4. $\mathbb{E}(\epsilon_{it}\epsilon_{js}) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}N\sum_{t=1}^{T}\sum_{s=1}^{T}|\tau_{ij,ts}| \leq M$;
5. for every $(t,s)$, $\mathbb{E}(|N^{-1/2}\sum_{i=1}^{N}(\epsilon_{is}\epsilon_{it} - \mathbb{E}(\epsilon_{is}\epsilon_{it}))|^4) \leq M$.

(**A4**) $\mathbb{E}(N^{-1}\sum_{i=1}^{N}||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0\epsilon_{it}||^2) \leq M$.

Under assumptions (A2) to (A4) the factor model here considered is the approximate factor model in the sense of [8]. The approximate factor model in contrast with the standard orthogonal factor model, allows some correlation in the idiosyncratic component.

Many different approaches have been proposed by the literature to estimate the factor model. In a small $N$ setting, one may write the factor model in a state space form, assume normality and use the maximum likelihood approach. However, since the number of parameters increases with $N$, this approach requires an intensive computational effort. More details can be found in [9].

Another possible approach to estimate (4) is to consider the least square approach by minimizing the squared sum of the residuals. That is, the estimates of $\lambda$ and $F$ are obtained by solving the following optimization problem

$$V(\tilde{F},\tilde{\Lambda}) = \underset{\Lambda,F}{\mathrm{argmin}}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \tilde{\lambda}_i\tilde{F}_t)^2, \tag{6}$$

where $\tilde{F}$ and $\tilde{\lambda}$ are the hypothetical values of the factors and their corresponding loadings. Minimizing (6) in respect to $\tilde{F}$ is equivalent to maximizing $\mathrm{Tr}(\tilde{\Lambda}'X'X\tilde{\Lambda})$ subject to $\tilde{\Lambda}'\tilde{\Lambda}/N = I$. In this context, the solution of (6) is obtained by setting $\hat{\Lambda}$ equal to the eigenvectors corresponding to the $r$ largest eigenvalues of $X'X$. Thus, the PC estimator of $F$ is

$$\hat{F} = X'\hat{\Lambda}/N. \tag{7}$$

A common issue related to the big data context arrives when the number of variables $N$ is much larger than the number of samples $T$. In this scenario, the rank of $\hat{\Sigma}_X$ is no more than $\min\{N,T\}$. However, as noted by [2], one might use the eigenvectors associated with the $T \times T$ $XX'$ matrix. This approach is called asymptotic principal component analysis (ACPCA) and provides a consistent estimator of the common factors under the following additional assumptions

(**A5**)

1. $\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}\epsilon_{is} \to 0, n \neq s$;
2. $\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}^2 \to \sigma^2$, for all $t$, as $N \to \infty$.

Thus, concentrating out $\hat{\Lambda}$, minimizing (6) in respect to $\tilde{\Lambda}$ is equivalent to maximizing $\mathrm{Tr}(\tilde{F}'X'X\tilde{F})$ subject to $\tilde{F}'\tilde{F}/T = I$. In this context, the solution of (6) is setting $\tilde{F}$ equal to the eigenvectors corresponding to the $r$ largest eigenvalues of $XX'$ yields the The APC estimator of $F$.

As in [10] and [1] this paper considers the setting when both $N$ and $T \to \infty$. The space spanned by $\hat{F}$ and $\tilde{F}$ are equivalent. Therefore, they can be used interchangeably depending on the sizes of $N$ and $T$ to achieve a computationally simpler approach.

Now, the estimation of the number of factors is addressed. Supposing that the factors are observed, [1] proposed information criteria for the estimation of the number of factors in approximate factor models. They are

$$\begin{aligned} IC_{p1}(k) &= ln(\hat{V}_k) + k(\tfrac{N+T}{NT})\ln(\tfrac{NT}{N+T}); \\ IC_{p2}(k) &= ln(\hat{V}_k) + k(\tfrac{N+T}{NT})\ln(\min\{N,T\}^2); \\ IC_{p3}(k) &= ln(\hat{V}_k) + k(\tfrac{\ln(\min\{N,T\}^2)}{\min\{N,T\}^2}), \end{aligned} \tag{8}$$

where $\hat{V}_k$ is the minimized value of $V(.)$ in (6) and $k$ is a number of estimated factors.

Since outliers are present in the data, any approach that makes use of (6) will also be affected. Next session presents the approximate factor model with additive outliers and develops a robust methodology to coherently estimate the number of factors when additive outliers are present.

## 3 Outliers and robust estimation

It is supposed that the observed process $X_t$ results from the contamination of $Z_t$ by additive random outliers, i.e.,

$$X_t = Z_t + \Omega\delta_t, \tag{9}$$

where $\Omega = \text{diag}(\omega_1, ..., \omega_N)$ and $\omega_i, i = 1, ..., N$, is the magnitude of the outliers which affects $Z_{it}$, $\delta_t = (\delta_{1t}, ..., \delta_{Nt})'$ is a random vector indicating the occurrence of an outlier at time $t$. It is assumed that $X_t$ and $\delta_t$ are uncorrelated processes and that $\mathbb{P}(\delta_{it} = -1) = \mathbb{P}(\delta_{it} = 1) = p_i/2$, $\mathbb{P}(\delta_{it} = 0) = 1 - p_i$ for $i = 1, \ldots, N$ where $0 \leq p_i < 1$. Then $\mathbb{E}(\delta_{it}) = 0$ and $\mathbb{V}ar(\delta_{it}) = p_i$. It is also assumed that $\mathbb{C}ov(\delta_t, \delta_t) = \Sigma_\delta = \text{diag}(p_1, ..., p_N)$ and that $\mathbb{C}ov(\delta_t, \delta_{t+h}) = 0$ when $h \neq 0$.

It follows from (9) that the effects of additive outliers on the level of the process is $\mathbb{E}(Z_t) = \mathbb{E}(X_t)$. The effect of additive outliers on the autocovariance function of the process is $\Gamma_Z(0) = \Gamma_X(0) + \Omega\Sigma_\delta\Omega'$, with $\Sigma_X = \Gamma_X(0)$ and $\Sigma_Z = \Gamma_Z(0)$. $\Gamma_Z(h) = \Gamma_X(h)$ when $h \neq 0$.

In view of (4), the factor model with additive outliers is

$$Z = F\Lambda' + \epsilon + \Omega\delta. \tag{10}$$

As can be seen from (10), the outliers that additively influence $X$ are not within the factors. The model under study here is in accord with the outlier models considered by [3] and [11].

The effect of the additive outliers on the covariance structure of the factor model is

$$\Sigma_Z = \Lambda\Sigma_F\Lambda' + \Sigma_\epsilon + \Omega\Sigma_\delta\Omega'. \tag{11}$$

However, it is not possible to decompose and correctly eliminate the occurrence of additive outliers from the observed series $Z$ in a real data scenario. Therefore, the eigenvalues and their corresponding eigenvectors are affected, and, consequently, the number of factors as well the factors themselves. Therefore, in order to mitigate this issue, a robust methodology is here proposed.

Some approaches have been discussed in order to transform the standard factor model robust against additive outliers. From the optimization problem point of view, i.e., context of (6), one could replace the least square estimates by some robust alternative, e.g, a different loss function such as least absolute deviation ([12]), singular value threshold ([3] and [13]) or Huber loss function ([14]). The latter was considered by [15] in factor models to perform a robust regression of the data onto the observed covariates before carrying out the PCA estimation procedure.

It is here proposed to robustify $\hat{F}$ and $\tilde{F}$ by replacing the traditional $N \times N$ or $T \times T$ covariance matrices by their corresponding robustified version.

### 3.1 Robust estimation of the covariance matrix from the robust $M$-cross-periodogram

It is known that a given zero-mean stationary univariate time series $X_t, t = 1, \ldots, T$, can be represented as a sum involving $T$ sines and cosines at the Fourier frequencies $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$. The classical periodogram of $X_t$ at frequency $\lambda_k$ is

$$I_T^X(\lambda_k) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_t \exp(-\mathrm{i}t\lambda_k) \right|^2 .$$

As discussed in [16], one alternative way to derive the periodogram function $I_T^X(.)$ is based on the Least Square (LS) estimates of a bi-dimensional vector $\boldsymbol{\beta'} = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model

$$X_i = c'_{Ti}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)} \cos(i\lambda_j) + \beta^{(2)} \sin(i\lambda_j) + \varepsilon_i \ , \ 1 \leq i \leq T, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{12}$$

where $\varepsilon_i$ denotes the deviation of $X_i$ from $c'_{Ti}\boldsymbol{\beta}$ and $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] < \infty$. In the sequel, $(\varepsilon_i)$ is assumed to be a function of a stationary Gaussian process.

It supposed that

$$\varepsilon_i = G(\eta_i), \tag{13}$$

where $G$ is a non null real-valued and skew symmetric measurable function (*i.e.* $G(-x) = -G(x)$, for all $x$) and $(\eta_i)_{i\geq 1}$ is a stationary Gaussian process with zero mean and unit variance. Additional assumption of $(\eta_i)_{i\geq 1}$ is given in (A10).

It can be shown that

$$I_T^X(\lambda_k) = \frac{T}{8\pi} ||\hat{\boldsymbol{\beta}}(\lambda_k)||^2 = \frac{T}{8\pi} \left( \hat{\beta}_1(\lambda_k)^2 + \hat{\beta}_2(\lambda_k)^2 \right), \tag{14}$$

where $\hat{\boldsymbol{\beta}}(\lambda_k)$ is the least squares regression solution

$$\hat{\boldsymbol{\beta}}(\lambda_k) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{t=1}^{T} (X_t - C'_t(\lambda_k)\boldsymbol{\beta})^2, \tag{15}$$

with the regressors $C_t(\lambda_k) = [\cos(t\lambda_k), \sin(t\lambda_k)]'$.

The periodogram is robustified by replacing the least squares regression with the $M$-regression.

Let now $\psi(.)$ be a function satisfying the following assumptions:

(**A6**) $0 < \mathbb{E}[\psi^2(\varepsilon_1)] < \infty$ .

(**A7**) The function $\psi$ is absolutely continuous with its almost everywhere derivative $\psi'$ satisfying $\mathbb{E}[|\psi'(\varepsilon_1)|] < \infty$ and such that the function $z \mapsto \mathbb{E}[|\psi'(\varepsilon_1 - z) - \psi'(\varepsilon_1)|]$ is continuous at zero.

(**A8**) $\psi$ is non-decreasing, $\mathbb{E}[\psi'(\varepsilon_1)] > 0$ and $\mathbb{E}[\psi'(\varepsilon_1)^2] < \infty$.

(**A9**) $\psi$ is skew symmetric, *i.e.* $\psi(-x) = -\psi(x)$, for all $x$.

(**A10**) Let $\eta_t$, $t \in \mathbb{Z}$, be i.i.d. standard Gaussian random variables and let $a_j$ be real numbers such that $\sum_{j\geq 0} |a_j| < \infty$ and $a_0 = 1$. Then,

$$\varepsilon_i = \sum_{j\geq 0} a_j \eta_{i-j}.$$

(**A11**) $\psi$ is the Huber function that is $\psi(x) = \max[\min(x, c), -c]$, for all $x$ in $\mathbb{R}$, where $c$ is a positive constant.

The $M$-estimator $\hat{\boldsymbol{\beta}}_\psi(\lambda_k)$ is defined as the solution of

$$\sum_{t=1}^{T} C_t(\lambda_k)\psi(X_t - C_t'(\lambda_k)\boldsymbol{\beta}) = 0, \tag{16}$$

where $\psi$ is defined by

$$\psi(x) = \begin{cases} x, & \text{if } |x| \le c, \\ c\,\text{sign}(x), & \text{if } |x| > c, \end{cases}$$

and $c$ is some positive constant, see [16]. In the following, $c = 1.345$ is adopted to ensure an efficiency of 95% for the regression estimator in Gaussian case.

Similarly to (14), the robust $M$-periodogram is defined by

$$I^X_{M,T}(\lambda_k) = \frac{T}{8\pi}||\hat{\boldsymbol{\beta}}_\psi(\lambda_k)||^2 = \frac{T}{8\pi}\left(\hat{\beta}_{1,\psi}(\lambda_k)^2 + \hat{\beta}_{2,\psi}(\lambda_k)^2\right). \tag{17}$$

For the univariate context, the asymptotic properties of $\boldsymbol{\beta}_\psi$ are established for the short and long-range dependence frameworks in [16], [17] and [18] .

Let now $X_1, Y_2, \ldots, X_T$ be a sample observation of a bivariate, $N = 2$, second order stationary time series $X_t$. The cross-periodogram is defined by

$$I^X_{T,ij}(\lambda_k) = \frac{1}{2\pi}\sum_{h=-(T-1)}^{T-1} \hat{\gamma}^X_{T,ij}(h)\exp(-\mathrm{i}h\lambda_k), \tag{18}$$

where, $i, j = 1, 2$, and $\hat{\gamma}^X_{T,ij}(.)$ is the standard sample estimator of the cross-covariance function.

In view of (14), the cross-periodogram at frequency $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$. defined by (18) may be written as by

$$I^X_{T,ij}(\lambda_k) = \begin{cases} \frac{T}{2\pi}\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{1,X_j}(\lambda_k) & \lambda_k = 0 \\ \frac{T}{8\pi}(\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{1,X_j}(\lambda_k) + \hat{\beta}_{2,X_i}(\lambda_k)\hat{\beta}_{2,X_j}(\lambda_k) - & \\ \qquad \mathrm{i}(\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{2,X_j}(\lambda_k) - \hat{\beta}_{1,X_j}(\lambda_k)\hat{\beta}_{2,X_i}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2,} \end{cases}$$

where $\hat{\beta}_{1,X_i}(\lambda_k)$ and $\hat{\beta}_{2,X_i}(\lambda_k)$ are defined by (15) and $X_t$ is replaced by $X_{it}$, $i = 1, 2$.

Likewise, the $M$-cross-periodogram is defined by

$$I^X_{M,T,ij}(\lambda_k) = \begin{cases} \frac{T}{2\pi}\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{1,X_j,\psi}(\lambda_k) & \lambda_k = 0 \\ \frac{T}{8\pi}(\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{1,X_j,\psi}(\lambda_k) + \hat{\beta}_{2,X_i,\psi}(\lambda_k)\hat{\beta}_{2,X_j,\psi}(\lambda_k) - & \\ \qquad \mathrm{i}(\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{2,X_j,\psi}(\lambda_k) - \hat{\beta}_{1,X_j,\psi}(\lambda_k)\hat{\beta}_{2,X_i,\psi}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2.} \end{cases}$$

where $\hat{\beta}_{1X_i,\psi}(\lambda_k)$ and $\hat{\beta}_{2X_i,\psi}(\lambda_k)$, are defined by (16) and $X_t$ is replaced by $X_{it}$, $i = 1, 2$.

Therefore, the $M$-periodogram matrix is define by

$$I^X_{M,T}(\lambda_k) = [I^X_{M,T,ij}(\lambda_k)]^2_{i,j=1} = \begin{bmatrix} I^X_{M,T,11}(\lambda_k) & I^X_{M,T,12}(\lambda_k) \\ I^X_{M,T,21}(\lambda_k) & I^X_{M,T,22}(\lambda_k) \end{bmatrix}. \tag{19}$$

Let $\Gamma^X_T$ be the covariance matrix of the first $T$, observations from $X_t$ with absolutely summable autocovariance function and let $f^X(.)$ be its spectral density matrix. Let $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$, and $D_T$ be an $2T \times 2T$ matrix,

$$D_T = [D_{T,ij}]^2_{i,j=1} = \begin{bmatrix} D_{T,11} & D_{T,12} \\ D_{T,21} & D_{T,22} \end{bmatrix}, \tag{20}$$

where

$$D_{T,ij} = \text{diag}[I^X_{M,T,ij}(\lambda_0), I^X_{M,T,ij}(\lambda_1), \ldots, I^X_{M,T,ij}(\lambda_{(T-1)})]. \tag{21}$$

Define a transformation matrix $H_T$ by

$$H_T = \begin{bmatrix} G_T & 0 \\ 0 & G_T \end{bmatrix}, \tag{22}$$

where $G_T$ is an $T \times T$ matrix with rows given by

$$g_{T,k} = T^{-1/2}[1, e^{-\mathrm{i}\pi k/T}, e^{-\mathrm{i}\pi 2k/T}, \ldots, e^{-\mathrm{i}\pi(T-1)k/T}], k = 0 \ldots, T-1. \tag{23}$$

Let $H^*_T$ be the conjugate transpose of $H_T$. Thus, we robustly estimate by $\Gamma^X_T$

$$\hat{\Gamma}^X_{M,T} = 2\pi H^*_T \hat{D}_{M,T} H_T. \tag{24}$$

The lag-$h$, $h = 0, \ldots, N-1$, robust sample cross-covariance function $\hat{\gamma}^X_{M,T,ij}(h)$ is extracted from the first row of $\hat{\Gamma}^X_{M,T,ij}$ for $i, j = 1, 2$. Finally, the lag-$h$ autocovariance and autocorrelation matrices are constructed estimating all $(i, j)$th elements for $i, j = 1, 2$. Thus, the robust autocovariance matrix function is:

$$\hat{\Gamma}^X_{M,T}(h) = \begin{bmatrix} \hat{\gamma}^X_{M,T,11}(h) & \hat{\gamma}^X_{M,T,12}(h) \\ \hat{\gamma}^X_{M,T,21}(h) & \hat{\gamma}^X_{M,T22}(h) \end{bmatrix} \tag{25}$$

It should be noted that the PCA or APCA procedure is calculated from the covariance matrix function at lag $h = 0$.

## 4 Simulation study

This section reports simulation results related to the performance of the proposed methodology for finite sample size. As in [1], the data generating process (DGP) is

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta}\epsilon_{it}, \tag{26}$$

where the factors are $T \times r$ matrices of $N(0, 1)$ random variables. The contaminated data generating process (CDGP) with additive outliers is

$$Z_{it} = X_{it} + \omega\delta_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta}\epsilon_{it} + \omega\delta_{it}. \tag{27}$$

For the simulations, $r = 1, 3$ and 5 and the maximum number of factors is 8. $N = 50, 100, 200$ and 500. $T = 50, 100, 200$, 500 and 1000. Two scenarios are considered: (i) the samples are uncontaminated ($p_i = 0, i = 1, \ldots, N$), and (ii) the samples are contaminated ($p_i \neq 0$). When $p_i \neq 0$, $\omega_1 = 15$ and $\omega_i = 0$, $i = 2, \ldots, N$, i.e, the contamination occurs only in the first series of the random vector with the probability of occurrence given in the tables. The reported empirical results are based on 1000 replications. The simulations were performed using the `R` programming language [19].

The first objective of this empirical study is to verify the performance of the three information criteria for estimating the number of factors estimated by APCA method as given in (8) under influence of additive outlier model (9). In this scenario, the estimated number of factors is expected to increase. The Averages of $\hat{r}$ are reported in Tables 1, 2 and 3, for $r = 1, 2$ and 5, respectively.

From Tables 1, 2 and 3, the effect of additive outliers in factor models appears by comparing the estimated number of factors when $p_i = 0$ with the case $p_i \neq 0$. When $p_i = 0$, the results are in accord with the ones in [1]. When $p_i \neq 0$, as expected, the increment of variability due to the presence of outliers leads to increase the number of estimated factors for all information criteria

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 1.00 | 1.00 | 4.90 | 1.21 | 1.18 | 5.38 | 1.70 | 1.66 | 6.16 |
| 50 | 100 | 1.00 | 1.00 | 1.00 | 1.30 | 1.26 | 1.38 | 1.88 | 1.85 | 1.93 |
| 50 | 200 | 1.00 | 1.00 | 1.00 | 1.40 | 1.38 | 1.45 | 1.97 | 1.96 | 1.98 |
| 50 | 500 | 1.00 | 1.00 | 1.00 | 1.49 | 1.47 | 1.52 | 2.00 | 2.00 | 2.00 |
| 50 | 1000 | 1.00 | 1.00 | 1.00 | 1.54 | 1.53 | 1.56 | 2.00 | 2.00 | 2.00 |
| 100 | 50 | 1.00 | 1.00 | 1.00 | 1.16 | 1.14 | 1.21 | 1.60 | 1.57 | 1.69 |
| 100 | 100 | 1.00 | 1.00 | 1.00 | 1.22 | 1.18 | 1.38 | 1.80 | 1.75 | 1.92 |
| 100 | 200 | 1.00 | 1.00 | 1.00 | 1.30 | 1.26 | 1.40 | 1.95 | 1.93 | 1.99 |
| 100 | 500 | 1.00 | 1.00 | 1.00 | 1.35 | 1.34 | 1.44 | 2.00 | 1.99 | 2.00 |
| 100 | 1000 | 1.00 | 1.00 | 1.00 | 1.41 | 1.40 | 1.46 | 2.00 | 2.00 | 2.00 |
| 200 | 50 | 1.00 | 1.00 | 1.00 | 1.10 | 1.10 | 1.13 | 1.46 | 1.44 | 1.51 |
| 200 | 100 | 1.00 | 1.00 | 1.00 | 1.16 | 1.13 | 1.24 | 1.68 | 1.64 | 1.81 |
| 200 | 200 | 1.00 | 1.00 | 1.00 | 1.19 | 1.15 | 1.44 | 1.84 | 1.79 | 1.97 |
| 200 | 500 | 1.00 | 1.00 | 1.00 | 1.28 | 1.25 | 1.40 | 1.99 | 1.98 | 2.00 |
| 200 | 1000 | 1.00 | 1.00 | 1.00 | 1.29 | 1.27 | 1.38 | 2.00 | 2.00 | 2.00 |
| 500 | 50 | 1.00 | 1.00 | 1.00 | 1.03 | 1.03 | 1.03 | 1.21 | 1.19 | 1.24 |
| 500 | 100 | 1.00 | 1.00 | 1.00 | 1.05 | 1.05 | 1.06 | 1.35 | 1.32 | 1.45 |
| 500 | 200 | 1.00 | 1.00 | 1.00 | 1.08 | 1.07 | 1.15 | 1.58 | 1.53 | 1.76 |
| 500 | 500 | 1.00 | 1.00 | 1.00 | 1.11 | 1.08 | 1.39 | 1.88 | 1.81 | 2.00 |
| 500 | 1000 | 1.00 | 1.00 | 1.00 | 1.14 | 1.12 | 1.30 | 1.98 | 1.97 | 2.00 |

**Table 1.** Averages of $\hat{r}$ for $p_i = 0, 0.01$ and $0.05$ when $r = 1$.

for the percentage of contamination of 1% and 5%. In general, it is noted that IC2 is less affected than the others.

The second objective is to verify and to compare the performance of the estimated number of factors using the information criteria when the standard APCA method is replaced by the robust methodology suggested in section 3. Let $\hat{r}^M$ denote the estimated number of factors considering the robust methodology. The primary interest here is to find out if the robust proposed methodology is competitive in the absence of contamination and if it still provides reliable results in a scenario where the data is contaminated. The results are reported in Tables 4, 5 and 6, for $r = 1, 2$ and 5, respectively.

From Tables 4, 5 and 6, when $p_i = 0$ it is noted that the reported values are in accord with the ones from Tables 1, 2 and 3. This indicates that the proposed robust method may still be considered in a scenario where the occurrence of outliers is uncertain. On the other hand, when there are outliers, i.e, $p_i \neq 0$ in the tables, the results are also close to ones when $p_i = 0$ of Tables 1, 2 and 3. Thus, in a scenario where there are outliers presented in the data, the robust methodology still provides useful results.

Others simulations with different degrees of contamination and data generating process present similar conclusions and are available upon request. The results presented in this section motive the application of the proposed methodology to a real data problem.

## 5 Conclusions

In this paper, a robust FA method for high-dimensional with additive outliers is proposed. The simulations show that additive outliers increase the number of factors estimated by the standard information criteria. The information criteria applied to the robustified estimation method presents better performance and is an alternative method when there is any evidence of atypical observations in the multivariate time series data.

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 3.00 | 3.00 | 7.24 | 3.25 | 3.23 | 7.39 | 3.79 | 3.75 | 7.67 |
| 50 | 100 | 3.00 | 3.00 | 3.00 | 3.40 | 3.38 | 3.46 | 3.93 | 3.92 | 3.96 |
| 50 | 200 | 3.00 | 3.00 | 3.00 | 3.58 | 3.56 | 3.61 | 3.99 | 3.99 | 3.99 |
| 50 | 500 | 3.00 | 3.00 | 3.00 | 3.73 | 3.73 | 3.76 | 4.00 | 4.00 | 4.00 |
| 50 | 1000 | 3.00 | 3.00 | 3.00 | 3.85 | 3.84 | 3.86 | 4.00 | 4.00 | 4.00 |
| 100 | 50 | 3.00 | 3.00 | 3.00 | 3.22 | 3.19 | 3.25 | 3.74 | 3.71 | 3.80 |
| 100 | 100 | 3.00 | 3.00 | 3.00 | 3.35 | 3.32 | 3.47 | 3.90 | 3.88 | 3.97 |
| 100 | 200 | 3.00 | 3.00 | 3.00 | 3.46 | 3.44 | 3.54 | 3.98 | 3.97 | 4.00 |
| 100 | 500 | 3.00 | 3.00 | 3.00 | 3.64 | 3.62 | 3.69 | 4.00 | 4.00 | 4.00 |
| 100 | 1000 | 3.00 | 3.00 | 3.00 | 3.73 | 3.72 | 3.77 | 4.00 | 4.00 | 4.00 |
| 200 | 50 | 3.00 | 3.00 | 3.00 | 3.15 | 3.15 | 3.17 | 3.64 | 3.61 | 3.68 |
| 200 | 100 | 3.00 | 3.00 | 3.00 | 3.23 | 3.21 | 3.31 | 3.83 | 3.80 | 3.89 |
| 200 | 200 | 3.00 | 3.00 | 3.00 | 3.37 | 3.31 | 3.57 | 3.95 | 3.93 | 3.99 |
| 200 | 500 | 3.00 | 3.00 | 3.00 | 3.54 | 3.50 | 3.68 | 4.00 | 4.00 | 4.00 |
| 200 | 1000 | 3.00 | 3.00 | 3.00 | 3.61 | 3.59 | 3.70 | 4.00 | 4.00 | 4.00 |
| 500 | 50 | 3.00 | 3.00 | 3.00 | 3.10 | 3.09 | 3.10 | 3.43 | 3.42 | 3.45 |
| 500 | 100 | 3.00 | 3.00 | 3.00 | 3.12 | 3.12 | 3.17 | 3.65 | 3.63 | 3.71 |
| 500 | 200 | 3.00 | 3.00 | 3.00 | 3.20 | 3.18 | 3.30 | 3.83 | 3.82 | 3.92 |
| 500 | 500 | 3.00 | 3.00 | 3.00 | 3.34 | 3.28 | 3.66 | 3.98 | 3.98 | 4.00 |
| 500 | 1000 | 3.00 | 3.00 | 3.00 | 3.48 | 3.45 | 3.68 | 4.00 | 4.00 | 4.00 |

**Table 2.** Averages of $\hat{r}$ for $p_i = 0, 0.01$ and $0.05$ when $r = 3$.

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 5.00 | 5.00 | 7.93 | 5.27 | 5.25 | 7.95 | 5.84 | 5.82 | 7.98 |
| 50 | 100 | 5.00 | 5.00 | 5.00 | 5.43 | 5.42 | 5.49 | 5.96 | 5.95 | 5.98 |
| 50 | 200 | 5.00 | 5.00 | 5.00 | 5.63 | 5.62 | 5.66 | 6.00 | 6.00 | 6.00 |
| 50 | 500 | 5.00 | 5.00 | 5.00 | 5.84 | 5.83 | 5.86 | 6.00 | 6.00 | 6.00 |
| 50 | 1000 | 5.00 | 5.00 | 5.00 | 5.93 | 5.93 | 5.94 | 6.00 | 6.00 | 6.00 |
| 100 | 50 | 5.00 | 5.00 | 5.00 | 5.25 | 5.24 | 5.28 | 5.79 | 5.77 | 5.83 |
| 100 | 100 | 5.00 | 5.00 | 5.00 | 5.38 | 5.35 | 5.50 | 5.95 | 5.92 | 5.99 |
| 100 | 200 | 5.00 | 5.00 | 5.00 | 5.55 | 5.53 | 5.63 | 6.00 | 5.99 | 6.00 |
| 100 | 500 | 5.00 | 5.00 | 5.00 | 5.76 | 5.76 | 5.80 | 6.00 | 6.00 | 6.00 |
| 100 | 1000 | 5.00 | 5.00 | 5.00 | 5.87 | 5.86 | 5.90 | 6.00 | 6.00 | 6.00 |
| 200 | 50 | 5.00 | 5.00 | 5.00 | 5.21 | 5.20 | 5.23 | 5.67 | 5.66 | 5.71 |
| 200 | 100 | 5.00 | 5.00 | 5.00 | 5.30 | 5.29 | 5.38 | 5.89 | 5.88 | 5.94 |
| 200 | 200 | 5.00 | 5.00 | 5.00 | 5.47 | 5.43 | 5.63 | 5.98 | 5.97 | 6.00 |
| 200 | 500 | 5.00 | 5.00 | 5.00 | 5.69 | 5.67 | 5.79 | 6.00 | 6.00 | 6.00 |
| 200 | 1000 | 5.00 | 5.00 | 5.00 | 5.81 | 5.80 | 5.87 | 6.00 | 6.00 | 6.00 |
| 500 | 50 | 5.00 | 5.00 | 5.00 | 5.13 | 5.13 | 5.14 | 5.56 | 5.56 | 5.59 |
| 500 | 100 | 5.00 | 5.00 | 5.00 | 5.20 | 5.19 | 5.23 | 5.79 | 5.77 | 5.82 |
| 500 | 200 | 5.00 | 5.00 | 5.00 | 5.32 | 5.30 | 5.41 | 5.93 | 5.92 | 5.97 |
| 500 | 500 | 5.00 | 5.00 | 5.00 | 5.51 | 5.46 | 5.76 | 6.00 | 6.00 | 6.00 |
| 500 | 1000 | 5.00 | 5.00 | 5.00 | 5.69 | 5.66 | 5.84 | 6.00 | 6.00 | 6.00 |

**Table 3.** Averages of $\hat{r}$ for $p_i = 0, 0.01$ and $0.05$ when $r = 5$.

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 1.00 | 1.00 | 1.02 | 1.01 | 1.01 | 1.04 | 1.07 | 1.05 | 1.27 |
| 50 | 100 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.06 | 1.04 | 1.15 |
| 50 | 200 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.05 | 1.04 | 1.10 |
| 50 | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.08 | 1.07 | 1.09 |
| 50 | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.18 | 1.18 | 1.19 |
| 100 | 50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.02 | 1.01 | 1.03 |
| 100 | 100 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.02 | 1.01 | 1.08 |
| 100 | 200 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 1.05 |
| 100 | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 1.02 |
| 100 | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 1.02 |
| 200 | 50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 |
| 200 | 100 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 200 | 200 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 |
| 200 | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 200 | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 100 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 200 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 500 | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table 4.** Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and $0.05$ when $r = 1$.

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 3.00 | 3.00 | 3.02 | 3.01 | 3.01 | 3.04 | 3.07 | 3.05 | 3.07 |
| 50 | 100 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.06 | 3.04 | 3.05 |
| 50 | 200 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.05 | 3.04 | 3.07 |
| 50 | 500 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.08 | 3.07 | 3.09 |
| 50 | 1000 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.08 | 3.07 | 3.09 |
| 100 | 50 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.02 | 3.01 | 3.03 |
| 100 | 100 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.02 | 3.01 | 3.08 |
| 100 | 200 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.01 | 3.05 |
| 100 | 500 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.01 | 3.02 |
| 100 | 1000 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.01 | 3.02 |
| 200 | 50 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 |
| 200 | 100 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.00 | 3.00 |
| 200 | 200 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.01 | 3.01 |
| 200 | 500 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 200 | 1000 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 500 | 50 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 500 | 100 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 500 | 200 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 500 | 500 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 500 | 1000 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |

**Table 5.** Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and $0.05$ when $r = 3$.

| T | N | $p_i = 0$ | | | $p_i = 0.01$ | | | $p_i = 0.05$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 | IC1 | IC2 | IC3 |
| 50 | 50 | 5.00 | 5.00 | 5.02 | 5.01 | 5.01 | 5.04 | 5.07 | 5.05 | 5.23 |
| 50 | 100 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 | 5.06 | 5.04 | 5.05 |
| 50 | 200 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.05 | 5.04 | 5.10 |
| 50 | 500 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.08 | 5.07 | 5.10 |
| 50 | 1000 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.18 | 5.08 | 5.89 |
| 100 | 50 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.02 | 5.01 | 5.03 |
| 100 | 100 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.02 | 5.01 | 5.08 |
| 100 | 200 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 | 5.01 | 5.05 |
| 100 | 500 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 | 5.01 | 5.02 |
| 100 | 1000 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 | 5.01 | 5.02 |
| 200 | 50 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 |
| 200 | 100 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 200 | 200 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.01 |
| 200 | 500 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 200 | 1000 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 500 | 50 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 500 | 100 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 500 | 200 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 500 | 500 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| 500 | 1000 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |

**Table 6.** Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and $0.05$ when $r = 5$.

## References


1. Bai, J., Ng, S.: Determining the number of factors in approximate factor models. Econometrica **70**(1) (2002) 191–221
2. Connor, G., Korajczyk, R.A.: Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of financial economics **15**(3) (1986) 373–394
3. Bai, J., Ng, S.: Principal components and regularized estimation of factor models. arXiv preprint arXiv:1708.08137 (2017)
4. Reisen, V.A., Sgrancio, A.M., Lévy-Leduc, C., Bondon, P., Monte, E.Z., Cotta, H.H.A., Ziegelmann, F.A.: Robust factor modelling for high-dimensional time series: An application to air pollution data. Applied Mathematics and Computation **346** (2019) 842–852
5. Stock, J.H., Watson, M.W.: Forecasting using principal components from a large number of predictors. Journal of the American statistical association **97**(460) (2002) 1167–1179
6. Geweke, J.: The dynamic factor analysis of economic time series. Latent variables in socio-economic models (1977)
7. Sargent, T.J., Sims, C.A., et al.: Business cycle modeling without pretending to have too much a priori economic theory. New methods in business cycle research **1** (1977) 145–168
8. Chamberlain, G., Rothschild, M.: Arbitrage, factor structure, and mean-variance analysis on large asset markets (1982)
9. Stock, J.H., Watson, M.W.: New indexes of coincident and leading economic indicators. NBER macroeconomics annual **4** (1989) 351–394
10. Bai, J.: Inferential theory for factor models of large dimensions. Econometrica **71**(1) (2003) 135–171
11. Baragona, R., Battaglia, F., et al.: Outliers in dynamic factor models. Electronic Journal of Statistics **1** (2007) 392–432
12. Kristensen, J.T.: Factor-based forecasting in the presence of outliers: Are factors better selected and estimated by the median than by the mean? Studies in Nonlinear Dynamics & Econometrics **18**(3) (2014) 309–338
13. Fan, J., Liao, Y., Mincheva, M.: Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4) (2013) 603–680
14. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics. Springer (1992) 492–518
15. Fan, J., Ke, Y., Liao, Y.: Robust factor models with covariates. (2016)
16. Reisen, V., Lévy-Leduc, C., Taqqu, M.: An m-estimator for the long-memory parameter. Journal of Statistical Planning and Inference **187** (2017) 44 – 55
17. Fajardo, F., Reisen, V.A., Lévy-Leduc, C., Taqqu, M.: M-periodogram for the analysis of long-range-dependent time series. Statistics **52**(3) (2018) 665–683
18. Reisen, V., Lévy-Leduc, C., Cotta, H., Bondon, P., Ispany, M.: An overview of robust spectral estimators. (2019)
19. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2019)

# Monotonicity Assumptions for Recession Forecasting

David Kelley*

Federal Reserve Bank of Chicago

**Abstract.** This paper applies Gaussian processes classification to predict recessions one year ahead. It shows that while the commonly used probit model can produce poorly calibrated estimates in this case, replacing the single index assumption of the probit model with a Gaussian processes model provides estimates that are well calibrated. Results from Gaussian process models restricted to be monotonically related to the indicator imply that breaking the monotonicity assumption and not just the linearity assumption of a single index model is required for well-calibrated conditional probability estimates.

## 1 Introduction

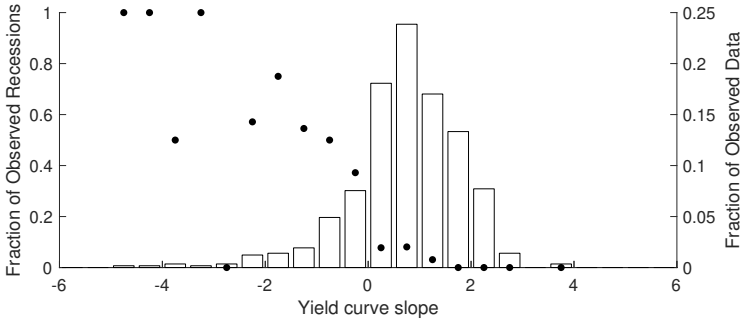
Economic forecasting inherently deals with the uncertainty of the future. For a given indicator, we are interested in the information conveyed by an observation of that variable. To generate a conditional probability, economic forecasters often turn to probit models. While more sophisticated recession prediction models have been developed, in practice probit models are commonly used (Estrella and Mishkin, 1996; Engstrom and Sharpe, 2018; Bauer and Mertens, 2018; Benzoni et al., 2018). The logit model (or logistic regression) makes a slight modification to the functional form in estimation, but both of these approaches inherently assume that estimating regression parameters inside a link function generates appropriate conditional probabilities. While estimating the regression parameters of these models can change the location and scale of the inferred probabilities, the shape of the posterior is assumed to be known.

This analysis examines this assumption and questions its validity. After introducing Gaussian process (GP) models as a more flexible alternative to the probit model, it will assess how well calibrated are the estimated probabilities of both the probit model and three versions of the Gaussian process model.

For simplicity and consistency with prior empirical applications, the focus will be on predicting recessions as defined by the NBER Business Cycle Dating Committee 12 months ahead with univariate, cross-sectional models. While this analysis excludes more sophisticated methods, it clearly documents a flaw in a

* Thanks to Luca Benzoni, Scott Brave, R. Andrew Butters, and Will Lee for helpful comments. The views expressed herein are my own and do not necessarily represent those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Email: David.Kelley@chi.frb.org.

**Fig. 1.** Observed fraction of recessions (dots) and distribution of indicator values (bars)




model commonly used in practice. Each model under consideration will be illustrated using the near-term forward spread of Engstrom and Sharpe (2018), defined as the difference between the 6-quarter forward and 1-quarter zero coupon U.S. Treasury yields (using fitted yields from Gurkaynak et al. (2007)). As a simple summary, the the observed fraction of recessions 12 months ahead according to the value of this indicators is plotted in Figure 1 along with the distribution of observations of the indicator. Notable in this figure is that while the observed faction of recessions by indicator value is generally increasing as the indicator value declines, the relationship is non-monotonic. Also, the bulk of observations of the slope of the yield curve occur between -2 and 3, but there are a few observations in the tails beyond these values.

## 2 Recession Prediction with Probit Models

For an indicator $x_t$, a generalized version of the probability of a recession 12 months ahead is

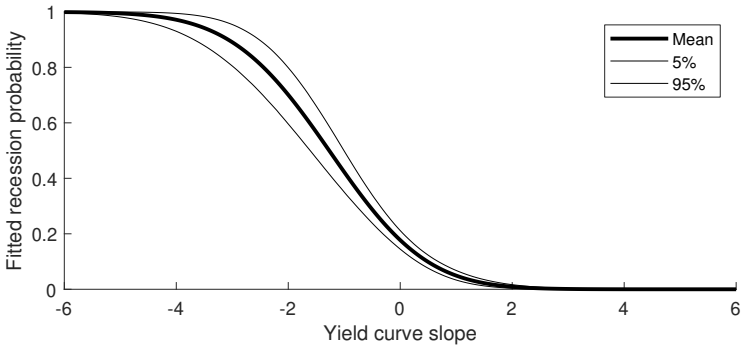
$$P(r_{t+12}|x_t) = g(f(x_t, \theta)) \tag{1}$$

where $g(\cdot) : \mathbb{R} \to [0, 1]$ is a parameter-free function and $f(\cdot, \theta) : \mathbb{R} \to \mathbb{R}$ is a transformation of the indicator to the appropriate location and scale that depends on unknown parameters $\theta$. Most commonly, $f(x_t, \theta) = \alpha + \beta x_t$ where $\theta = \{\alpha, \beta\}$, otherwise known as a single index model. Other choices for $f(\cdot)$ are possible including nonlinear and nonparametric estimates.

The probit model is a single index model where $g(\cdot)$ is the cumulative standard normal function.[1] It has an underlying latent variable representation (Gel-

[1] Throughout, $g(\cdot)$ is assumed to be the cumulative normal distribution so that the single-index model is a probit model. Results throughout are broadly consistent if the function $g(\cdot)$ is the logistic transformation that implies the logit model in a single index context.

**Fig. 2.** Fitted probabilities of probit model




man et al., 2013, p. 408) that assumes normally distributed errors such that

$$u_t \sim \mathcal{N}(a + bx_t, 1) \tag{2}$$

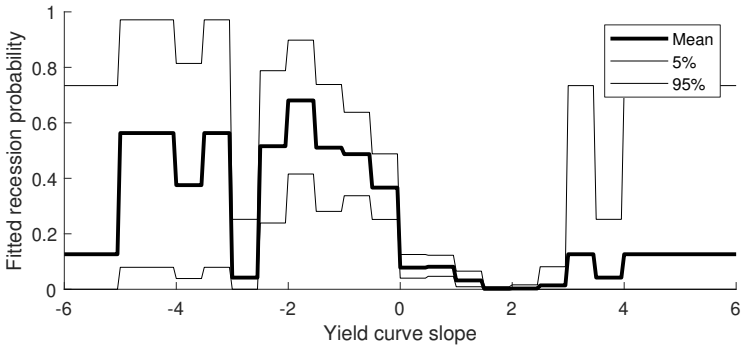
$$r_t = \begin{cases} 1 \text{ if } u_t > 0 \\ 0 \text{ if } u_t < 0 \end{cases} \tag{3}$$

This parameterization allows for efficient inference on the underlying parameters via a Gibbs sampler. The familiar fitted probability of a recession conditional on an observed indicator value is plotted in Figure 2, presented for comparison to later models. Note that the fitted probability approaches one somewhat quickly for larger negative values of the slope of the yield curve.

## 3 A Binomial Model of Recession Prediction

The simplest departure from a parametric estimate of recession probabilities is to consider $P(r_{t+12}|x_t)$ in the context of a binomial sampling model. To do so, bin the realizations of an indicator with a bin width 0.5. By specifying a prior probability of recession in any period of a Beta(0.13, 0.87) distribution, the prior matches the observed frequency of recessions with a "prior observation" weight of one. The posterior in this case is known to follow a Beta($0.13 + \sum_t r_t$, $0.87 + T - \sum_t r_t$) distribution. While suboptimal due to the discrete binning procedure, the fitted posterior probability conditional on the indicator value in Figure 3 reveals three things. First, the probability of recession conditional on the near-term forward spread is non-monotonically decreasing in the value of the indicator. As the indicator increases from its lowest values, the probability of recession decreases to a local low near -3, rises again to a local peak near -2, falls again to a local low near +2, then rises as it is more influenced by the prior for bins with relatively few observations at higher values. This implies that a monotonically decreasing recession probability requires some prior information or a different likelihood function. Second, as the indicator moves further from the center of it's distribution, the uncertainty around the recession probability

**Fig. 3.** Fitted probabilities of beta-binomial model




increases substantially. In the area slightly greater than zero where most indicator values are concentrated there is relative certainty but considerably more uncertainty away from this region. Third, the estimated probability is never higher than 70%. Despite low values of the slope of the yield curve always being associated with a recession one year ahead, there are few observations at these values leading to only moderately high probabilities of recession.

## 4 Gaussian Process Models

While modeling choices in Equation 1 have generally focused around $g(\cdot)$, the benefit of explicitly parameterizing $f(\cdot, \theta)$ allows for consideration of models that break the assumption of a single index model and introduce some of the flexibility of the binomial sampling model. The Gaussian process (GP) model is one possible replacement that does so. For a comprehensive introduction to Gaussian process models, see Rasmussen and Williams (2006).

A Gaussian process model is one in which the full sample of a dependent variable $Y$ is jointly normally distributed according to a mean and variance that are specified as a function of independent variables in $X$. Given the observed data, the mean of Y will be a function $m(X)$ and variance of Y will be a function $K(X, X')$. We write the output as distributed according to the Gaussian process,

$$Y \sim \mathcal{GP}(m(X), K(X, X')) \tag{4}$$

Given that the full dependent variable is jointly normally distributed for observed data $(X, Y)$ and given a new observation of the independent variable $x^*$, the distribution of the full data is also jointly normal,

$$\begin{bmatrix} Y \\ y^* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(x^*) \end{bmatrix}, \begin{bmatrix} K(X, X') & K(X, x^*) \\ K(x^*, X') & K(x^*, x^*) \end{bmatrix} \right) \tag{5}$$

Conditional normality gives the distribution of the unobserved dependent variable $y^*$:

$$\mathbb{E}(y^*|X, Y, x^*) = m(x^*) + K(X, x^*)K(X, X')^{-1}(Y - m(X)) \tag{6}$$

$$\text{Var}(y^*|X, Y, x^*) = K(x^*, x^*) - K(X, x^*)K(X, X')^{-1}K(x^*, X') \tag{7}$$

For simplicity, consider a classification problem similar to the probit model where $g(\cdot)$ is the cumulative standard normal function implying

$$P(r_{t+12}|x_t) = \Phi(\mathcal{GP}[m(x_t), K(x_t, x_t)]) \tag{8}$$

Since we model the observed recession classifications $r_{t+12}$, direct estimation of the conditional quantities in Equations 6 and 7 is no longer feasible. We therefore require an approximation of the posterior, using either the Laplace approximation or expectation propagation using the package developed in Vanhatalo et al. (2013).

The modeling specifications for Gaussian process models come in the choice of functional forms for $m(\cdot)$ and $K(\cdot, \cdot)$. The mean function $m(\cdot)$ is usually assumed to be zero but we will also consider the case where it encodes prior information in a linear form with parameters fixed by introspection,

$$m(x_t) = a + bx_t \tag{9}$$

Attention generally focuses on the covariance function $K(\cdot, \cdot)$. The most commonly used form, and the one used here, is the squared exponential,

$$K(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right) \tag{10}$$

which implies that data near each other have high covariance while those further away have lower covariance. The parameter $l$ determines how data with different input values affect the final Gaussian process by specifying the relative influence of points that are more distant. The parameter $\sigma$ gives the covariance of two outputs with the same observed input $x$. The parameters in the covariance function ($l$ and $\sigma$) are unobserved and must be fitted to the data, done so by maximizing the marginal posterior. For all models, the prior on $l$ is specified as a mean-zero $t$-distribution with 4 degrees of freedom and a standard deviation of 1, and the prior for $\sigma$ is uninformative.

### 4.1 Gaussian Processes with Constant Mean Priors

The simplest model is one where the GP mean function is zero, but this model produces an unconditional probability of recession equal to 50%. Since recessions are rare events, the prior is instead specified so that the unconditional probability matches the observed fraction of recessions of roughly 13%. The fitted probability in Figure 4 show the appeal of using Gaussian processes in this application. While the fitted probabilities generally follow the contour of the beta-binomial

**Fig. 4.** Fitted probabilities of flat prior Gaussian process model

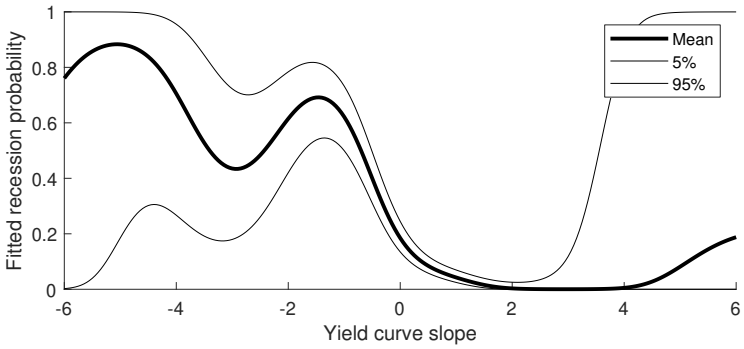
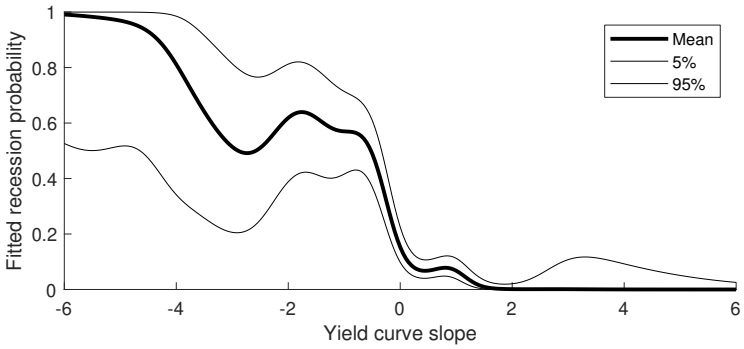



**Fig. 5.** Fitted probabilities of linear prior Gaussian process model

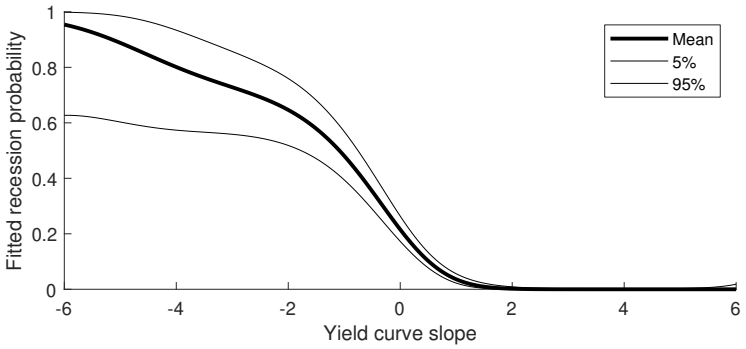



model in Figure 3, they now vary smoothly over the range of observed values. Additionally, the estimated uncertainty appears reasonable as the 90% credible interval is small where many indicator values have been observed and much wider where the data are relatively sparse.

### 4.2 Gaussian Processes with Linear Mean Priors

The economic intuition that the yield curve embodies financial market participants' expectations of future short rates motivates the use of a mildly informative prior. The slope parameter is set to -0.1 to provide only a weak prior connection between the indicator and the probability of recession. The intercept parameter of $m(\cdot)$ is then set such that that the mean of the indicator produces a prior probability of recession equal to the unconditional observed fraction of recessions. The results in Figure 5 generally resemble the results of the model with a constant mean prior with three differences: (1) changes in the fitted probability are somewhat more abrupt as the fitted length scale parameter $l$ is substantially smaller than for the constant mean prior model, (2) the mean probability is closer to one and zero for extreme values of the indicator as a result of the prior,

**Fig. 6.** Fitted probabilities of monotonic Gaussian process model




and (3) the uncertainty has been slightly reduced, particularly for extreme values of the indicator. While the prior has reduced the posterior variance slightly, the 90% credible intervals are still suitably wide enough to allow additional data to influence the estimates substantially.

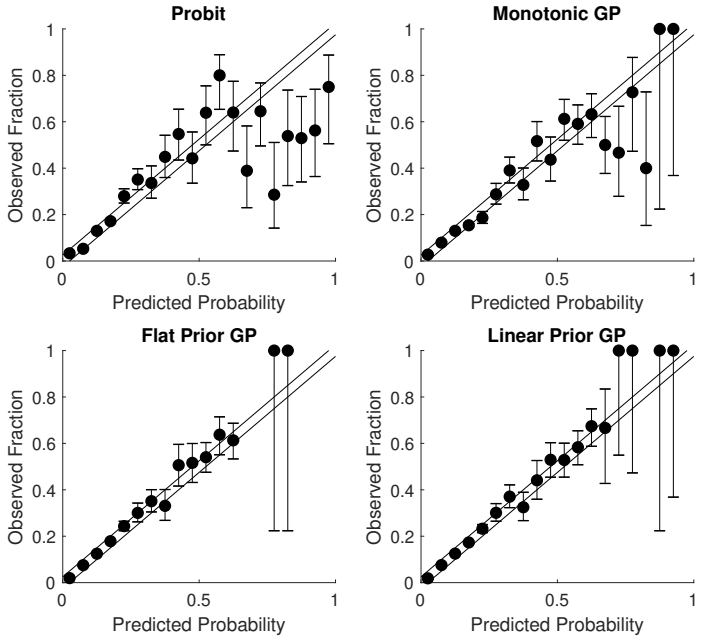
### 4.3 Monotonic Gaussian Processes

Finally, to facilitate comparison to the probit model consider an additional specification in which the the output from the GP is restricted to be a monotonic function of $x_t$ according to Riihimki and Vehtari (2010). In this case, the mean function is taken as $m(x_t) = 0$ but the model restricts all outputs such that $x_1 < x_2 \Rightarrow f(x_1) > f(x_2)$. To do so, the standard GP model is first fit similarly to the other GP models. This model then iteratively augments the likelihood by adding observations that enforce a non-negative derivative at the point that has the largest negative derivative until the output is monotonic. The results in Figure 6 are substantially different than the other two Gaussian process results. The mean response is somewhat similar to the probit model but with a lower estimated probability for more negative values. Even with an uninformative prior, the addition of the monotonicity assumption has substantially reduced the uncertainty for extreme values of the indicator as non-local data are effectively allowed more influence on the estimated probabilities.

## 5 Model Calibration

Given the approach of directly modeling a posterior probability, of substantial interest for these models is how well calibrated they are – i.e., when a model gives a 40% chance of a recession, does a recession occur 40% of the time? Since recessions are relatively rare events, most models make few, if any, predictions with high probability. To reduce the uncertainty in this exercise, each model is fitted to each series in FRED-MD (McCracken and Ng, 2015), a monthly panel of

**Fig. 7.** Calibration plots with 5% bins




common U.S. macroeconomic time series.[2] Each model is estimated separately for each series, and all indicator variables have been sign-normalized to have a negative correlation with the recession classification so that as the indicator decreases the probability of recession is increasing.

To construct the calibration plots, the fitted probabilities for each model across all series are organized into bins and the fraction of months in which a recession occurred is plotted as a dot. For a well calibrated model, this value will lie within the boundaries of the bin, shown as the two offset 45° lines. To test this, the 90% credible intervals is computed for each bin as the posterior of a binomial sampling model with a Beta(0.5, 0.5) prior. For example, for a 5% wide bin centered at 50%, a credible interval that falls entirely above 52.5% or entirely below 47.5% is good evidence that the model is poorly calibrated.

Figure 7 presents calibration plots for 5% bins of the probit, monotonic GP, flat prior GP, and linear prior GP models. Unsurprisingly, the unrestricted Gaussian process models appear well-calibrated as none of the 90% credible intervals exclude the appropriate bins. The most notable difference between these two models is that the model with an informative prior has more points of high esti-

[2] For simplicity, the only series used are those with no missing values prior to the last month in which all series were observed. The resulting panel contains 93 series out of the 128 series in the full FRED-MD panel. All series are transformed according to the recommended specification before being demeaned and standardized.

mated probability due to the additional information of the prior but it remains well calibrated. On the other hand, the probit model appears to deviate from the 45° line at higher fitted probabilities. Given the manner in which this model is used as an early warning indicator of coming recessions, these are exactly the regions where poor calibration should be concerning. This would indicate that the times when macroeconomic policymakers are likely to be more concerned about a recession are the same times when the probit model provides unreliable estimates. The calibration of the monotonic GP model appears somewhat better but still has three bins in which the 90% credible interval excludes the associated bin edges. Compared to the unrestricted models, this is evidence of notably worse calibration.

## 6 Conclusion

The suitability of the probit model to recession classification appears to crucially rely on whether the proportion of classified events is monotonically related to the indicator under consideration. For estimated recession probabilities one year forward, probit and monotonic GP models fitted to the FRED-MD data produce poorly calibrated results compared to the less restrictive GP models, implying that the monotonicity restriction itself is at least partially responsible for the poor calibration. While on its face the single index assumption of the probit model appears to be overly restrictive, these results suggest otherwise. They suggest that for those unwilling to undertake the complexity of GP models, more traditional nonlinear models may be capable of overcoming these calibration issues as well.

Admittedly, a monotonicity assumption is attractive for interpretive reasons, allowing for the simplification that movements in a given direction of an indicator have a common interpretation. However, this assumption may not always be appropriate. GP models with modestly informative priors produce a posterior predictive probability that incorporates this intuition while producing potentially more realistic conditional probabilities. When the data is uninformative, they reflect the notion that the econometrician knows something about how the indicator is correlated with economic activity. But when the data are informative, the estimated probabilities are allowed to match the data more closely, producing fitted probabilities with superior calibration.

The probit model still performs a role in mapping an indicator from an unbounded range to a $[0, 1]$ range, but many other methods are also capable of doing so. For those interested in probability estimates, calibration tests are a useful metric about the performance of a predictive model. For cases where the fitted probability of an event is meaningfully interpreted – such as recession prediction – these results favor GP models over probit models.

# Bibliography


Michael Bauer and Thomas Mertens. Economic forecasts with the yield curve. *Economic Letters*, 2018(07), 2018. URL https://www.frbsf.org/economic-research/publications/economic-letter/2018/march/economic-forecasts-with-yield-curve/.

Luca Benzoni, Olena Chyruk, and David Kelley. Why does the yield-curve slope predict recessions? *Chicago Fed Letter*, 2018(404), 2018. doi: 10.21033/cfl-2018-404.

Eric Engstrom and Steve Sharpe. The near-term forward yield spread as a leading indicator: A less distorted mirror. *Finance and Economics Discussion Series*, 2018(055), 2018. doi: 10.17016/feds.2018.055.

Arturo Estrella and Frederic S. Mishkin. The yield curve as a predictor of u.s. recessions. *Current Issues in Economics and Finance*, 2(7), 1996.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. 3rd edition, 2013.

Refet Gurkaynak, Brian Sack, and Jonathan Wright. The u.s. treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, 54(8):2291–2304, 2007.

Michael W. McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. *Federal Reserve Bank of St. Louis, Working Papers*, 2015(012), 2015. doi: 10.20955/wp.2015.012. URL https://doi.org/10.20955/wp.2015.012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology, 2006. ISBN 026218253X. URL http://www.gaussianprocess.org/gpml.

Jaakko Riihimki and Aki Vehtari. Gaussian processes with monotonicity information. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 645–652. PMLR, 2010. URL http://proceedings.mlr.press/v9/riihimaki10a.html.

Jarno Vanhatalo, Jaakko Riihimki, Jouni Hartikainen, Pasi Jylnki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14(Apr):1175–1179, 2013.

# The Tsallis Statistics Faces Social Problems in Developing Countries

Huber Nieto-Chaupis

Universidad Autonoma del Perú
Programa de Ingenierí a de Sistemas
hubernietochaupis@gmail.com

**Abstract.** We apply the well-known Tsallis statistics to model criminality and vehicle chaos in Lima city. Despite of the fact tha Perú has shown interesting indicators of a reduced generalized poverty, the apparition of high rates of criminality and high vehicle traffic is considered as a logic consequence due to the economical progress of the country. In this paper we argument that the geographical apparition of these social issues has as origin the phase transitions as seen in the Tsallis entropy, from a low to middle social-economical layer, fact that is not contemplated in the regional and urbanistic evolution of large cities such as Lima with a population of around 10M of habitants.

## 1 Introduction

The development of a country is seen in terms of economic and social progress. This progress becomes clear in those ones whose economic growth has surpassed the expected ones such as the predicted by the World Bank [1]. Although the economic growth is crucial for a homogeneous wellness of the societies, there are additional factors that are inside the social and education territories that to some extent are important indicators to anticipate a full growth either in the middle or long term. In this paper, we focus on the apparition of social anomalies in Lima city as consequence of the overpopulation and people behavior. Concretely, we focus on the cases of (i) Criminality (ii) Vehicle Traffic, that are social problems of global interest since the conjunction of both occurrences as undeniable part of a reality. As it is well-known, points (i) and (ii) share in common the geographic aspect so the knowledge in advance of these events might be an advantage to reconfigure policies to guarantee the security of people in particular of those that are in the layers of vulnerability as is seen in largest Latin American cities as Lima city in Peru, for example. As done in [2], the presence of events containing this duality in Lima city was studied with entropy of Shannon by which geographical areas in the urban areas were systematically identified through the usage of sophisticated algorithms that use random comparisons in the sense that the coincidence of both (i) and (ii) is done by following stochastic rules. The usage of physics-inspired formulations such as the Shannon entropy, has as justification the fact that the wellness of a society is dictated by an equilibrium that is perceived as a fully subjective fact.

Certainly, the apparition of these both issues (i) and (ii) might degrade substantially the quality of life of people, in particular of elderly people that are residing in the social layers where poverty is a common factor. In this paper we focus on the application of the well-known Tsallis entropy that emphasizes the concept of non-additivity in the sense that the events of different nature might not be seen as part of a sum, but instead it includes a pseudo-event that characterizes the conjunction of a couple of events, for instance.

In fact, the main motivation for the usage of the Tsallis entropy is that the Tsallis concept [3][4] adjust well to the case of the conjunction of criminality and vehicle traffic when both situations are seen as product of a single reality but are engaged to a subjective view by which we can denominate the product of the interaction or concurrence of both events. In this manner the aim of this paper concentrates in - Identify geographical areas in Lima city with a high risk of criminality and traffic. - Adjudicate probabilities to the identified events with the Tsallis entropy. - Identify in a quantitative manner the correlation of both events through the definition of a local equilibrium function that is translated as the extra term that breaks the additivity of statistical systems.

Although there is not an absolute mechanism that predicts the concurrence of events that violates the social equilibrium, the usage of the Tsallis entropy might be beneficial to the extent that one can associate not only street criminality and traffic but also additional events that are part of the genuine origin of these situations: corruption. In fact, the presence of corruption as a subjective force that distorts the multiple functionalities of the crucial social and economic variables in modern societies, has serious implications in the continuous degradation of the society, particularly in Latin American where most of the social and political changes have had its origin in the diverse dynamics of corruption.

Therefore, in the language of the Tsallis entropy, societies might then exhibit a kind of additive or also non-additive statistics [5][6] by which the conjunction of negative events against their quietness would give as result the apparition of subsequent episodes that characterizes the social deformation of large cities. In second section we briefly review the concepts of the Tsallis entropy [7][8]. In third section we propose phenomenological models to be applied to the case of Lima city. In fourth section we present the results. Last section is devoted to the conclusion of this paper.

## 2 Brief Review to the TSallis Entropy

Consider a set events whose individual probability is defined by being equals to $p_k$ with the condition that $\sum_K p_K = 1$. Then the Tsallis entropy [8] is written as follows,

$$S(p_k) = \frac{1}{q-1}\left(1 - \sum_k p_k^q\right) \tag{1}$$

where q is known the entropic-index. For the continuous case

$$S_q[p(x)] = \frac{1}{q-1}\left(1 - \int p^q(x)dx\right) \tag{2}$$

where the one expects that

$$0 < \left|\int p^q(x)dx\right|^2 < 1 \tag{3}$$

On the other hand, as mentioned previously the non-additivity features the Tsallis entropy. For instance, consider two independent systems A and B, then the corresponding probability density is written as

$$p(A,B) = p(A)p(B) \tag{4}$$

For this case the Tsallis entropy is defined

$$S_q(A,B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B) \tag{5}$$

clearly the extra term that breaks the additivity is of particular importance for us. The parameter (1-q) that measures the breaking of the additivity when q=1 the system returns to a full scenario of additivity [9], however in the context of Tsallis entropy it is commonly referred as a pseudo-additivity [10].

In virtue of (4) we can write down the Tsallis entropy for a specific scenario of two concurrent events such as criminality and vehicle traffic. In this way

$$s[f_q(x), g_q(x)] = \frac{1}{q-1}\left(1 - \int f^q(x)dx\right) + \frac{1}{p-1}\left(1 - \int g^p(x)dx\right) \tag{6}$$

The extra term that gives account of the non-additivity of the entropy [11] can be written as

$$N[p,q,x] = \frac{\mathcal{H}(q)}{pq-q-p+1}\left(1 - \int f^q(x)dx\right)\left(1 - \int g^p(x)dx\right) \tag{7}$$

That is actually the product or a pure join in probability [11][12]. As seen above $H(p,q) = 1$ recover the additivity of the system statistics. We left that $H(p,q)$ is to be arbitrary in the sense that is a function of integers number to be employed whatever the case the Tsallis entropy is used. Thus $H(p,q)$ is an input function that would describe the system.

## 3 Applications and Phenomenological Models

Fundamental equations of the Tsallis entropy (5) and (6) are adjusted in order to be applied systematically to the phenomenon of the concurrence of street criminality and vehicle chaos in Lima city.

**MODEL - I**

For instance, we write down the entropy for the system $f_q(x)$ as a difference,

$$S[f_q(x)] = \frac{1}{q-1} - \frac{1}{q-1}\int f^q(x)dx \tag{8}$$

Now we can modify the integration to test the simplest scenarios of application, so that we have that

$$S[f_q(z)] = \frac{1}{1-q}\int_{\text{MIN}}^{\text{MAX}} (\text{Exp}[-(x-qz)^2])^q dx \tag{9}$$

Where for the sake of simplicity we have used the Gauss function whose center is moved as to the product $qz$ by which we have introduced an extra variable $z$ so that the integration runs over $x$ and the result is given in terms of $z$. The upper limit Max is understood as being a large number.

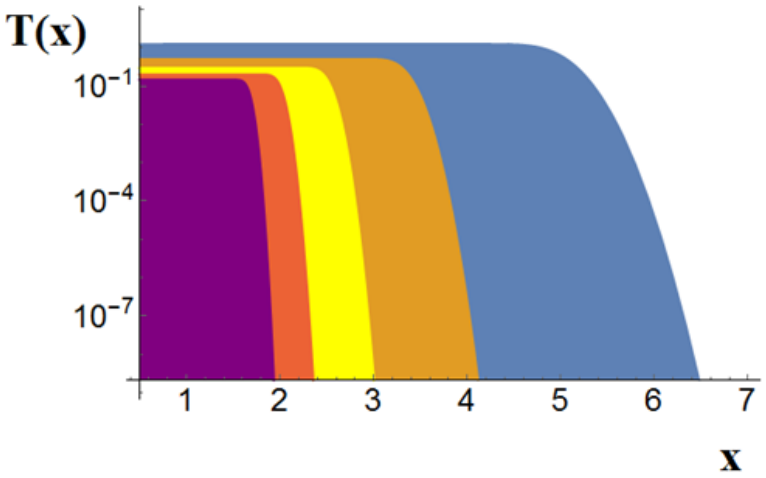



**Fig. 1.** The Tsallis entropy according to (8) for different values of the integer number q.

In Fig.1 is seen the different distributions of Tsallis where in reality all of them have the morphology of an Erf function but flipped to the right. Our next task is the projection of this model to recognize the critic zones in Lima city. The simplest model is when we will use the pseudo-additivity

**MODEL - II**

Now we extend the mathematical formulation of the Tsallis entropy as seen in section 2, to one that includes characteristics of the scenarios to be applied [13]. One of them is when the argument of the Gaussian profile is deformed by including the product of the continue variables xz, fact that changes the morphology of the resulting integrations. Thus, our proposal reads as follows

| Variable / Parameter | Tsallis | Lima City |
|---|---|---|
| x | All possible probabilities | All possible zones of risk |
| z | Distributed probabilities | Identified zones of potential risk |
| q | Level of entropy | States of Risk |
| Min | Minimal Probability | Boundary for minimal risk |
| Max | Maximal Probability | Boundary for top risk |

**Fig. 2.** Meaning of quanties in the tsallis entropy and its application to lima city

$$S[f_q(z)] - \frac{1}{q-1} = \frac{1}{1-q}\int_0^{10} (\mathrm{Exp}[-(x-qz-xz)^2])^q dx \qquad (10)$$

Being it still a crude model to be applied to a realistic social modeling. We generate additional models below.

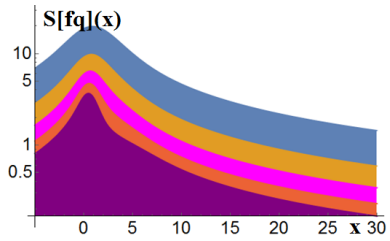



**Fig. 3.** The Tsallis entropy according to (9) for different values of the integer number q.

**MODEL - III** This is actually an extension of II with the insertion of a "width", as seen below to be 1.0001.

$$S[f_q(z)] - \frac{1}{q-1} = \frac{1}{1-q}\int_0^{10} \left(\frac{\mathrm{Exp}[-(x-qz-xz)^2]}{0.0001}\right)^q dx \qquad (11)$$

**MODEL - IV** And finally a model with a width being this small as 0.2,

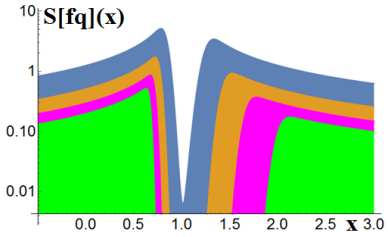



**Fig. 4.** The Tsallis entropy according to (10) for different values of the integer number q.

$$S[f_q(z)] - \frac{1}{q-1} = \frac{1}{1-q}\int_0^{10}\left(\frac{\text{Exp}[-(x-qz-xz)^2]}{0.2}\right)^q dx \qquad (12)$$

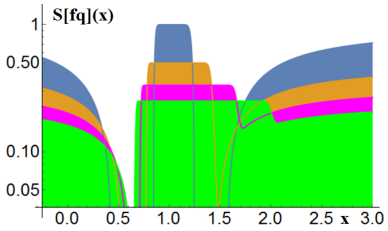



**Fig. 5.** The Tsallis entropy according to (11) for different values of the integer number q.

## 4 Results

In Fig.5 the map of Lima is shown [12]. In essence, it covers center, a substantial portion of north and east zones. In according to official reports, the perimeter of map has been limited to a 60% of normal area,

In Fig.6, is superimposed squares and circles in according to the results found after of applying the Tsallis entropy. As rule, we have used an elementary algorithm that compares a random number with the resulting Tsallis number obtained for the case that corresponds. Left-side column is a location code that is employed to identify a geographical zone inside the perimeter of Lima city. The column of middle lists the resulting probabilities of risk in according to the Tsallis entropy as done in equations (7-11). In right-side column is specified the concurrence is it is dual or single. In brackets is given the color for that recognized area. These are in according to the colors blue, yellow and white.

**Fig. 6.** A portion of map of Lima city in according to Google Earth.

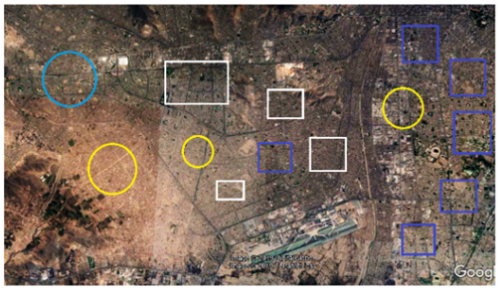


**Fig. 7.** Identification of risk areas where duality occurs in the highest probabilities in according to Table II (Fig. 8)

The numerical evaluations of the Tsallis integrations were performed through a Monte-Carlo-like method that requires of a comparison between a random number and the resulting Tsallis probability. While one expects equilibrium in terms of low risk, this can be broken by the presence of a latent risk in adjacent areas fact that degrades the level of wellness in people that belong to such areas.

The risk that an elder people gets a dual event is estimate as follows:

$$D = \frac{N}{T} \times S[f(x)] \times \frac{A}{S} \tag{13}$$

For example for the area of north edge of Lima city one gets D = 0.15 0.80 0.20 = 0.024. That is interpreted as at is 2 per 100 elder people is under a high risk. The zones of high risk in percent belongs to CRB (0.32), SJL (33.2), CAL (8.1), PP (3.2) as was officially published in [14].

## 5 Conclusion

In this paper, we have used the well-known Tsallis entropy to identify risk zones by using the central equation o Shannon that yields locations with the con-

| Location Code | Risk Probability | Concurrence |
|---|---|---|
| 3 - 9 | 76% | Dual (y) |
| 10 - 12 | 92% | Single (b) |
| 18 - 21 | 80% | Dual (w) |
| 36 – 39 | 62% | Dual (b) |
| 42 - 45 | 77% | Single (y) |
| 57 - 64 | 90 % | Dual (w) |

**Fig. 8.** Resulting tsallis probabilities to recognize areas of concurrence: criminality and traffic.

currence of traffic and street criminality. We have derived a phenomenological relation as written in Eq.(12) by which we can extract an estimate of the possible values of risk in those zones of duality. We have compared our results with the ones obtained from official reports having identified focus of violence in the order of the 90%.

## References

1. http://siteresources.worldbank.org/INTGLOMONREP2007/Resources/3413191-1179404785559/Chp1 GMR07 webPDF-corrected-may-14-2007-4.pdf.
2. S. Furuichi, On uniqueness Theorems for Tsallis entropy and Tsallis relative entropy, IEEE Transactions on Information Theory Year: 2005, Volume: 51 , Issue: 10, Pages: 3638 3645.
3. H. Suyari ; M. Tsukada, Law of error in Tsallis statistics IEEE Transactions on Information Theory Year: 2005 , Volume: 51 , Issue: 2 Pages: 753 757.
4. ] A. M. Mathai and Hans J. Haubold Mittag-Leffler functions to pathway model to Tsallis statistics Integral Transforms and Special Functions, Volume 21, 2010 - Issue 11.
5. Christian Beck, Generalised information and entropy measures in physics Contemporary Physics, Volume 50, 2009 - Issue 4.
6. Naiju M. Thomas, On the Ratios of Pathway Random Variables Communications in Statistics - Theory and Methods, Volume 43, 2014 - Issue 23.
7. Marcelo Fernandes and Breno Nri, Nonparametric Entropy-Based Tests of Independence Between Stochastic Processes Econometric Reviews, Volume 29, 2009 - Issue 3.
8. Vikas Kumar, Characterization results based on Dynamics Tsallis cumulative residual entropy Communications in Statistics - Theory and Methods, Volume 46, 2017 - Issue 17.
9. Christophe Schinckus, When Physicists Invade Economics Interdisciplinary Science Reviews, Volume 37, 2012 - Issue 2.

10. Tao Wen, Wen Jiang M, Measuring the complexity of complex network by Tsallis entropy Physica A: Statistical Mechanics and its Applications, In press, accepted manuscript, Available online 2 May 2019, Article 121054.
11. Vikas Kumar, Rekha, A quantile approach of Tsallis entropy for order statistics Physica A: Statistical Mechanics and its Applications, Volume 503, 1 August 2018, Pages 916-928.
12. Camilla Cal, Maria Longobardi, Jafar Ahmadi, Some properties of cumulative Tsallis entropy Physica A: Statistical Mechanics and its Applications, Volume 486, 15 November 2017, Pages 1012-1021.
13. Taiki Takahashi, A social discounting model based on Tsallis statistics, Physica A: Statistical Mechanics and its Applications, Volume 389, Issue 17, 1 September 2010, Pages 3600-3603.
14. https://www.pqs.pe/actualidad/noticias/ Robery in Lima city.

# Common trends in producers' expectations: implications for GDP forecasting in Uruguay.

Bibiana Lanzilotta, Lucía Rosich and Juan Gabriel Brida
Instituto de Economía (IECON), FCEA, UdelaR, GIDE-DMMC, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay

## Abstract

Data preprocessing methods: Data decomposition, seasonal adjustment, singular spectrum analysis, detrending methods, etc., Econometric models, Preferable Oral, Real macroeconomic monitoring and forecasting, Real time macroeconomic monitoring and forecasting

Abstract: This paper examines the interdependence between expectations and economic growth in Uruguay, for the last two decades (1998-2018).

To this aim, this research considers the expectation surveys collected by the "Cámara de Industrias del Uruguay" and macroeconomic series of National Account System.

The study achieved two main findings. Firstly, that there is a long-run relationship between producers' expectations and Uruguayan GDP growth controlling by the usual production factors. Secondly, by estimating multivariate structural models we found that there is a common level between the expectation indicators of the different industrial sectors grouped according to production specialization and international trade insertion. The expectation indicator of which the others depends is the one of the more tradable industries. The level component of this indicator of expectations drives the level component of the other groups.

The research shows that expectation indicators could be an accurate leader indicator of sectoral and aggregate growth. Moreover, the driver of the producers' expectations is the aggregate indicator of expectation of the more tradable industries.

# Latent precursors of delayed river ice-jam shattering: An anthropogenic factor

Alexandre Chmel[1][000-0002-5959-5331], Lyubov Banshchikova[2, 3]

[1] Ioffe Institute, 26 Polytekhnicheskaya str., 194021 St Petersburg, Russia
chmel@mail.ioffe.ru
[2] State Hydrological Institute, 199053 St Petersburg. Russia
[3] SPO "Gidrotekhproekt", 199178 St Petersburg, Russia

**Abstract** Dynamics of the ice-jam flooding was analyzed statistically as the process in the non-equilibrium system with the varying conservation. The distributions of the daily average water heights during spring breakups in the basin of Lena river were constructed. It has been found that the water heights during water rising in jam-free years were distributed in a random manner that is in accordance with an exponential (Poissonian-like) function. In contrast, the water height distributions in the periods of ice-jam flooding followed a power law in common with various multiscale hydrological phenomena. The analysis of the height distributions showed some latent perturbations in the water-ice system caused by natural causes or blasting intended to destroy jams. The efficiency aspect of the blasting actions during ice-jam flooding was considered.

**Keywords** spring backup, water height distribution, ice jam, blasting.

## 1 Introduction

The flowing water and consolidating ice (ice-jam) constitute a unified dynamic system, in which a multiplicity of local motions, ruptures, and compactizations of ice cover fragments, leads the system to its thermodynamic attractor as a non-equilibrium statistical system. The large-scale hydrological phenomena show some aspects of both organization and randomness [1]. The periodicity and persistence of various processes can be assessed from the time series of related hydrological events [2] such as the occurrence of floods [3] and return periods of rainfall [4, 5], which exhibit scale-invariant (power law) behavior. Meanwhile, to our best knowledge, there are no evidences in literature of scaling features in the process of ice-jam formation.

In this communication, we considered the statistics of a few water rises prior to ice-jams in Lena river including a catastrophic ice-jam near the town of Lensk (2001) complicated by an anthropogenic intervention through blasting. Time series were analyzed from the viewpoint of the manifestation of a self-organizing process in the non-equilibrium system with varying conservation.

Datasets from a hydrometric station (HS) situated near the town of Lensk at the distance of 2508 km from the river mouth were analyzed. The records cover periods of spring floods from 2001 to 2014.

## 3 Statistics of water level variation

### 3.1 Spring breakup without ice-jam

Figure 1 shows a size distribution of the daily average water heights recorded at the HS 3030 in the period of spring breakup in 2011, which was not followed by an ice-jam. The height distribution was calculated in the form of the dependence $N(H>h)$ versus $h$ where $N$ is the number of days when the water-level $H$ was higher than the value of $h$, which goes through the values of all registered water levels (horizontal coordinate). The same data were plotted both in semi-logarithmic and double logarithmic coordinates.

In semi-logarithmic coordinates (Fig. 1a), the data points fall upon a straight line with a slope $a$:

$$\log_{10}N(H>h) \propto -ah. \qquad (1)$$

The relation (1) is equivalent to the exponential law of Poissonian type:

$$N(H>h) \propto \exp(-ah), \qquad (1a)$$

which is indicative of random events occurring independently from each other. The same distribution plotted in the log-log coordinates (Fig. 1b) diagramed a complicated curve, which cannot be approximated by a simple function.

### 3.2 Spring breakups with ice-jams

Figure 2 shows two different presentations of the $N(H>h)$ versus $h$ distributions obtained in 2014 when a spring breakup culminated with an ice-jam. The distribution represented in the semi-logarithmic coordinates showed a

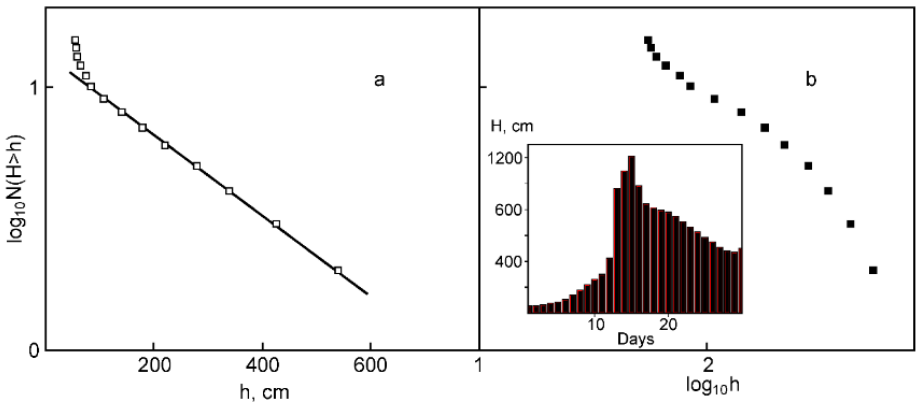



Figure 1. The water height distribution during spring breakup, which happened without an ice-jam; a – semi-logarithmic coordinates; b – log-log coordinates. A straight line follows Eq. (1). Cut-in: water level histogram covering a period from May 1 till May 30.

complicated curve. In contrast, the distribution plotted in the log-log coordinates exhibited a broken log-linear dependence

$$\log_{10}N(H>h) \propto -b\log_{10}(h) \qquad (2)$$

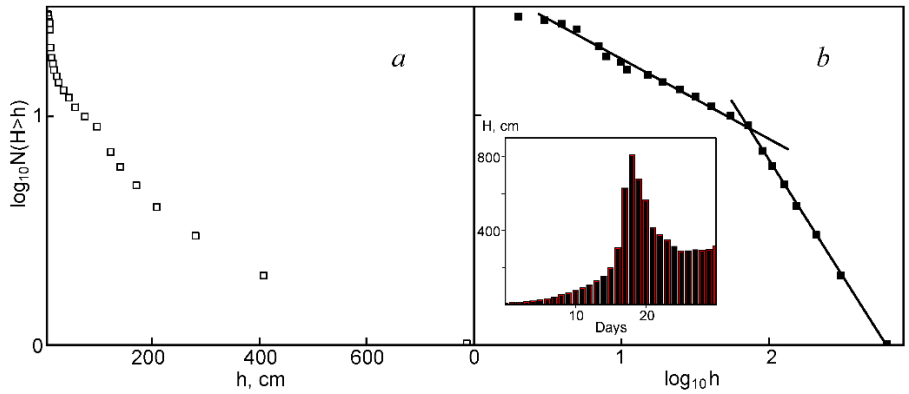



Figure 2. The water height distribution during spring breakup in 2014, which happened with an ice-jam; a – semi-logarithmic coordinates; b – log-log coordinates. Straight lines follow Eq. (2). Cut-in: water level histograms covering the period from April 1 till May 9.

consisted of the two straight portions characterized by different values of *b*. The relation (2) represents the power law function:

$$N(H>h) \propto h^{-b}, \qquad (2a)$$

which is specific for dynamics of open, non-equilibrium, multiple-event systems. Contrary to the rapidly descending exponential function, the power law provides long-range interactions between elements/events, which interdepend both in space and time [6](Blöschl and Sivapalan 1995). As a rule, the *b*-value increased a few days before ice-jamming.

The *b*-value characterizes a relative amount of "large" and "small" events in the $N(H>h)$ versus *h* distribution: the higher *b*-value the more contribution of smaller events. The above-mentioned increase of this parameter (Fig 2) means the prevailing of smaller changes in water height just before the jam.

### 3.3 Blasting

An example of the histogram of the water level variation during a catastrophic flood caused by a steady ice-jam downstream the town of Lensk is depicted in Fig. 3 together with the water height distribution. This flood required the full public evacuation. Three series of blasting were applied to the growing jam.

Explosive have been used to remove ice-jams for well over a hundred years ([7] Hanamoto et al., 1986), and some attempts to enhance its efficiency are continuing up to now [8], though the practice of blasting has been largely reduced since the 20th century due to limited success. In most cases, the ice-jam blasting is implemented as a last resort [9]. In this particular case, the primary action was performed by a terrestrial blasting team when the water level exceeded the critical height over 390 cm. The water height histogram exhibits a ravine in a short period of time after this blasting. However, the advance of ice stopped shortly with forming a new large scale (80 km in long) ice-jam downstream from the town of Lensk. The water level began to rise rapidly inundating the residential area and neighboring lands. The second blasting (2-d, four tons of explosive agent) performed by an aircraft was fully ineffective (Fig. 3a). The most powerful blasting (16 tons of explosive agent) was applied by a helicopter Mi-6. The blasting did not result in the breaking of ice-jam, but the log-linear water height distribution showed a slope variation after the action (Fig. 3b). The ice-jam broke up spontaneously within 20 hours after the most powerful final blasting. After the breakup of the jam, the system water-ice decomposed, and the used above analysis of water level variations became inapplicable to free flowing. The height distribution during the lowering of water levels after the ice-jam looks like almost vertical plot.

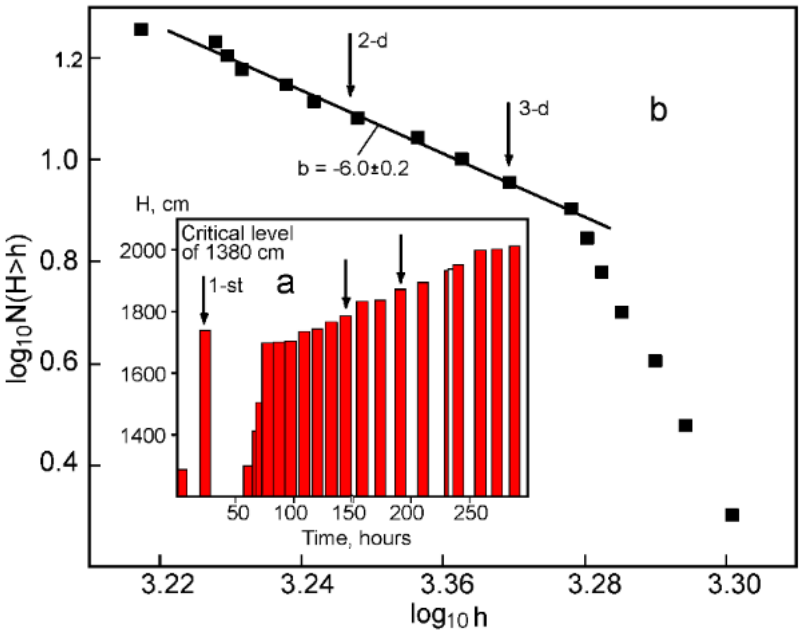



Figure 3. Water height histogram (a) and water height distribution (b) covering a period from May 14 till May 18. Arrows indicate blasting. Straight lines follow Eq. (2).

## 4. Discussion

The crucial condition of the power law dependence of the number of events on their intensities is the non-equilibrium state of the statistical system. In an equilibrium system that does not have any energy intake, possible fluctuations of the state decay quickly in accordance with the exponential law such as a relation (1a). In the latter case, the system's dynamics proceeds in a random manner. In this study, the water level distribution of this kind was obtained for a jam-free flood (Fig. 1). The power law distribution manifested itself only in the floods induced by ice-jams. In our statistical analysis of many spring floods, we obtained that in the most of ice-jams, the *b*-value exhibited a sudden increase a few days/hours before the jam disappearing. To relate changes in the *b*-value with the actual state of the hydrological system, one should take into account that the scale invariance expressed through a power law takes place only in sufficiently conservative medium, which would be able to retain some incoming energy. Christensen and Olami [10] used a spring-block model for the computer simulation of the seismic activity, in which the energy conservation in the system was tuned with a certain parameter of conservation. The model calculations showed that the lower the level of conservation, the higher the *b*-value. Moreover, there existed a critical value of the parameter, below which a power low dependence transformed into an exponential decay.

In the case of ice-jam formation, the energy of the water-ice system grows due to the elastic deformation of the congealing and densifying ice cover. The long-term data show that prior to almost each jam some limited ice movements and openings take place. The ice cover loses partially its connectedness — the conservation of the water-ice system reduces. As a result, the fragmentation of the ice cover leads to the increase of the *b*-value, which manifests itself in a break of the log-linear dependence. After the breakup of the jam, the system becomes highly

dissipative — the $N(H>h)$ versus $h$ dependence follows an exponential law. A $b$-value variation could display itself not only in a natural way but being caused by an anthropogenic factor. The blasting is the latter case.

The statistical analysis of the water height time series might reveal some ill-detectable perturbations in the evolving ice jam including its response to the actions of blasting. Figure 3 presents an example of three consequential applications of the ice-jam blasting. All blasting series were recognized as disadvantageous.

From the viewpoint of the main goal of the actions, the final blasting in 2001 (HS 3030, Lensk) was unsuccessful. Meanwhile, the slope of the $N(H>h)$ versus $h$ dependence after the third blasting increased notably thus signalizing a certain change in the flood dynamics. One can suppose that final blasting induced some inner ice cover rearrangements, which accelerated the jam breakup next day.

## 5. Conclusion

Dynamics of the ice-jam evolution was considered as the multi-scale process in an open, non-equilibrium system with varying conservation. The water rise time series during spring breakups were found to be random (Poissonian-like) when no ice-jams occurring. The water height daily distributions during ice-jam flooding followed a power law similarly to many natural large-scale hydrological processes. The power law exponent ($b$-value) depends on the actual ability of the water-ice system to retain the energy accumulated in the elastically deformed ice-cover.

In many cases, the $b$-value increased steeply a few days before the jam. The effect was caused by the variation of the energy balance in the system water-ice due to some natural or anthropogenic factors. The fragmentation of ice disturbs the ice cover connectedness thus reducing the degree of conservation of the system. The congealing and densifying of the ice cover produce an opposite outcome. The actions of blasting could cause both these effects. The analysis of the water height distributions enables to disclose some latent perturbations in the water-ice system caused by natural causes or blasting.

## References


1. Gutknecht, D.: Grundphanomene hydrologischer Prozesse. Zür. Geogr. Schr. 53, 25–38 (1993).
2. Machiwal, D., Jha, M.K.,: Time series analysis of hydrologic data for water resources planning and management: A review. J. Hydrol. Hydromech. 54, 237–257 (2006)
3. Mazzarella, A., Rapetti, F.: Scale-invariance laws in the recurrence interval of extreme floods: An application to the upper Po river valley (northern Italy). J. Hydrol. 288, 264–271 (2004).
4. Dickman, R.: Fractal rain distributions and chaotic advection. Brazilian J. Phys. 34, 337–346 (2004).
5. Peñate, I., Martín-González, J.M., Rodríguez, G., Cianca, A.: Scaling properties of rainfall and desert dust in the Canary Islands. Nonlin. Process. Geoph. 20, 1079–1094 (2013).
6. Blöschl. G., Sivapalan, M.: Scale issues in hydrological modelling: A review. Hydrol. Process. 9, 251–290 (1995).
7. Hanamoto, B., Perham, R., Rand, J.: Ricer Lake Ice Engineering. In: Ashnon, G.D. (ed.) Ice control. Water Resources Publications. LLC. Highlands Ranch, Collorado, ch. 7 (1986).
8. Wang, T., Guo, X., Fu, H., Guo, Y., Peng. X., Wu, Y., Li, J., Xia, Y.: Effects of water depth and ice thickness on ice cover blasting for ice jam flood prevention: A case study on the Heilong river, China. Water 10, 700 (13 pages) (2018).
9. Eichelberger, J., Gavrileva, T.: Reducing spring flood impacts to wellbeing of communities of the Nord. U.S. – Russia peer-to-peer dialogue initiative 2015-2016. Quarterly report October-December 2015.
10. Christensen, K., Olami, Z.: Variation of the Gutenberg-Richter b-values and non-trivial temporal correlations in a spring-block model for earthquakes. J. Geophys. Res. 97, 8729–8735 (1992).

# Keyword Index

# Author Index