ITISE 2018

International Conference on Time Series and Forecasting

PROCEEDINGS OF PAPERS

Volumen 1

ITISE 2018
International Conference on Time Series and Forecasting

Proceedings of Papers
19-21 September 2018
Granada (Spain)

Editors and Chairs
Olga Valenzuela
Fernando Rojas
Héctor Pomares
Ignacio Rojas

I.S.B.N: 978-84-17293-57-4
Legal Deposit: Gr 1165-2018
Edit and Print: Godel Impresiones Digitales S.L.

Preface

We are proud to present the set of final accepted papers for the fourth edition of the ITISE 2018 conference "International work-conference on Time Series" held in Granada (Spain) during September, 19-21, 2018.

The ITISE 2018 (International work-conference on Time Series) 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 2018 is to create a friendly environment that could lead to the establishment or strengthening of scientific collaborations and exchanges among attendees, and therefore, ITISE 2018 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 3.2, program committee members or additional reviewer). In this proceedings we are presetting the abstract of the contribution to be presented during ITISE-2018 (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. Dr. Peter M Robinson , Tooke Professor of Economic Science and Statistics Department of Economics, London School of Economics .

2. Prof Andrew C. Harvey, Emeritus Professor of Econometrics in the Faculty of Economics, University of Cambridge, and a Fellow of Corpus Christi College.

3. Prof. Salah Bourennane, Aix Marseille Univ, CNRS, Centrale Marseille, Institut Fresnel, Marseille, France.

4. Dr Karsten Webel, Deutsche Bundesbank, Central Office, Directorate General Statistics Germany.

5. Prof. Dr. Robert Kunst, Professor of Economics at the University of Vienna and affiliated with the IHS (Institute for Advanced Studies) .

6. Prof. Dr. Uwe Hassler, Applied Econometrics and International Economic Policy. Goethe University Frankfurt .

During ITISE 2018 several Special Sessions will be carried out. Special Sessions will be a very useful tool in order to complement the regular program with new and emerging topics of particular interest for the participating community. From the organization of ITISE, we would like to thank deeply the great work that the organizers of Special Sessions do. Thank you very much for your great effort and interest.

Special Sessions that emphasize on multi-disciplinary and transversal aspects, as well as cutting-edge topics are especially encouraged and welcome. and in this edition of ITISE 2018 are the following:

1. *Forecasting Evolution*, Prof. Philip Gerrish, School of Biology, Georgia Institute of Technology, 310 Ferst Dr, Atlanta, GA 30332 .

2. *Forecasting Climate Weather and Operation Impact on Reliability, Safety and Resilience of Critical Infrastructures*, Prof. Krzysztof Kolowrocki, Gdynia Maritime University, Poland, and Prof. Joanna Soszynska-Budny, Gdynia Maritime University, Poland

3. *Applications of time series for hydro-climatic data*, Prof. Bruno Remillardi, Professor at HEC Montral. Consultant at the National Bank of Canada and Prof. Bouchra R. Nasri .

4. *Times series analysis in geosciences*, Prof. Eulogio Pardo-Igzquiza, Professor at Instituto Geolgico y Minero de Espaa (IGME) and Prof. Francisco Javier Rodrguez-Tovar, Depart. Estratigrafa y Paleontologa, University of Granada, Spain.

5. *Forecasting in High Dimension and Complex/Big Data* , Prof. Dr. Luis Javier Herrera and Prof. Dr. Ignacio Rojas , Dep. Computer Architecture and Computer Technology, University of Granada, Spain

6. *Quantum Computing*, Prof. Peter Gloesekoetter, Fachbereich Elektrotechnik und Informatik, Stegerwaldstrae 39, 48565 Steinfurt, Germany. and Dr. Bernd Burchard, Elmos Semiconductor AG, Germany.

7. *Computational Intelligence methods for Time Series*, Prof. Dr. Hctor Pomares , Dep. Computer Architecture and Computer Technology, University of Granada, Spain and Prof. Dr. German Gutierrez , Dep. Computer Science, E.P.S. University Carlos III of Madrid, Spain

8. *Structural Time Series Models*, Prof. Dr. Fernando Rojas , Dep. Computer Architecture and Computer Technology, University of Granada, Spain

9. *Recent Developments on Time-Series Modelling*, Prof. Dr. Olga Valenzuela, Applied Mathematics, University of Granada, Spain

10. *Expert Systems with Time Series - Data*, Prof. Dr. Kalle Saastamoinen , Department of Military Technology, National Defence University,Helsinki, Finland

11. *Spatio-temporal brain dynamics in attention tasks*, Prof. Dr. Juan Manuel Grriz , University of Granada, Spain, and Prof. Dr. Pedro A. Valdes-Sosa , Cuban Neurosciences Center and Prof. Dr. Csar Germn Castellanos Dominguez , Universidad Nacional de Colombia

This new edition of ITISE was organized at the Universidad de Granada, with the help of the Spanish Chapter of the IEEE Computational Intelligence Society and 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 2018.

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, 2018
Granada

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

Program Committee

Bahram Abediniangerabi	University of Texas at Arlington
Sajjad Ahmad	University of Nevada, Las Vegas
Dorel Aiordachioaie	University Dunarea de Jos of Galati
Jose M. Amigo	Universidad Miguel Hernandez
Josu Arteche	University of the Basque Country UPV/EHU
Marcel Ausloos	GRAPES
José Luis Aznarte M.	Artificial Intelligence Department - UNED
Rosangela Ballini	IE - DTE - UNICAMP
Yukun Bao	School of Management, Huazhong University of Sci.&Tech.,
Ildar Batyrshin	Instituto Politecnico Nacional
Salah Bourennane	Institut Fresnel
Bernd Burchard	ELMOS
German Castellanos	Universidad Nacional de Colombia
João P. S. Catalão	University of Porto
Lee Chang-Yong	Kongju National University
Fuxia Cheng	Illinois State University
Paulo Cortez	University of Minho
Pierpaolo D'Urso	Sapienza University of Rome
Ricardo de A. Araújo	Laboratório de Inteligência Computacional do Araripe / Instituto Federal do Sertão Pernambucano
Francisco Estrada	1) Centro de Ciencias de la Atmósfera Universidad Nacional Autónoma de México; 2) Institute for Environmental Studies, VU University Amsterdam
Peter Gloesekoetter	Muenster University of Applied Sciences
Jesus Gonzalo	U. Carlos III de Madrid
Alberto Guillen	University of Granada
German Gutierrez	University Carlos III
Juan M. Gálvez	University of Granada
Ferda Halicioglu	Istanbul Medeniyet University
Marc Hallin	Université libre de Bruxelles
Uwe Hassler	Goethe University Frankfurt
Luis Herrera	University of Granada
Tzung-Pei Hong	Department of Computer Science and Information Engineering, National Univesity of Kaohsiung
Wei-Chiang Hong	School of Education Intelligent Technology, Jiangsu Normal University, China
Samrad Jafarian-Namin	Yazd University
Vinayakam Jothiprakash	none
Sreekanth K J	KISR
Rebecca Killick	Lancaster University
Alexey Koronovskiy	Saratov State University, Faculty of Nonlinear Processes
Dalia Kriksciuniene	Vilnius University
Clifford Lam	London School of Economics and Political Science
Elmar Lang	University of Regensburg
Hooi Hooi Lean	Universiti Sains Malaysia
Junsoo Lee	University of Alabama

Chunshien Li	Department of Information Management, National Central University
Carlos Lima	University of Brasilia
Hui Liu	Central South University, China & University of Rostock, Germany
Wieslaw M. Macek	Space Research Centre, Polish Academy of Sciences
Francisco Martínez-Álvarez	Universidad Pablo de Olavide
Anke Meyer-Baese	FSU
Janusz Miśkiewicz	University of Wrocław
Antonio Montañés	University Zaragoza
Miquel Montero	Universitat de Barcelona
Nageswara Rao Moparthi	V R SIDDHARTHA ENGINEERING COLLEGE
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
Eulogio Pardo-Iguzquiza	Geological Survey of Spain (IGME)
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
Bruno Remillardi	HEC Montreal
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
Gerhard Rünstler	European Central Bank
Kalle Saastamoinen	National Defence University of Finland
Leonid Sheremetov	Mexican Petroleum Institute
Ansgar Steland	RWTH Aachen University
Yixiao Sun	University of California San Diego
Leopold Sögner	Institute for Advanced Studies
Ryszard Tadeusiewicz	AGH University of Science and Technology, Krakow, Poland
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
Michael Wolf	University of Zurich
Ruqiang Yan	Southeast University
Gilney Zebende	UEFS
Wei-Xing Zhou	ECUST

Table of Contents

Energy Forecasting

Real macroeconomic monitoring and forecasting (Part. I)

Atmospheric Science Forecasting

Advanced econometric methods

Health Forecasting

Econometric models (Part.I)

Computational Intelligence methods for Time Series

Nonstationarity Time Series

Real Macroeconomic Monitoring and Forecasting (Part.II)

Advanced methods in Forecasting

Econometric models (Part.II)

Advanced in Time Series and Forecasting. Virtual Presentation

Note on Whittle Type Estimation under Long Memory and Nonstationarity*

Ying Lun Cheung and Uwe Hassler
Goethe University Frankfurt†

August 29, 2018

Abstract

We consider six variants of (local) Whittle estimators of the fractional order of integration d. They follow a limiting normal distribution under stationarity as well as under (a certain degree of) nonstationarity. Experimentally, we observe a lack of continuity of the objective functions of the two fully extended versions at $d = 1/2$ that has not been reported before. It results in a pile-up of the estimates at $d = 1/2$ when the true value is in a neighbourhood to this half point. Consequently, studentized test statistics may be heavily oversized. The other four versions suffer from size distortions, too, although of a different pattern and to a different extent.

Keywords: Discontinuity; fractional integration.
JEL classification: C12 (hypothesis testing), C22 (time series models)

1 Introduction

Strong persistence and long memory are considered as stylized facts in many empirical time series from a variety of fields. For surveys from economics and finance over political sciences to hydrology see e.g. Baillie (1996), Box-Steffensmeier and Tomlinson (2000) and Montanari (2003), respectively. Fractional integration of order d is the most widely used model to capture strong

*The authors thank Mehdi Hosseinkouchack for helpful comments.
†Campus Westend, Economics and Business Administration, 60323 Frankfurt, Germany

persistence and long memory. Estimated values of d vary between 0 and 1, with $d = 1/2$ separating the reign of stationarity from nonstationarity. The properties of the so-called (local) Whittle estimator depend on the true d_0, which is a nuisance, since this parameter is not known a priori; see Robinson (1995), Velasco (1999) and Velasco and Robinson (2000). This is not the case with the exact local Whittle estimator proposed by Shimotsu and Phillips (2005) and the tapered local Whittle estimator by Velasco (1999); the former, however, is burdened by the necessity of mean estimation, see Shimotsu (2010), while the latter is plagued by variance inflation. Similarly, the fully nonstationarity-extended (local or parametric) Whittle estimators that have been proposed by Abadir, Distaso, and Giraitis (2007) and Shao (2010), respectively, do not depend on the true d_0, although the half point is not covered by their theory. Abadir et al. (2007, p. 1366), however, "conjecture that our asymptotic expansions are valid for these points", where 'these points' are exactly the half points $1/2, 3/2, \ldots$. This conjecture is contrasted with the statement by Faÿ, Moulines, Roueff, and Taqqu (2009, p. 165) that "the approximation of the extended local Whittle function is not continuous at $d = 1/2, 3/2, \ldots$". Our findings support the second view. Moreover, extensive experimental evidence on the behavior of altogether five Whittle type estimators is provided. It sheds some doubt on the reliability of statistical inference in the empirically relevant range from long memory to nonstationarity.

The rest of this paper is organized as follows. The next section becomes precise on the Whittle type estimators. Section 3 demonstrates by means of computer experiments the lack of continuity of the objective functions of fully nonstationarity-extended Whittle estimators. Section 4 discusses the consequences for the estimates, and compares with the other three versions. Section 5 shows the effects on test statistics. A more detailed summary is provided in the final section.

A word on notation before we begin: Throughout this paper, $\Rightarrow$ stands for weak convergence as the sample size n diverges, and $\lfloor x \rfloor$ returns the largest integer smaller than or equal to $x \geq 0$, $x \in \mathbb{R}$. Further, i denotes

the imaginary unit with $\mathrm{i}^2 = -1$, and the Landau symbol $o(\cdot)$ has its usual meaning.

2 Model and Estimators

For ease of exposition, we maintain a very simple and overly restricted model and do not give details of the assumptions discussed in the literature; see also Hassler (2019) for a recent survey. We assume a covariance stationary process $\{\xi_t\}$ with finite second moments that is fractionally integrated of order d_ξ. In the frequency domain this amounts to the assumption of a spectrum characterized by the following behavior at the origin:

$$f_\xi(\lambda) = b_0|\lambda|^{-2d_\xi} + o(|\lambda|^{-2d_\xi}) \,, \quad \lambda \to 0 \,, \; b_0 > 0 \,, |d_\xi| < 1/2 \,. \tag{1}$$

More generally, we allow for difference-stationary processes $\{y_t\}$ by

$$\Delta^p y_t = \xi_t \,, \quad p \in \{0, 1, 2, ...\} \,, \; t = 1 - p, 2 - p, \ldots, n \,, \; d = p + d_\xi \,, \tag{2}$$

where $\Delta = 1 - B$ are the usual differences, such that $\{y_t\}$ is fractionally integrated of order $d = p + d_\xi$ (often called of type I since the work by Marinucci and Robinson (1999)).

Since Whittle estimation is settled in the frequency domain, we define the periodogram at the Fourier frequencies $\lambda_j = 2\pi j/n$,

$$I_y(\lambda_j) := |w_y(\lambda_j)|^2 \text{ with } w_y(\lambda_j) := \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} y_t e^{\mathrm{i}t\lambda_j}, \tag{3}$$

where $w_y(\lambda_j)$ is the usual discrete Fourier transform (DFT) evaluated from a stretch of data of length n. Note, that the model (1) is semiparametric or local in that only an assumption within an arbitrarily small neighborhood of frequency 0 is made. Consequently, the estimation is restricted to a neighborhood of the origin, which requires the choice of a bandwidth m with $m = o(n)$.

Following Phillips (1999), Abadir et al. (2007) considered the so-called extended periodogram at the Fourier frequencies:

$$I_y(\lambda_j;d) := \left|w_y(\lambda_j;d)\right|^2, \quad w_y(\lambda_j;d) := w_y(\lambda_j) + k(\lambda_j;d)\,. \tag{4}$$

The extension $w_y(\lambda_j;d)$ depends on the nonzero correction term $k(\lambda_j;d)$ for $d \geq 1/2$:

$$k(\lambda_j;d) := \begin{cases} 0\,, & d \in [-\frac{1}{2},\frac{1}{2}) \\ \frac{e^{\mathrm{i}\lambda_j}}{\sqrt{2\pi n}}\frac{y_n - y_0}{1-e^{\mathrm{i}\lambda_j}}\,, & d \in [\frac{1}{2},\frac{3}{2}) \end{cases}\,. \tag{5}$$

Note that one now assumes observed data from y_0 through y_n, where y_0 enters through $k(\cdot;d)$, while the standard DFT $w_y(\cdot)$ builds on $y_1,\ldots,y_n$ only. An alternative, equivalent representation of the extended periodogram has been given by Abadir, Distaso, and Giraitis (2011, p. 190). The objective function to be minimized becomes

$$U_n(d) := \ln\left\{\frac{1}{m}\sum_{j=1}^{m} j^{2d} I_y\left(\lambda_j;d\right)\right\} - \frac{2d}{m}\sum_{j=1}^{m}\ln j\,.$$

The minimizer has been called fully extended local Whittle estimator, in short FELW: $\widehat{d}_{FELW} = \arg\min U_n\left(d\right)$. Let $d_0 = p_0 + d_{\xi,0}$ denote the true parameter value. Under some further conditions spelled out in Abadir et al. (2007, Cor. 2.1), $\widehat{d}_{FELW}$ follows

$$\sqrt{m}\left(\widehat{d}_{FELW} - d_0\right) \;\Rightarrow\; \mathcal{N}\left(0,\frac{1}{4}\right) \quad \text{as } \frac{1}{m} + \frac{m}{n} \to 0 \tag{6}$$

for $-\frac{1}{2} < d_0 < \frac{3}{2}$, $d_0 \neq \frac{1}{2}$. Note that we restrict the parameter range to the empirically most relevant region of $(-1/2, 3/2)$. Abadir et al. (2007) are more general, allowing for $d_0 = p_0 + d_{\xi,0}$ with $p_0 \in \{-1,0,1,\ldots\}$, which of course requires to generalize the correction term $k(\lambda_j;d)$ accordingly.

If the extended periodogram in U_n is replaced by the conventional periodogram $I_y(\lambda_j)$, then the usual local Whittle (LW) estimator results: $\widehat{d}_{LW}$. Distributional properties have been established by Robinson (1995, Thm. 2) under stationarity and have been generalized Velasco (1999, Thm. 3) up to 3/4. More precisely, the limiting law form (6) holds for $\widehat{d}_{LW}$ in the range

$-1/2 < d_0 < 3/4$. Beyond 3/4, this limiting distribution breaks down according to Velasco (1999), although the estimator remains consistent up to $d_0 = 1$.

Velasco (1999) discussed tapering of the data in order to extend the range of validity in (6). We employ the so-called Žurbenko-Kolmogorov taper of order $P = 3$, see also Žurbenko (1979). Given a bandwidth m as before, the summation in the objective function is now over only $m/3$ harmonic frequencies. Minimization yields the so-called tapered local Whittle estimator, $\widehat{d}^{(3)}_{TLW}$. It holds for true $-1/2 < d_0 \leq 5/2$ that

$$\sqrt{\frac{m}{3}}\left(\widehat{d}^{(3)}_{TLW} - d_0\right) \;\Rightarrow\; \mathcal{N}\left(0, \frac{\Phi_3}{4}\right), \quad \Phi_3 \approx 1.00354\,, \tag{7}$$

under some additional assumptions spelled out in Velasco (1999, Thm. 6). A nice byproduct of tapering is that a potential quadratic time trend is removed from the series without having to estimate it. The price for the nice features of tapered LW is of course the variance inflation: $\mathrm{Var}(\widehat{d}^{(3)}_{TLW}) \approx 3\,\mathrm{Var}(\widehat{d}_{LW})$.

Now, we turn to the parametric Whittle estimation. To simplify the exposition, we treat only the special case of fractionally integrated noise (FIN) under homoskedasticity. There is no reason to expect that a more complicated setting would remove the lack of continuitiy or the size distortions reported in the next sections. Hence, $\{\varepsilon_t\}$ is white noise, $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{E}(\varepsilon_t^2) = \sigma^2$, and $\mathrm{E}(\varepsilon_t\varepsilon_s) = 0$, $t \neq s$. Further, $\{\xi_t\}$ behind (2) becomes FIN:

$$\xi_t = \Delta^{-d_\xi}\varepsilon_t\,, \qquad |d_\xi| < 1/2\,,\ t \in \mathbb{Z}\,. \tag{8}$$

The (pseudo-) spectrum of $\{\xi_t\}$ (and $\{y_t\}$) is given by

$$f(\lambda) = |1 - e^{\mathrm{i}\lambda}|^{-2d_0}\frac{\sigma^2}{2\pi}\,, \quad \lambda \in (0, \pi]\,,$$

with d_0 from (2). Shao (2010) adopted the idea by Abadir et al. (2007) to replace the periodogram by the extended periodogram. For our case of fractionally integrated noise the objective function becomes

$$L_n(d) := \frac{2\pi}{n}\sum_{j=1}^{n-1}|1 - e^{\mathrm{i}\lambda_j}|^{2d} I_n\left(\lambda_j; d\right).$$

The minimizer is called nonstationarity-extended parametric Whittle (NEPW) estimator: $\widetilde{d}_{NEPW} = \arg\min L_n(d)$. Under conditions given by Shao (2010, Thm. 3.2) it holds that

$$\sqrt{n}\left(\widetilde{d}_{NEPW} - d_0\right) \;\Rightarrow\; \mathcal{N}\left(0,\, \frac{6}{\pi^2}\right), \quad \text{as } n \to \infty, \tag{9}$$

with the parameter range $d_0 \in \left(-\frac{1}{2}, \frac{1}{2}\right) \cup \left(\frac{1}{2}, \frac{3}{2}\right)$. Again, a generalization for $d_0 = p_0 + d_{\xi,0}$ with $p_0 \in \{0, 1, 2, \ldots\}$ is possible. The variance expression in (9) follows from Hassler (2019, Cor. 8.2 b).

Classical parametric Whittle estimation of d results when inserting $I_y(\lambda_j)$ for the extended periodogram in L_n: $\widetilde{d}_W$. It has been pioneered by Fox and Taqqu (1986), Dahlhaus (1989) and Giraitis and Surgailis (1990) for the region of stationarity, and distributional results have been extended to nonstationarity by Velasco and Robinson (2000). Under conditions given by Velasco and Robinson (2000, Thm. 1 and 2), (9) continues to hold for $\widetilde{d}_W$, however, in the range $-1/2 < d_0 < 3/4$; again, consistency is guaranteed up to $d_0 = 1$. Velasco and Robinson (2000) also discussed a tapered version of the Whittle estimator along the lines of Velasco (1999); we omit details.

Finally, we turn to the exact local Whittle estimator (ELW), which is semiparametric again. The idea is to replace $j^{2d} I_y(\lambda_j)$ by the periodogram of the fractionally differenced data, $I_{\Delta_+^d y}(\lambda_j)$. The model behind this procedure is a so-called type II process where the filter is truncated at the beginning of the sample. Note that the previous four estimators are robust with respect to means different from zero, i.e. (6) and (9) continue to hold when replacing y_t by $\mu + y_t$. This no longer holds true for ELW. Therefore, Shimotsu (2010) suggested to demean the observations prior to differencing, where appropriate demeaning without knowing the true d_0 a priori is a delicate issue. Shimotsu (2010, Sect. 4.1) proposed to use a preliminary or first-step consistent estimator. To that end, we employ $\widehat{d}^{(3)}_{TLW}$ as recommended by Shimotsu (2010, p. 511). We abbreviate this two-step ELW estimator by 2ELW, $\widehat{d}_{2ELW}$. Shimotsu (2010, Thm. 3) proved that (6) continues to hold, however, for the wider range $-\frac{1}{2} < d_0 < \frac{3}{2}$, *without excluding* the half point. In fact, the range can be even extended to $-1 < d_0 < \frac{7}{2}$.

3 Discontinuity of extended Whittle estimation

All experimental results in this paper are based on simulation with 5000 repetitions for each experiment performed with matlab. The true model is fractionally integrated noise,

$$(1-B)^{d_0} y_t = \varepsilon_t \sim iid(0,1)\,, \tag{10}$$

where the independent innovations are from a standard normal distribution. For the semiparametric estimators LW, FELW and 2ELW the bandwidth is chosen as $m = \lfloor n^{0.65} \rfloor$, following recommendations by Abadir et al. (2007, p. 1363) and Shimotsu (2010, p. 515).

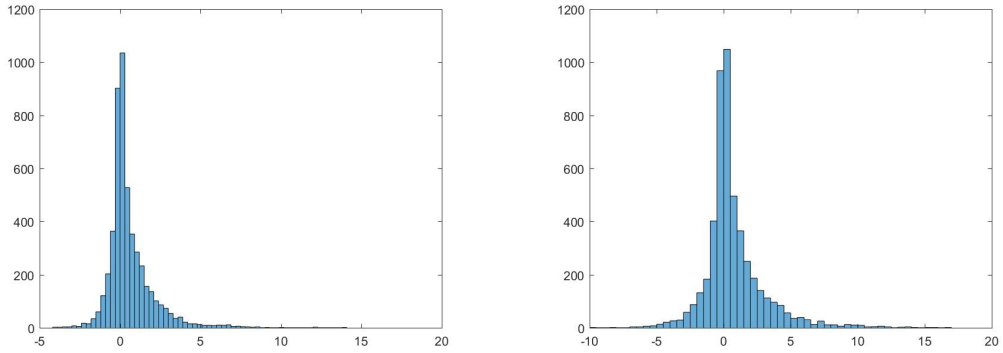
To begin with, we choose $d_0 = 1/2$ and observed three cases for the objective functions $L_n(d)$ and $U_n(d)$: The objective function drops when approaching the half point from the left (case 1), it drops coming from the right (case 2), or there is no obvious discontinuity. A glimpse into how often these cases occur is given in Figure 1 for NEPW. To that end, we separate the first and the second branch of the correction term, $k_b(\lambda_j) := k(\lambda_j; b)$ for $b = 0, 1$, and analogously for the extended periodogram: $I_b(\lambda_j) := |w_y(\lambda_j; b)|^2$, $b = 0, 1$. We define two corresponding branches $L_n(d; b)$, such that the criterion function becomes

$$L_n(d) = \begin{cases} L_n(d;0) = \frac{2\pi}{n} \sum_{j=1}^{n-1} |1 - e^{\mathrm{i}\lambda_j}|^{2d} I_0(\lambda_j)\,, & -\frac{1}{2} \le d < \frac{1}{2} \\ L_n(d;1) = \frac{2\pi}{n} \sum_{j=1}^{n-1} |1 - e^{\mathrm{i}\lambda_j}|^{2d} I_1(\lambda_j)\,, & \frac{1}{2} \le d < \frac{3}{2} \end{cases} .$$

We now evaluate the difference at the half point, $L_n(1/2;0) - L_n(1/2;1)$. With true $d_0 = 1/2$, we report the absolute frequencies of these differences. If $L_n(1/2;0) - L_n(1/2;1) > 0$, this falls into the category of case 1. According to Figure 1, this occurs more often and with larger jumps than case 2 ($L_n(1/2;0) - L_n(1/2;1) < 0$). We note that the potential of discontinuity does not vanish with growing sample size, it stands out for $n = 5000$ just as strongly as for $n = 500$. Cheung and Hassler (2018) provided a similar picture for FELW. Further, similar pictures arise when simulating with $d_0 = 0.4$

or $d_0 = 0.6$. More interesting than the discontinuity itself, however, is how this affects the estimates and resulting test statistics.

Figure 1. Absolute frequencies of $L_n(1/2;0) - L_n(1/2;1)$ for $n = 500$ (left) and $n = 5000$ (right)

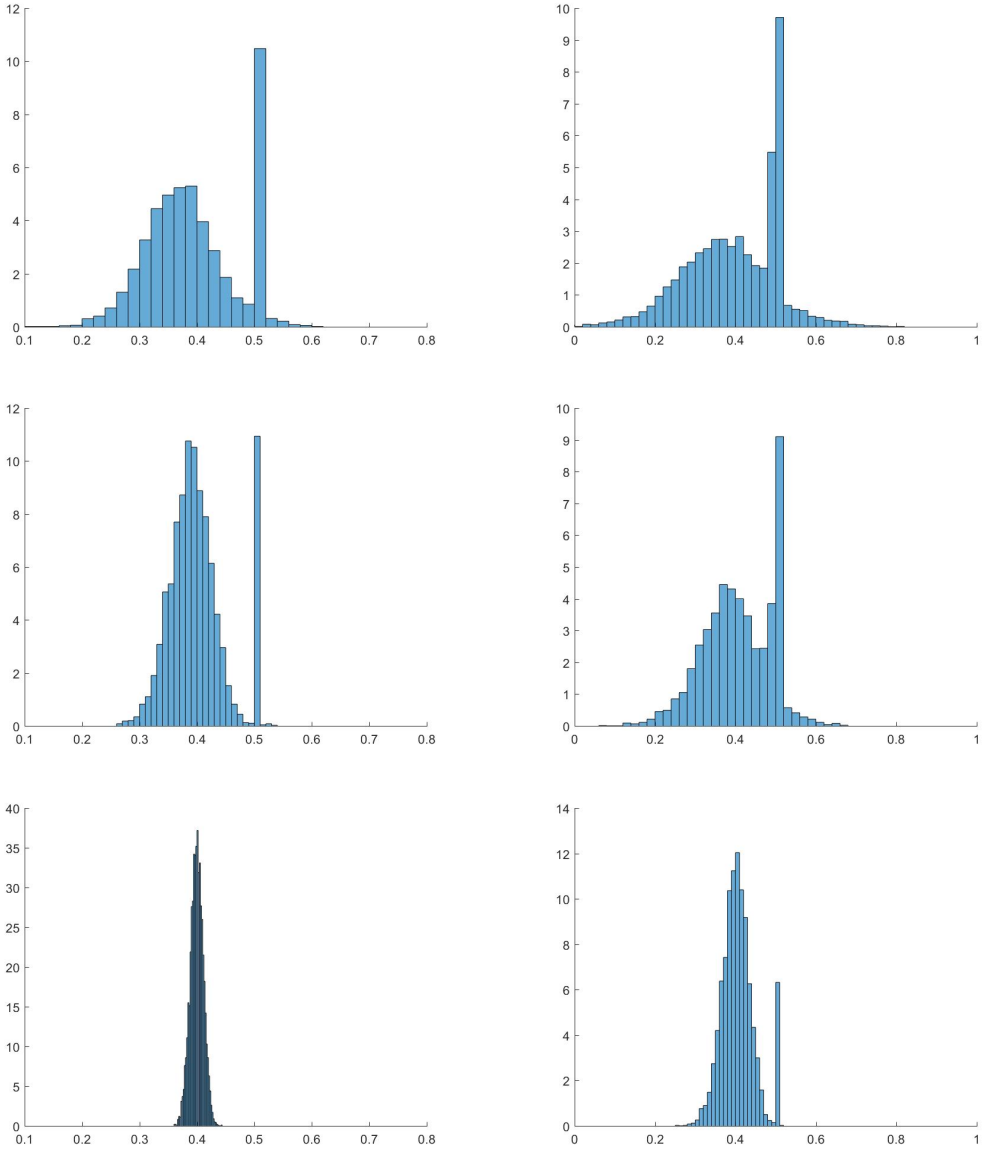


Note The model is $(1-B)^{1/2} y_t = \varepsilon_t$

4 Finite sample distribution

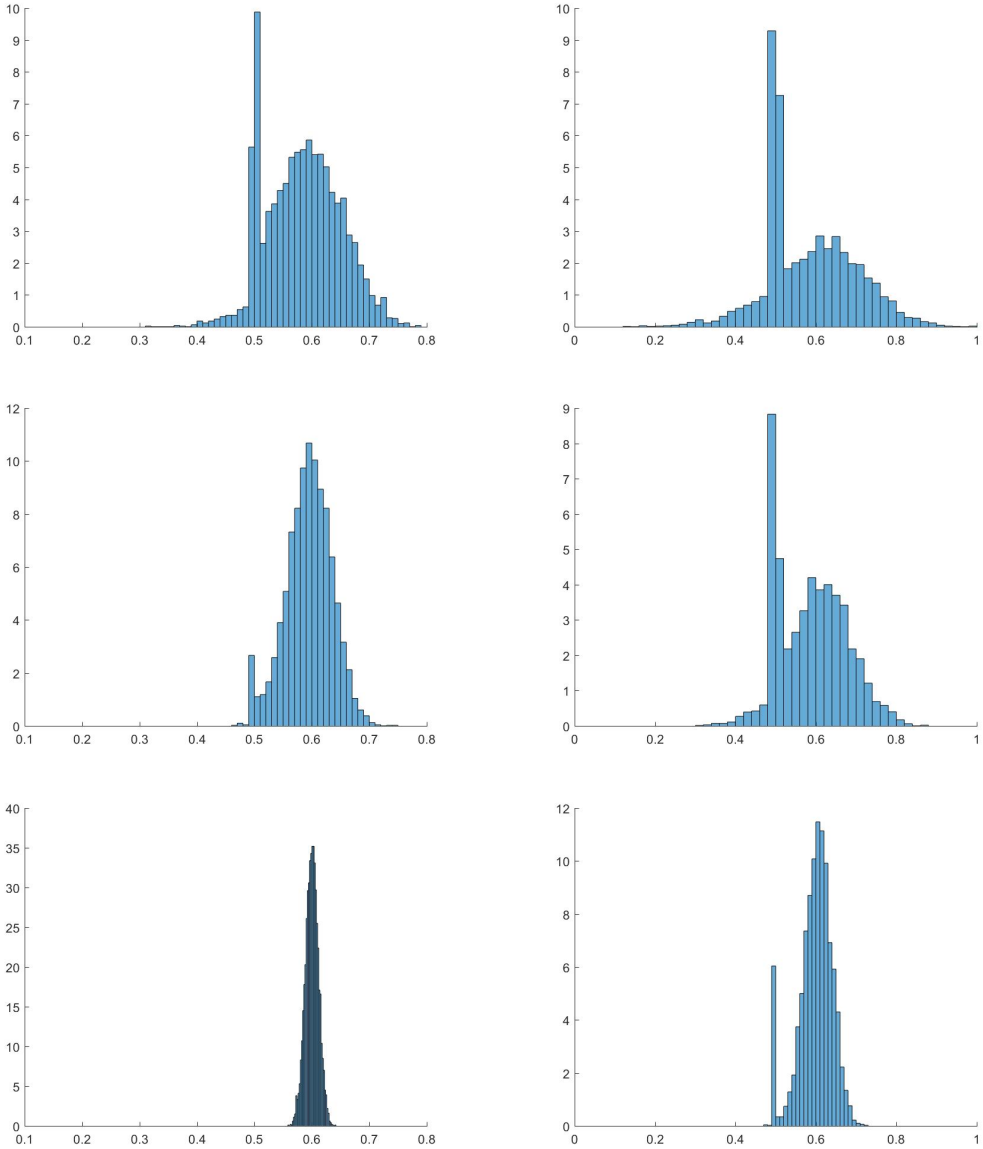
We continue with extended Whittle estimators and FIN with true d_0 close to the half point: $d_0 = 0.4$ and $d_0 = 0.6$. We report the histograms of the estimates in Figures 2 and 3 for $n = 200, 500$ and $n = 5000$ (top to bottom). For $n = 200$, we observe that $\widetilde{d}_{NEPW}$ picks $d = 1/2$ much too often, which is still true for $n = 500$, although less pronounced. For a very large sample size, $n = 5000$, the estimates spread nicely around the true values of $d_0 \in \{0.4, 0.6\}$. The distortions are more extreme for the semiparametric estimator, which is less efficient and converges more slowly. Even for $n = 5000$, $\widehat{d}_{FELW}$ suffers from the lack of continuity in $U_n(d)$ picking too often the borderline case $d = 1/2$. To sum up: both estimators are affected by a potential discontinuity of the objective functions, which become visible in the histograms of the estimates for d_0 "close" to the half point, where closeness is of course relative to the sample size and relative to the efficiency of the estimators.

Figure 2. Histogram of $\widetilde{d}_{NEPW}$ (left) and of $\widehat{d}_{FELW}$ (right) for $n = 200, 500$ and $n = 5000$ (top to bottom)



Note The model is $(1 - B)^{0.4} y_t = \varepsilon_t$

Figure 3. Histogram of $\widetilde{d}_{NEPW}$ (left) and of $\widehat{d}_{FELW}$ (right) for $n = 200, 500$ and $n = 5000$ (top to bottom)



Note The model is $(1 - B)^{0.6} y_t = \varepsilon_t$

We want to close this section with a remark drawn from further Monte Carlo experiments that are not reported here but available upon request. When extending the range of nonstationarity and allowing for $p = 2$ in (2), a similar discontinuity like in the previous section and a similiar pile-up in the histrograms like in this section are observed at the next half point, $d = 3/2$.

Finally, histograms for the other estimators, local Whittle, the parametric Whittle, and (two-step) exact local Whittle are presented in Cheung and Hassler (2018): No pile-up occurs since the objective functions are continuous.

5 Approximating Gaussianity

Robinson (1995), Abadir et al. (2007) and Shao (2010) reported Monte Carlo bias and mean squared error for their estimators, see also Hauser (1999). They did not, however, investigate the frequencies of rejection when performing a test at level α under the null hypothesis. We now turn to this issue building on (9) and (6). A two-sided test at level α hence rejects H_0: $d = d_0$ in the parametric cases if

$$|t_\ell| > z_{1-\alpha/2}\,,\ \widetilde{t}_\ell := \pi\sqrt{\frac{n}{6}}\left(\widetilde{d}_\ell - d_0\right)\,,$$

where the label stands for Whittle or nonstationarity-extended parametric Whittle, $\ell \in \{W, NEPW\}$, respectively, and $z_{1-\alpha/2}$ is the quantile from the standard normal distribution. Similarly, inference may rely for the semi-parametric estimators on $2\sqrt{m}\left(\widehat{d}_\ell - d_0\right)$, where $\ell \in \{LW, FELW, 2ELW\}$ for local Whittle, fully extended local Whittle, and (two-step) exact local Whittle. An alternative variance approximation can be motivated the Fisher information, see Cheung and Hassler (2018) for a discussion. It amounts to replacing m by

$$\nu := \sum_{j=1}^{m} \nu_j^2 \quad \text{with } \nu_j := \ln j - \frac{1}{m}\sum_{\ell=1}^{m} \ln \ell\,. \tag{11}$$

The decision rule then becomes

$$|t_\ell| > z_{1-\alpha/2}\,,\ \widehat{t}_\ell := 2\sqrt{\nu}\left(\widehat{d}_\ell - d_0\right)\,, \quad \ell \in \{LW, FELW, 2ELW\}\,.$$

According to Cheung and Hassler (2018), $2\sqrt{\nu}\left(\widehat{d}_\ell - d_0\right)$ clearly outperforms $2\sqrt{m}\left(\widehat{d}_\ell - d_0\right)$ in terms of size distortions. Note, however, that under tapering the decision rule is according to (7)

$$2\sqrt{\frac{m}{3\Phi_3}}\left|\widehat{d}^{(3)}_{TLW} - d_0\right| > z_{1-\alpha/2}\,.$$

We report frequencies for $\alpha = 0.01, 0.05, 0.10$ and $n = 200, 500, 1000$ with series generated from model (10). The first table contains results for the fully extended local Whittle estimation and nonstationarity-extended parametric Whittle estimator. Typically, two massive distortions (printed in bold) stand out in each row symmetrically around 1/2. They can be attributed to the pile-up effect discussed in the previous section. Their location is determined by the nominal level of the test and by the sample size, depending on whether the pile-up occurs in the relevant tail of the distribution or not. The strongest distortions are not necessarily very close to 1/2 but occur between 0.25 and 0.75. In this range rejections of hypotheses as well as confidence intervals may not be considered as reliable. Moreover, we observe that t_{NEPW} has a poorer size performance than t_{FELW}.

Next, we turn to the remaining Whittle type estimators, presenting evidence for a larger range of d_0 up to 1. Table 2 contains the result for the parametric Whittle estimator $\widetilde{d}_W$, where the limiting law (9) is guaranteed only up to 0.75. Indeed, up to $d_0 \leq 0.5$, the size distortions are moderate at least for $n = 1000$. In the subrange $0.55 \leq d_0 \leq 0.85$, however, the distortions grow with growing d_0. For $d_0 = 1$, the experimental sizes are actually pretty close to the nominal ones, which seems to be coincidential and is not covered by theory.

Similarly to $\widetilde{d}_W$, the local Whittle estimator satisfies the Gaussian law (6) only up to 0.75. Generally, its size performance according to Table 3 is superior to the parametric estimator. Surprisingly, it is even not so bad for

Table 1. Frequency of rejection for two-sided tests at level α

	d_0										
n	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
	FELW: t_{FELW}										
	$\alpha = 1\%$										
200	1.70	1.82	1.78	1.68	1.66	1.66	1.52	1.56	1.80	1.40	1.10
500	2.26	**5.12**	1.50	1.60	1.40	1.42	1.80	1.34	1.24	**6.68**	3.26
1000	1.70	2.20	**7.22**	1.00	1.72	1.22	1.36	1.24	**8.06**	3.14	1.94
	$\alpha = 5\%$										
200	**9.82**	5.58	5.82	6.30	5.70	6.10	5.46	5.46	4.92	4.26	**12.24**
500	6.30	9.00	**14.60**	5.32	5.02	5.72	5.60	5.10	**18.76**	11.24	7.64
1000	6.20	5.98	**11.18**	3.92	5.70	5.22	5.12	4.86	**12.42**	7.50	6.18
	$\alpha = 10\%$										
200	14.86	**18.94**	9.60	9.68	10.04	10.62	9.78	9.46	9.18	**21.36**	16.84
500	11.12	13.78	**18.74**	9.00	9.22	10.14	9.90	9.10	**23.20**	15.70	12.66
1000	11.42	11.22	15.82	**22.82**	10.02	9.40	9.76	**28.24**	16.84	12.46	11.64
	NEPW: t_{NEPW}										
	$\alpha = 1\%$										
200	4.16	5.98	**12.62**	4.30	4.28	3.94	3.26	2.28	**6.20**	4.28	3.82
500	2.30	2.12	3.30	**12.92**	2.28	2.08	1.52	**4.26**	2.36	2.32	2.26
1000	2.08	2.12	2.12	**5.40**	1.92	2.12	1.54	**2.54**	1.86	1.88	2.02
	$\alpha = 5\%$										
200	11.02	12.58	**18.54**	10.40	10.64	10.12	8.20	5.78	**12.22**	10.80	10.74
500	7.24	7.26	8.10	**17.16**	7.16	6.80	5.00	**9.34**	7.86	7.82	7.84
1000	6.92	6.80	6.84	9.56	**30.68**	6.62	**15.98**	7.28	6.74	6.80	6.84
	$\alpha = 10\%$										
200	17.38	18.68	24.26	**35.94**	16.26	15.34	12.20	**22.28**	17.88	17.54	17.24
500	13.88	13.82	14.24	**22.04**	12.52	11.88	9.04	**15.18**	13.60	13.28	13.50
1000	11.92	11.90	12.06	14.38	**34.42**	11.06	**20.56**	12.32	11.76	11.80	11.84

Note The model is $(1-B)^{d_0} y_t = \varepsilon_t$

$d_0 \geq 0.75$. Nevertheless, most people will be reluctant to rely on inference that is not justified by asymptotic theory. Therefore, it is interesting to look at exact local Whittle results in Table 4. Experimental sizes of t_{2ELW} are close to the nominal ones for $d_0 \geq 0.75$ or $d_0 \leq 0.4$, at least for $n = 1000$. For d_0 around 0.5, the test is oversized, while it is undersized for $0.6 \leq d_0 \leq 0.7$, and the distortion seems to even grow with n. Maybe, this is attributable to problems with demeaning for values of d in this range. All in all, however, inference with t_{2ELW} seems to be most reliable. It is worth emphasizing that we report results for simulated type I processes; for type II processes, 2ELW works even better. Further, we want to stress that t_{LW} and t_{2ELW} both build on replacing m by ν, since we collected simulation evidence not reported here that this improves the size behaviour. Finally, we turn to the tapered LW estimator, $\widehat{d}^{(3)}_{TLW}$, see Table 5. For $n = 1000$, the experimental size is remarkably close to the nominal one, and also for $n = 500$ the size distortions are mild in comparison with the previous tables. The price one pays for this fine behaviour is of course the variance inflation discussed above. Note that replacing m by ν would result for $\widehat{d}^{(3)}_{TLW}$ in conservative tests where

the experimental size is below the nominal one.

Before closing this section, it must be stressed that Tables 1 through 5 have been produced under the ideal model of fractionally integrated noise, $(1-B)^{d_0}y_t = \varepsilon_t$, such that the parametric estimators draw from a correctly specified model. Similarly, the semiparametric estimators are not confounded by additional short memory components that are known to induce a bias with distorted inference in finite samples.

Table 2. Frequency of rejection for two-sided t_W at level α

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
$d_0 \backslash n$	200	500	1000	200	500	1000	200	500	1000
0.25	3.62	2.34	1.44	10.68	7.86	6.20	17.02	13.66	12.08
0.3	3.60	1.96	1.72	9.82	7.88	6.38	16.14	14.06	11.22
0.35	3.10	2.02	1.70	9.60	6.88	6.92	15.62	13.14	12.76
0.4	3.50	2.02	1.50	10.04	7.30	6.46	16.22	13.56	11.94
0.45	3.36	2.14	1.76	10.06	7.94	6.64	16.68	13.64	12.22
0.5	3.54	2.44	1.48	9.86	8.18	6.78	16.30	14.48	12.50
0.55	2.90	2.08	1.90	9.38	8.22	7.82	15.72	14.36	13.48
0.6	3.32	2.62	2.52	10.52	8.88	8.58	16.98	15.80	14.98
0.65	3.64	2.62	3.24	10.18	9.58	9.90	17.12	16.32	16.20
0.7	4.56	3.70	4.64	12.02	10.98	11.38	19.32	17.34	18.08
0.75	3.60	4.00	6.48	11.26	12.04	14.06	18.10	18.76	20.94
0.8	3.54	4.52	7.28	11.18	12.70	16.14	17.44	19.72	22.92
0.85	3.10	4.28	8.20	10.04	12.40	17.32	16.38	20.26	24.82
0.9	2.48	2.48	5.90	8.94	9.50	14.96	14.84	16.16	22.74
0.95	2.20	1.52	1.78	7.36	6.64	7.78	12.86	11.86	14.20
1	2.54	1.66	1.42	7.86	5.68	5.02	13.16	10.78	9.80

6 Concluding remarks

We provide an experimental study on the behavior of six Whittle type estimators of the parameter d, which denotes the fractional order of integration. The focus is on the empirically most interesting (neighborhood of the) half

Table 3. Frequency of rejection for two-sided t_{LW} at level α

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
$d_0 \backslash n$	200	500	1000	200	500	1000	200	500	1000
0.25	1.90	1.82	1.40	6.42	6.42	5.84	11.28	11.28	11.12
0.3	1.88	1.32	1.30	6.10	5.86	6.18	11.06	11.10	11.60
0.35	2.12	1.36	1.56	6.60	6.30	6.08	11.84	11.80	11.72
0.4	1.92	1.22	1.46	6.42	6.28	6.46	11.46	11.52	12.08
0.45	2.12	1.64	1.22	6.72	6.44	5.90	12.04	11.98	11.54
0.5	1.96	1.64	1.54	6.94	6.90	5.80	12.62	12.58	11.22
0.55	1.92	1.64	1.48	7.10	6.44	5.90	12.94	12.20	11.62
0.6	1.32	1.48	1.96	6.22	7.06	7.12	11.70	12.76	13.16
0.65	2.16	1.74	1.90	7.42	6.90	6.92	13.18	12.32	13.26
0.7	2.22	2.14	2.50	7.38	7.92	8.50	13.14	13.92	15.22
0.75	1.78	2.04	3.02	6.56	7.42	9.62	12.18	12.96	15.98
0.8	1.74	2.26	2.54	6.82	7.72	9.40	12.16	13.14	15.68
0.85	1.68	1.74	2.04	5.98	6.94	8.04	10.72	12.80	14.46
0.9	1.64	0.92	1.22	5.54	4.54	6.48	11.18	9.44	12.62
0.95	1.52	1.04	0.96	5.02	4.70	4.50	9.40	9.44	9.30
1	1.32	1.28	0.84	4.74	4.50	3.68	8.56	8.54	7.88

point 1/2. We begin with two different extended Whittle objective functions. At $d = 1/2$, a discontinuity may occur in the objective function of the fully nonstationarity-extended (local or parametric) Whittle estimators by Abadir et al. (2007) or Shao (2010), respectively. This results in a pile-up of the frequency distribution of the estimates at 1/2 when the true d_0 is "close" to this value. This has consequences when testing for d_0 with a studentized statistic. Experimentally, we observe massive size distortions even for large sample sizes. The location of the strongest distortion depends on the nominal level of the test and the sample size, since the pile-up causes distortions only if located in the tail of the distribution. In fact, the maximum size distortions may occur for true values of d_0 between 0.25 and 0.75. The local Whittle, parametric Whittle and exact local Whittle estimators partly suffer from severe size distortions, too: Not as bad as the fully nonstationarity-

Table 4. Frequency of rejection for two-sided t_{2ELW} at level α

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
$d_0 \backslash n$	200	500	1000	200	500	1000	200	500	1000
0.25	1.88	1.56	1.36	5.98	6.36	5.66	11.16	10.94	10.82
0.3	3.24	1.30	1.24	6.86	5.58	6.00	11.52	10.44	11.08
0.35	3.14	2.10	1.40	9.82	6.10	5.74	14.12	11.48	11.34
0.4	2.34	3.02	2.52	11.54	8.60	7.20	18.98	13.84	11.78
0.45	1.70	2.22	2.82	8.32	11.18	9.40	17.18	18.40	15.18
0.5	1.88	1.32	1.90	6.46	8.78	10.00	13.46	17.88	18.26
0.55	1.80	1.32	1.06	6.04	5.40	6.54	10.98	12.02	13.68
0.6	1.06	1.30	1.12	5.00	5.62	4.86	9.78	9.84	9.98
0.65	1.50	1.08	0.68	5.66	4.42	3.38	10.28	8.22	6.80
0.7	1.50	1.08	0.94	5.10	4.28	4.30	8.94	8.64	8.68
0.75	1.16	1.06	1.36	4.22	4.30	5.50	7.70	8.70	10.56
0.8	1.10	0.86	1.58	4.62	5.32	6.14	9.22	10.86	12.18
0.85	0.90	1.36	1.70	4.40	6.66	6.02	10.12	12.28	10.74
0.9	1.24	1.50	1.16	6.12	5.50	5.72	12.70	10.54	10.92
0.95	1.56	1.64	1.64	6.72	5.98	5.72	12.00	10.72	11.04
1	1.96	1.58	1.26	6.66	6.24	5.28	11.58	11.46	10.12

extended versions in their worst cases, but sometimes worse than the fully nonstationarity-extended versions in their best cases. The most reliable inference of these three seems to arise from the (two-step) exact local Whittle estimator. In particular, we find for these semiparametric estimators a superior approximation of the limiting law when replacing the bandwith m by $\sum_{j=1}^{m}(\ln j - m^{-1}\sum_{\ell=1}^{m}\ln \ell)^2$. Generally, however, confidence intervals and rejections of hypotheses in the parameter range of interest considered here have to be interpreted with some care. Things are different for the sixth estimator, tapered local Whittle. Here, the size performance is satisfactory in large samples, however, it comes at a price of considerable variance inflation.

Table 5. Frequency of rejection for two-sided $2\sqrt{\frac{m}{3\Phi_3}}\left(\widehat{d}^{(3)}_{TLW} - d_0\right)$ at level α

	$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 10\%$		
$d_0 \backslash n$	200	500	1000	200	500	1000	200	500	1000
0.25	3.80	1.92	0.88	10.14	6.46	4.64	16.34	11.22	9.84
0.3	3.50	1.88	0.90	10.20	6.58	4.50	17.10	12.06	9.28
0.35	3.64	1.62	1.04	10.96	6.40	5.48	17.18	11.36	10.48
0.4	3.36	2.30	0.96	9.80	6.50	4.46	15.32	11.86	9.32
0.45	3.76	1.94	1.32	10.62	6.58	5.08	16.44	12.20	10.22
0.5	3.88	1.70	1.12	10.72	6.56	5.56	17.30	11.82	9.90
0.55	3.20	1.68	1.28	9.44	6.52	5.18	16.64	11.82	10.36
0.6	3.34	1.72	0.94	10.22	6.40	4.74	16.48	11.50	9.44
0.65	3.40	1.62	1.26	10.06	6.12	5.50	16.38	11.16	10.58
0.7	3.44	1.36	1.04	10.38	5.58	4.76	16.28	11.14	9.54
0.75	3.12	1.76	1.20	9.90	6.34	5.34	15.94	11.66	10.04
0.8	3.40	1.90	1.02	10.10	6.34	4.68	16.76	11.74	9.48
0.85	3.36	1.96	1.08	10.34	6.74	4.92	17.10	11.68	9.98
0.9	3.80	1.66	1.02	10.54	5.90	5.20	17.32	10.86	10.56
0.95	3.24	1.62	1.08	10.50	5.86	5.44	17.30	11.66	10.32
1	3.34	1.50	1.04	10.58	6.10	5.10	16.60	11.74	10.04

References

Abadir, K. M., W. Distaso, and L. Giraitis (2007). Nonstationarity-extended local Whittle estimation. *Journal of Econometrics 141*, 1353–1384.

Abadir, K. M., W. Distaso, and L. Giraitis (2011). An I(d) model with trend and cycles. *Journal of Econometrics 163*, 186–199.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Box-Steffensmeier, J. M. and A. R. Tomlinson (2000). Fractional integration methods in political science. *Electoral Studies 19*, 63–76.

Cheung, Y. L. and U. Hassler (2018). Whittle type estimation under long

memory and nonstationarity. Technical report, Social Science Research Network.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics 17*, 1749–1766.

Faÿ, G., E. Moulines, F. Roueff, and M. S. Taqqu (2009). Estimators of long memory: Fourier versus wavelets. *Journal of Econometrics 151*, 159–177.

Fox, R. and M. S. Taqqu (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics 14*, 517–532.

Giraitis, L. and D. Surgailis (1990). A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probability Theory and Related Fields 86*, 87–104.

Hassler, U. (2019). *Time Series Analysis with Long Memory in View*. Wiley.

Hauser, M. A. (1999). Maximum likelihood estimators for ARMA and ARFIMA models: A Monte Carlo study. *Journal of Statistical Planning and Inference 80*, 229–255.

Marinucci, D. and P. M. Robinson (1999). Alternative forms of fractional Brownian motion. *Journal of Statistical Planning and Inference 80*, 111–122.

Montanari, A. (2003). Long-range dependence in hydrology. In P. Doukhan, G. Oppenheim, and M. S. Taqqu (Eds.), *Theory and Applications of Long-range Dependence*, pp. 461–472. Birkhäuser.

Phillips, P. C. B. (1999). Discrete Fourier transforms of fractional processes. Cowles Foundation Discussion Papers 1243, Cowles Foundation for Research in Economics, Yale University.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics 23*, 1630–1661.

Shao, X. (2010). Nonstationarity-extended Whittle estimation. *Econometric Theory 26*, 1060–1087.

Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and trend. *Econometric Theory 26*, 501–540.

Shimotsu, K. and P. C. B. Phillips (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics 33*, 1890–1933.

Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis 20*, 87–127.

Velasco, C. and P. M. Robinson (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association 95*, 1229–1243.

Žurbenko, I. G. (1979). On the efficiency of estimates of a spectral density. *Scandinavian Journal of Statistics 6*, 49–56.

An overall seasonality test based on recursive feature elimination in conditional random forests

Karsten Webel[1] and Daniel Ollech[1]

Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt, Germany
{karsten.webel,daniel.ollech}@bundesbank.de

Abstract. To check an observed time series for presence of seasonality, different statistical tests tailored to particular manifestations of seasonality have been developed over recent decades. We suggest a generic scheme to construct an overall seasonality test from a given set of candidate tests. The main building blocks are (1) a recursive feature elimination algorithm in conditional random forests to identify the most informative candidate tests and (2) a final classification rule derived from a single pruned conditional inference tree based on the identified tests. Using a broad set of simulated ARIMA processes that is representative of the Bundesbank's time series database, we identify the modified QS and Friedman tests as being most informative when computed on the residuals obtained from fitting a non-seasonal ARIMA model to the original input series. The resulting classification rule has low type I and type II misclassification rates whilst avoiding excessive complexity.

Keywords: Conditional inference trees · Correlated predictors · JDemetra+ · Simulation study · Supervised machine learning.

1 Motivation

Seasonality can be defined broadly as those movements of a time series that recur with similar intensity in the same period each year. Owing to the different ways seasonal behaviour can manifest itself in observed time series, this definition leaves some scope for interpretation and, accordingly, there exists a variety of seasonality tests tailored to particular manifestations. For example, the F-test on seasonal dummies checks for the presence of a time-constant seasonal pattern, whereas the modified QS test checks for significant autocorrelation at seasonal lags, allowing for a time-varying seasonal pattern.

Different seasonality tests often lead to consistent outcomes when being applied to the same time series, but occasional exceptions do exist. Fig. 1 shows three macroeconomic time series for Germany over the most recent complete six years: retail trade turnover of games and toys, the harmonised index of consumer prices (HICP) of tobacco and the consumer price index (CPI) of energy. Table 1 reports the verbal comments of the six seasonality tests implemented in JDemetra+ (JD+) regarding presence of seasonality in the entire series. The tests agree that the turnover series is seasonal, which is no surprise given the

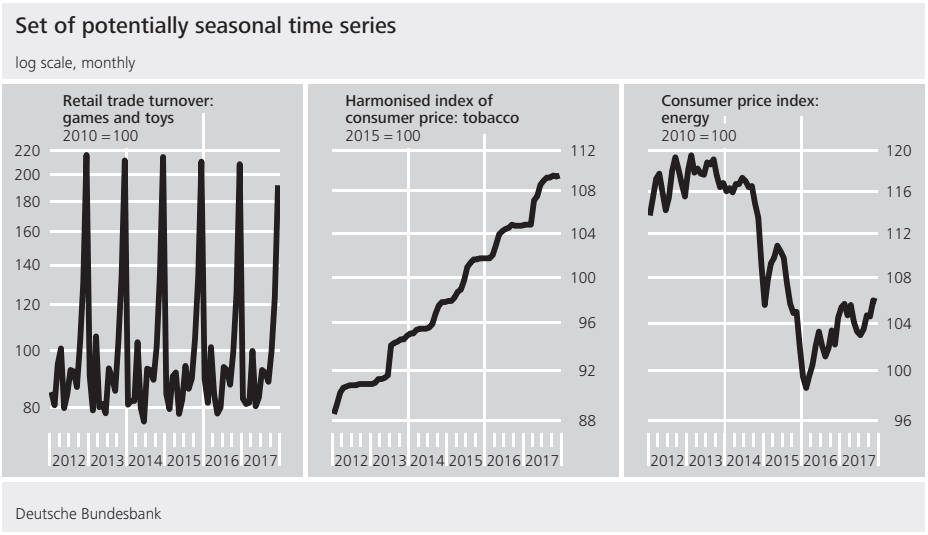


Fig. 1. Three macroeconomic time series for Germany.

distinct intra-year pattern with a prominent spike in December related to the Christmas business. The tests also agree that the HICP series is not seasonal, which again is no surprise given the dominant long-term upward movement of tobacco prices driven by occasional level shifts (in different months) related to VAT increases. However, the tests disagree for the CPI series, which may be explained by the energy prices' volatile nature that hampers the identification of recurring movements.

In case of consistent results, we do not necessarily need the entire set of tests to reach a final decision about the seasonal nature of the time series in question. In case of inconsistent results, we cannot reach a final decision as straightforwardly as in the previous case as we first need to combine the individual test results appropriately. Either way, we aim at reducing the dimension of a classification problem by identifying strong predictors from a whole set, which in general can be done with the aid of random forests [9]. However, standard clas-

Table 1. Comments on presence of seasonality in the time series shown in Fig. 1.

JD+ seasonality test	Retail trade turnover: games and toys	HICP: tobacco	CPI: energy
Modified QS test	Present	Not present	Perhaps present
Friedman test	Present	Not present	Perhaps present
Kruskal-Wallis test	Present	Not present	Present
Test for seasonal peaks	Present	Not present	Not present
Periodogram test	Present	Not present	Present
F-test on seasonal dummies	Present	Not present	Present

sification trees are likely to give biased variable importance measures (VIMs) in case of correlated predictors, such as the outcomes of different seasonality tests. Conditional inference trees are suitable in this situation [9] and we use them along a recursive feature elimination (RFE) path to identify the most informative seasonality tests from a large set of candidate tests. To this end, we simulate a vast set of ARIMA processes that is representative of the Bundesbank's macroeconomic time series database. Eventually, we construct an overall seasonality test based upon the identified tests.

The remainder of this paper is organised as follows. Section 2 provides basic information on random forests, conditional inference trees and VIMs. Section 3 describes the candidate seasonality tests and the overall test, including the RFE scheme. Section 4 reports the design and the results of the simulation study. Section 5 briefly illustrates the overall test using the three time series shown in Fig. 1. Finally, Section 6 concludes.

2 Conditional random forests

Conditional random forests (CRF) differ from classical random forests [1] in two respects: variable selection and VIMs. Let $\mathcal{L} = (\mathbf{XY})$ be the training data, where $\mathbf{X} = (\mathbf{X}_1 \dots \mathbf{X}_p)$ is a set of p predictors with $\mathbf{X}_j = (x_{1j}, \dots, x_{Nj})^\top$ for all $j \in \{1, \dots, p\}$ and $\mathbf{Y} = (y_1, \dots, y_N)^\top$ is a vector of categorical responses with $y_i \in \{1, \dots, K\}$ for all $i \in \{1, \dots, N\}$. For $b \in \{1, \dots, B\}$, a bootstrap sample $\mathcal{L}_b$ is drawn with replacement from $\mathcal{L}$, and an unpruned decision tree $\mathcal{T}_b$ is grown for the sample within the conditional inference framework [7]. As opposed to the classical approach, the binary split of any terminal node is created according to the following generic algorithm:

1. Test the global null hypothesis of independence between $\mathbf{Y}$ and each $\mathbf{X}_j$ given the node's case weights. Stop if this hypothesis cannot be rejected. Otherwise, find the predictor $\mathbf{X}_{j*}$ with the strongest association to $\mathbf{Y}$ by means of standardised linear statistics within the permutation test framework [13].
2. Find the optimal binary split of the node using $\mathbf{X}_{j*}$ and pre-specified splitting criteria, such as two-sample linear statistics [7], that are in line with the criteria applied in Step 1. Calculate the daughter nodes' case weights.

Tree growing is stopped if a pre-specified minimum number of observations in terminal nodes, n_{min}, or the stop criterion mentioned in Step 1 of the generic algorithm is reached. Finally, the forest classification is given by the unweighted majority vote of the tree classifications.

As for the classical approach, subsets of $\mathcal{L}$ can be used for validation purposes. Let $\mathcal{O}_b = \mathcal{L} \setminus \mathcal{L}_b$ be the "out-of-bag" (OOB) data of the b-th bootstrap sample, that is the training data not selected in $\mathcal{L}_b$. The CRF's performance can then be judged by means of misclassification rates in the OOB samples. Alternatively, external validation (VAL) data can be considered.

The misclassification rates in the OOB samples can also be used to evaluate the importance of any predictor $\mathbf{X}_j$ in terms of its contribution to the forest classification. The basic idea is to permute the values of $\mathbf{X}_j$ in the OOB samples in

order to mimic absence of $\mathbf{X}_j$. Then, the difference between the misclassification rates before and after permutation can be seen as the mean decrease in accuracy of $\mathbf{X}_j$. However, this VIM is likely to be biased if a purely random permutation scheme is applied in the presence of correlated predictors. The conditional permutation scheme [14, 15] prevents ex ante the overestimation of seemingly influential predictors $\mathbf{X}_j$ that in fact are not strongly associated with $\mathbf{Y}$ but appear as such due to a high positive correlation with a truly influential predictor. To this end, the values of $\mathbf{X}_j$ are randomly permuted only within subgroups of $\mathbf{X}_j^c = (\mathbf{X}_1 \dots \mathbf{X}_{j-1}\mathbf{X}_{j+1} \dots \mathbf{X}_p)$, resulting in the conditional permutation scheme $\pi(\cdot)|\mathbf{X}_j^c$. Let $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)|\mathbf{X}_j^c})$ denote the predicted classes of y_i obtained from $\mathcal{T}_b$ before and after conditional permutation of the values of $\mathbf{X}_j$ in $\mathcal{O}_b$. The respective VIM is then given by

$$\mathrm{VI}\left(\mathbf{X}_j\right) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{O}_b}\left[\frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)|\mathbf{x}_j^c})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)\}}}{|\mathcal{O}_b|}\right], \tag{1}$$

where for each tree $\mathcal{T}_b$ the permutation grid for $\mathbf{X}_j$ is defined by the cut-points of $\mathbf{X}_j^c$ in $\mathcal{T}_b$.

3 Seasonality tests

Let $\{z_t\}$ denote a weakly stationary time series of length T with τ observations per year. Whenever appropriate, the notation will be changed to $\{z_{ij}\}$, where z_{ij} is the observation of the i-th period, $i \in \{1, \dots, \tau\}$, in the j-th year. Furthermore, let $(pdq)(PDQ)$ abbreviate the ARIMA model

$$\phi_p(B)\,\Phi_P(B^\tau)\,\nabla_1^d\,\nabla_\tau^D x_t = \theta_q(B)\,\Theta_Q(B^\tau)\,\varepsilon_t,$$

where B is the Backshift operator, $B^k x_t = x_{t-k}$, $\nabla_k = 1 - B^k$, ϕ_p and Φ_P are the non-seasonal and seasonal autoregressive operators, θ_q and Θ_Q are the non-seasonal and seasonal moving average operators, and $\{\varepsilon_t\}$ is white noise with zero mean and finite variance. Then, a given ARIMA model is said to be non-seasonal (N-S) if $(PDQ) = (000)$ and seasonal (S) in any other case.

3.1 Candidate tests

We consider three branches of seasonality tests, each of which states absence of seasonality in $\{z_t\}$ as the null hypothesis:

- The six seasonality tests implemented in JD+ and considered in Table 1 have been described in detail by [9], along with respective intuitions. We consider five of them: the modified QS test (QS), the Friedman test (FT), the Kruskal-Wallis test (KW), the periodogram test (PD) and the F-test on seasonal dummies (SD).
- We also consider selected tests from the literature on medical statistics, which are primarily used for detection of periodic occurrences of diseases, and a seasonal unit root test. Basic information on these tests is provided below.

– The following tests are also applied to the residuals obtained from fitting a non-seasonal ARIMA model to the unadjusted data, where (pdq) is determined by the automatic model identification routine [8] with $p \leq 3$ and $q \leq 3$: the QS, FT, KW, Roger, Welch and Welch-Kruskal-Wallis tests. The short form notation of each of these tests is the short form notation of the respective original test plus the suffix "-R".

Welch test. The Welch (WE) test [16] is essentially a one-way ANOVA without repeated measures for heteroskedastic data. Let $\mu_i = \mathbb{E}(z_{ij})$, n_i the number of observations in the i-th period, $\bar{z}_i = n_i^{-1} \sum_{j=1}^{n_i} z_{ij}$ and $s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2$. Then, the null hypothesis is given by $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$ and the test statistic is

$$WE = \frac{(\tau - 1)^{-1} \sum_{i=1}^{\tau} w_i \left(\bar{z}_i - w^{-1} \sum_{i=1}^{\tau} w_i \, \bar{z}_i\right)^2}{1 + 2\,(\tau + 2)(\tau^2 - 1)^{-1} \sum_{i=1}^{\tau} (n_i - 1)^{-1} \left(1 - w^{-1} w_i\right)^2}, \tag{2}$$

where $w_i = n_i \,/ s_i^2$ and $w = \sum_{i=1}^{\tau} w_i$. Under H_0, (2) follows an F-distribution with f_1 and f_2 degrees of freedom, which are given by

$$f_1 = \tau - 1 \quad \text{and} \quad f_2 = \left[3\,(\tau^2 - 1)^{-1} \sum_{i=1}^{\tau} (n_i - 1)^{-1} \left(1 - \frac{w_i}{w}\right)^2 \right]^{-1}.$$

In addition to the original WE test, we also consider a variant based on the ranks of $\{z_{ij}\}$. Since the ranks are assigned as for the KW test, that is over the entire observation period, this variant is referred to as the Welch-Kruskal-Wallis (WEKW) test.

Edwards test. The Edwards (ED) test [3, 4] assumes that seasonality can be described by harmonic patterns. The null hypothesis is the same as for the WE test. Letting $n_i = n$ for each $i \in \{1, \ldots, \tau\}$, the test statistic is given by

$$ED = 8 \sum_{i=1}^{\tau} \tilde{z}_i^2 \cdot \frac{\left[\sum_{i=1}^{\tau} \tilde{z}_i \, \sin\left(2\pi i/\tau\right)\right]^2 + \left[\sum_{i=1}^{\tau} \tilde{z}_i \, \cos\left(2\pi i/\tau\right)\right]^2}{\left(\sum_{i=1}^{\tau} \tilde{z}_i\right)^2}, \tag{3}$$

where $\tilde{z}_i = \sqrt{\sum_{j=1}^{n} z_{ij}}$. Under H_0, (3) asymptotically follows a χ^2-distribution with two degrees of freedom. However, note that $\tilde{z}_i$ and, thus, the test statistic can be calculated only if $\sum_{j=1}^{n} z_{ij} > 0$ for the i-th period.

Roger test. The Roger (RO) test [12] is an advanced version of the ED test. Again, H_0 is specified as for the WE test and the test statistic is

$$RO = 2 \cdot \frac{\left[\sum_{i=1}^{\tau} z_i^\star \, \sin\left(2\pi i/\tau\right)\right]^2 + \left[\sum_{i=1}^{\tau} z_i^\star \, \cos\left(2\pi i/\tau\right)\right]^2}{\sum_{i=1}^{\tau} z_i^\star}, \tag{4}$$

where $z_i^\star = \sum_{j=1}^{n} z_{ij}$. Under H_0, (4) approximately follows a χ^2-distribution with two degrees of freedom. Note that $z_i^\star = \tilde{z}_i^2$ and, thus, the test statistic can be calculated without any restriction on the observations.

OCSB test. The OCSB test [10] is a seasonal unit root test. Dropping the stationarity assumption on $\{z_t\}$, we use the following representation [11]:

$$\nabla_1 \nabla_\tau z_t = \mu_t + \beta_1 \nabla_\tau z_{t-1} + \beta_2 \nabla_1 z_{t-\tau} + \sum_{j=1}^{J} \phi_j \nabla_1 \nabla_\tau z_{t-j} + \varepsilon_t, \qquad (5)$$

where $\{\varepsilon_t\}$ is white noise with zero mean and finite variance. Assuming validity of ∇_1, the null hypothesis reads $\beta_1 = 0$ and is tested using the standard t-test in the regression (5) for a pre-specified order J.

3.2 Construction of overall test

Bearing in mind potential general conflicts between model parsimony and accuracy, we aim at condensing a set of candidate seasonality tests into an overall test that balances interpretable results against a low overall misclassification rate. Our approach essentially comprises two steps:

- the identification of the most informative candidate tests based on misclassification rates and VIMs in a CRF framework along an RFE path and
- the derivation of a classification rule based on a pruned single conditional inference tree which considers only the most informative candidate tests.

Identification of most informative candidate tests. We use the following generic RFE algorithm to identify the most informative seasonality tests in a set of candidate tests:

1. Start with the entire set of candidate tests.
2. Train multiple CRFs on independent training data sets using the p-values of the remaining candidate tests as predictors and the seasonal nature of simulated ARIMA processes as categorical response.
3. For each candidate test, aggregate the VIM (1) over all CRFs. Store the mean and standard deviation of the overall misclassification rates computed for each CRF's external VAL data.
4. Eliminate the test with the smallest aggregated VIM from the set of candidate tests.
5. Repeat Steps 2 to 4 until a pre-defined minimum number of predictors, p_{min}, is reached.
6. Select the most informative tests taking into account the overall misclassification rates along the entire RFE path and the numbers of tests included.

Remark 1. This RFE algorithm differs from similar approaches [2, 5, 6] in three respects: first, we consider conditional instead of classical random forests (Step 2); second, we consider VIMs aggregated over multiple forests (Step 3) in order to stabilise the feature elimination in each recursion of Steps 2 to 4; third, we evaluate misclassification rates over external VAL instead of OOB data (Step 3).

Derivation of classification rule. Once the most informative seasonality tests have been identified in Step 6 of the generic RFE algorithm, their p-values are used again as predictors to grow a single conditional inference tree. Depending on the specified parameters, such as the number of identified tests, this tree is more or less likely to contain some redundancies in the sense that all branches after a particular node lead to the same classification. In this case, the single conditional inference tree is pruned to obtain a simplified but equivalent decision tree which always yields the same classification as the unpruned tree.

4 Simulation study

4.1 Design

We aim at simulating a set of seasonal and non-seasonal ARIMA models that closely mirrors the macroeconomic monthly time series stored in the Bundesbank's database in terms of their distributional properties. To this end, we first identify representative ARIMA models. Then, we use the simulation algorithm [9] that takes into account the shares of the identified ARIMA models and the multivariate distribution of the parameters within each identified ARIMA model class. Finally, we apply the seasonality tests described in Section 3.1 to the simulated ARIMA models and construct the overall test as outlined in Section 3.2.

ARIMA model identification. We randomly sample without replacement 3,300 macroeconomic time series that are seasonally adjusted each month and 10,600 monthly time series that are not seasonally adjusted at all from the database. For each sampled series, we run the automatic ARIMA model identification routines of the regARIMA and TRAMO pre-processors in JD+. For each identified model m of class $k \in \{\text{N-S}, \text{S}\}$, we calculate the model's regARIMA/TRAMO averaged share p_{mk} among all models in class k and the simulation weights $w_{mk} = \tilde{p}_{mk} / \sum_j \tilde{p}_{jk}$, where $\tilde{p}_{mk} = p_{mk} \cdot \mathbb{1}_{\{p_{mk} \geq 0.01\}}$.

ARIMA model simulation. For each identified model m of class k and each length $N \in \{60, 120, 240\}$, we simulate $100{,}000 \cdot w_{mk}$ ARIMA time series, using the same algorithm and parameter setup as in [9] and the 'gsarima' package under R 3.3.3. To this end, the length of the burn-in period for the simulation depends on the number of parameters in the AR(∞) representation of the model. Eventually, this yields a representative set of 600,000 simulated ARIMA models.

Candidate tests, RFE and CRF settings. We use a total of 18 seasonality tests as we consider three versions of the OCSB test: the OCSB1 test uses $J = 3$ in (5), the OCSB2 test essentially takes the sum over $j \in \{1, 2, 3, 12, 13\}$ in (5) and the OCSB3 test is the OCSB2 test with $\{\varepsilon_t\}$ being a seasonal MA(1) process instead of white noise. In general, the stationarity assumption of the candidate tests, if applicable, is met by differencing the input series once as in JD+.

We run the RFE algorithm with $p_{min} = 1$. In each recursion of Steps 2 to 4, 50 independent training data sets are randomly selected from the 600,000 simulated ARIMA models, where the size depends on the current number of candidate tests, varying between 800 and 8,000 due to computational restrictions. For each training data set, a CRF of $B = 100$ trees is grown with $n_{min} = 1$. In Step 3, the mean and median VIMs (1) and a random sample of 50,000 ARIMA series not selected in the respective CRF are considered as aggregated VIM and external VAL data, respectively.

Regarding the overall test, we grow the single conditional inference tree on a training data set of 50,000 simulated ARIMA time series with $n_{min} = 250$.

4.2 Results

Misclassification rates. Table 2 reports the misclassification rates of the candidate seasonality tests and CRFs[1] at the initial recursion of the RFE algorithm, allowing the following general statements:

[1] The rates are based on the entire set of simulated ARIMA models for the candidate tests and averaged over the respective forests for the OOB and VAL data.

Table 2. Misclassification rates of candidate seasonality tests and initial CRFs as a percentage (N-S = non-seasonal series, S = seasonal series).

		$\alpha = 0.01$								$\alpha = 0.05$							
		Simulated ARIMA time series															
		All		5-year		10-year		20-year		All		5-year		10-year		20-year	
Classifier		N-S	S	N-S	S	N-S	S	N-S	S	N-S	S	N-S	S	N-S	S	N-S	S
CRF	OOB	0.6	1.9	0.6	1.9	0.5	2.0	0.7	1.9	0.6	1.9	0.6	1.9	0.5	2.0	0.7	1.9
	VAL	0.6	1.9	0.6	2.0	0.5	1.8	0.6	1.8	0.6	1.9	0.6	2.0	0.5	1.8	0.6	1.8
QS		4.9	1.5	2.5	1.7	5.0	1.4	7.1	1.3	7.4	1.2	4.9	1.4	7.5	1.1	9.8	1.2
QS-R		0.3	8.4	0.2	7.5	0.2	7.8	0.6	9.8	1.1	7.0	0.9	6.1	0.9	6.4	1.4	8.3
FT		2.1	2.1	1.5	2.2	2.3	1.9	2.4	2.1	6.6	1.6	5.7	1.6	7.0	1.5	7.2	1.7
FT-R		0.8	2.1	0.4	2.3	0.8	2.0	1.4	2.1	4.2	1.6	3.2	1.6	4.2	1.5	5.2	1.7
KW		2.4	3.8	1.9	3.9	2.6	3.7	2.7	3.8	6.9	3.2	6.2	3.1	7.1	3.1	7.4	3.3
KW-R		0.7	2.1	0.3	2.2	0.7	2.0	1.1	2.2	4.0	1.6	3.1	1.6	4.1	1.5	4.7	1.8
PD		3.2	3.6	3.2	3.4	3.3	3.6	3.2	3.9	8.1	3.2	8.3	2.8	8.2	3.1	8.0	3.5
SD		4.0	2.7	4.4	2.5	4.1	2.7	3.7	2.9	9.1	2.2	9.7	2.0	9.2	2.2	8.5	2.5
RO		11.8	93.9	10.1	93.7	11.6	94.0	13.7	94.0	15.1	91.0	13.5	90.7	14.9	91.1	16.8	91.3
RO-R		16.7	78.7	15.6	82.3	17.2	77.6	17.2	76.4	19.9	71.7	18.8	74.9	20.6	70.6	20.4	69.8
ED		4.2	99.2	3.3	99.1	4.1	99.3	5.2	99.4	4.9	99.1	3.9	98.9	4.8	99.1	6.1	99.3
WE		3.5	3.8	4.3	3.7	3.2	3.7	3.0	4.0	8.9	3.3	10.8	3.0	8.2	3.2	7.7	3.6
WE-R		1.7	2.1	2.7	2.0	1.5	2.1	1.1	2.4	6.4	1.7	8.6	1.4	5.7	1.6	4.9	2.0
WEKW		5.6	3.4	9.0	3.1	4.5	3.4	3.4	3.7	12.0	2.9	17.5	2.7	10.1	2.9	8.3	3.2
WEKW-R		4.1	1.8	7.6	1.5	2.9	1.8	1.7	2.0	10.4	1.4	16.7	1.1	8.3	1.4	6.1	1.6
OCSB1		4.9	4.3	7.8	3.8	3.7	4.5	3.2	4.6	2.9	7.1	4.0	7.0	2.4	7.4	2.3	6.7
OCSB2		3.3	4.9	8.4	3.1	1.0	5.0	0.6	6.6	0.8	10.0	1.7	6.9	0.4	9.6	0.4	13.4
OCSB3		10.5	4.6	11.9	3.2	8.6	4.7	11.0	5.9	3.5	8.5	3.6	7.0	3.1	8.6	4.0	9.8

- The "-R" tests tend to have lower type I and type II misclassification rates than their standard implementations, except for the QS and RO tests.
- The ED and RO tests display unacceptably high misclassification rates, especially for seasonal series.[2] For that reason, they will not be considered for the RFE algorithm and construction of the overall seasonality test.
- For the OCSB tests, the order of augmentation strongly affects the misclassification rates: the OCSB1 test tends to have higher type I but lower type II misclassification rates than the OCSB2 test, especially for longer series; the OCSB3 test does not outperform either of them in general.
- The overall accuracy tends to be higher for $\alpha = 0.01$ mainly because the type I misclassification rates are noticeably lower while the type II misclassification rates are barely higher than for $\alpha = 0.05$.

RFE path. Table 3 reports the results of the RFE algorithm. In general, considering the mean or median VIMs (1) barely makes a difference as virtually all candidate test are eliminated in the same or adjacent recursions, except for the KW-R and FT-R tests which are eliminated in recursions 10 and 15, respectively. However, the interchangeability of these two tests does not surprise given their sensitivity to similar manifestations of seasonality and their almost identical type I and type II misclassification rates reported in Table 2.

Regardless of the aggregated VIM considered, the misclassification rate, averaged over external VAL data of 50 CRFs, gradually increases from 1.12% to 1.37% during RFE recursions 1 to 11, then decreases less gradually back to the

[2] Originally, both tests have been tailored to incidence data in epidemiology and are not scale invariant as they depend strongly on the level of the input series.

Table 3. Recursive elimination of candidate seasonality tests.

	Mean variable importance			Median variable importance		
RFE	Eliminated	Misclassification rate		Eliminated	Misclassification rate	
recursion	test	Average (%)	Std. dev.	test	Average (%)	Std. dev.
1	WE	1.12	0.087	WEKW	1.12	0.074
2	WEKW	1.12	0.070	WE	1.12	0.078
3	KW	1.14	0.077	PD	1.13	0.074
4	PD	1.14	0.067	KW	1.12	0.074
5	SD	1.14	0.078	SD	1.12	0.082
6	OCSB2	1.14	0.073	WEKW-R	1.13	0.080
7	WEKW-R	1.15	0.065	OCSB2	1.14	0.091
8	OCSB3	1.17	0.067	OCSB3	1.15	0.080
9	FT	1.19	0.083	FT	1.18	0.092
10	KW-R	1.20	0.067	FT-R	1.20	0.089
11	WE-R	1.37	0.116	WE-R	1.37	0.108
12	OCSB1	1.30	0.072	QS	1.30	0.092
13	QS	1.18	0.089	OCSB1	1.29	0.088
14	QS-R	1.12	0.043	QS-R	1.12	0.043
15	FT-R	1.40	0.039	KW-R	1.36	0.043

initial 1.12% during recursions 12 to 14 and, finally, soars to 1.40% and 1.36%, respectively, in the final recursion where only one candidate test is left. Taking additionally the respective standard deviations into account, a set of two seasonality tests seems to provide a convenient balance between accuracy and parsimony and, thus, to be a good starting point for deriving the classification rule of the overall test.

Classification rule. Sticking to the RFE path based on the mean VIM, which seems to be the more natural choice, the QS-R and FT-R tests are identified as the most informative seasonality tests. Fig. 2 shows the single unpruned conditional inference tree based upon these two tests and the classification of the training data, where all thresholds have been rounded to three decimal places. The tree actually consists of 12 nodes but due to some redundancies only three of them are relevant for the final classification. To see this, note, for example, that any series in the left daughter node of the initial node is classified as seasonal if p-value ≤ 0.011 holds for the QS-R test, regardless of all further branches. A

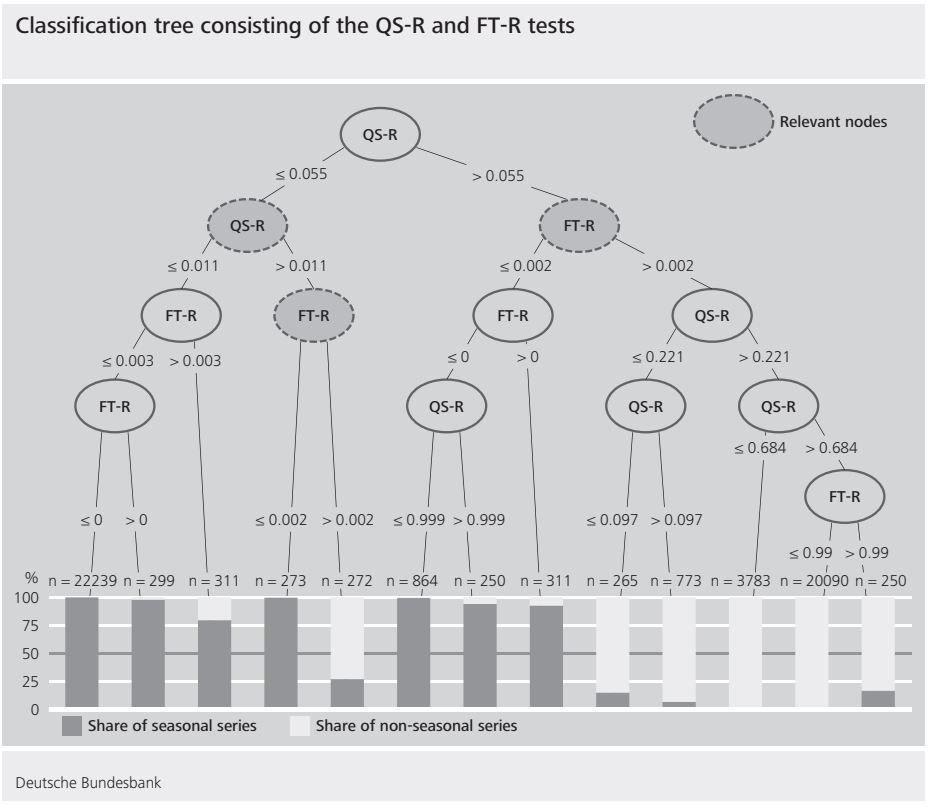


Fig. 2. Unpruned conditional inference tree based upon the QS-R and FT-R tests.

similar rule applies to the right daughter node of the initial node which classifies any series as seasonal (non-seasonal) if the p-value of the FT-R test is smaller (larger) than 0.002. Interestingly, the latter rule also holds for the right daughter node of the "QS-R" node mentioned before, making the entire right set of branches of the initial node virtually redundant.

Pruning the single conditional inference tree shown in Fig. 2, the equivalent classification rule of the overall seasonality test reads: a time series is classified as seasonal if

1. p-value ≤ 0.011 for the QS-R test OR
2. p-value > 0.011 for the QS-R test AND p-value ≤ 0.002 for the FT-R test.

The type I and type II misclassification rates of this overall test are given by 0.54% and 1.56%, respectively, for the training data and by 0.56% and 1.57%, respectively, for the VAL data, that is the simulated ARIMA processes not used for growing the single conditional inference tree.

Remark 2. This overall test is sensitive to both stable and moving seasonal pattern since the ANOVA-type rank test mainly covers fixed seasonality, while the QS-R test allows for a higher degree of time-varying seasonality.

Remark 3. Sticking to the RFE path based on the median VIM yields the same classification rule of the overall test with the KW-R test replacing the FT-R test and misclassification rates that are practically indistinguishably from the ones reported above.

5 Application

We now apply the overall test to the three macroeconomic time series displayed in Fig. 1. Table 4 reports the test's ingredients and final classifications. Unsurprisingly, the overall test confirms the common agreement of the six JD+ seasonality tests reported in Table 1 that the turnover series is seasonal and the HICP series is not. More interestingly, the CPI series, which was given disagreeing comments from the JD+ seasonality tests, is classified as non-seasonal since the p-values of the QS-R and FT-R tests are clearly higher than the respective thresholds of 0.011 and 0.002. Thus, the overall test even overrules the majority vote of the six JD+ seasonality tests.

Table 4. Overall seasonality test for the time series shown in Fig. 1.

	p-value		
Time series	QS-R test	FT-R test	Classification
Retail trade turnover: games and toys	0.000	0.000	Seasonal
HICP: tobacco	0.969	0.740	Non-seasonal
CPI: energy	0.038	0.017	Non-seasonal

6 Summary

We proposed a generic scheme to construct an overall seasonality test from a set of candidate tests with two building blocks: (1) recursive feature elimination in conditional random forests to identify the most informative candidate tests and (2) a pruned conditional inference tree based on the latter tests for the final classification. Using 600,000 simulated ARIMA processes that are representative of the Bundesbank's time series database and 18 candidate tests, we identified residual-based variants of the QS and Friedman tests as being most informative. The resulting classification rule is highly accurate and not overly complex.

References

1. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)
2. Díaz-Uriarte, R., de Andrés, S. A.: Gene Selection and Classification of Microarray Data Using Random Forest. BMC Bioinformatics **7**, Article 3 (2006)
3. Edwards, J. H.: Congenital Malformations of the Central Nervous System in Scotland. British Journal of Preventive & Social Medicine **12**(3), 115–130 (1958)
4. Edwards, J. H.: The Recognition and Estimation of Cyclic Trends. Annals of Human Genetics **25**(1), 83–87 (1961)
5. Genuer, R., Poggi, J.-M., Tuleau-Malot, C.: Variable Selection Using Random Forests. Pattern Recognition Letters **31**(14), 2225–2236 (2010)
6. Gregorutti, B., Michel, B., Saint-Pierre, P.: Correlation and Variable Importance in Random Forests. Statistics and Computing **27**(3), 659–678 (2017)
7. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics **15**(3), 651–674 (2006)
8. Hyndman, R. J., Khandakar, Y.: Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software **27**(3), 1–22 (2008)
9. Ollech, D., Webel, K.: Assessing the Informational Content of Seasonality Tests by Random Forests of Conditional Inference Trees. In: JSM Proceedings, Business and Economic Statistics Section, pp. 1889–1900. American Statistical Association, Alexandria, VA (2017)
10. Osborn, D. R., Chui, A. P. L., Smith, J. P., Birchenhall, C. R.: Seasonality and the Order of Integration for Consumption. Oxford Bulletin of Economics and Statistics **50**(4), 361–377 (1988)
11. Rodrigues, P. M. M., Osborn, D. R.: Performance of Seasonal Unit Root Tests for Monthly Data. Journal of Applied Statistics **26**(8), 985–1004 (1999)
12. Roger, J.: A Significance Test for Cyclic Trends in Incidence Data. Biometrika **64**(1), 152–155 (1977)
13. Strasser, H., Weber, C.: On the Asymptotic Theory of Permutation Statistics. Mathematical Methods of Statistics **8**, 220–250 (1999)
14. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A.: Conditional Variable Importance for Random Forests. BMC Bioinformatics **9**, Article 307 (2008)
15. Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T.: Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics **8**, Article 25 (2007)
16. Welch, B. L.: On the Comparison of Several Mean Values: An Alternative Approach. Biometrika **38**(3/4), 330–336 (1951)

Simulation-based selection of prediction models

Robert M. Kunst *

Institute for Advanced Studies Vienna, Josefstadter Strasse 39, 1080 Vienna,
and
University of Vienna, Oskar Morgenstern Platz, 1090 Vienna

Abstract. We target an assessment of the potential benefits of basing model selection decisions in a forecasting context on simulations that fuse data information and the structure hypothesized by tentative rival models. These procedures can be applied to any empirical forecasting problems. Our main focus is, however, on macro-economic forecasting. Our procedure aims at choosing among a small number of tentative forecast models in the presence of data. From models fitted to the data, pseudo-data are generated. Again, the models are applied to the pseudo-data and their out-of-sample performance is evaluated. The ultimate choice of the forecasting model is based on the relative performance of rival models in predicting 'their own data' and those of the rival model.

Keywords: Simulation, forecasting, time series

1 Introduction

Model selection is a standard problem of statistics (see, for example, [3]). In short, confronted with a finite sample of data, a researcher wishes to decide whether model A or model B has generated the data, assuming both A and B as parametric models with several free parameters to be estimated. For example, model A may be $X_t = a + bX_{t-1} + \varepsilon_t$, a first-order autoregression, and model B may be $X_t = a + bX_{t-1} + cX_{t-2} + \varepsilon_t$, a second-order autoregression. The target is to classify data correctly, even if the true model is B with a very small parameter value c, which makes it probable that model A is typically preferred. Very crudely, most model selection procedures are variants of restriction tests, in this case on the significance of the second-order coefficient c.

If the aim of modeling is forecasting, the specific aim modifies the problem.[1] This modification may be responsible for the often very skeptical attitude of prominent forecasters toward statistical inference, as expressed radically for instance by [2]. Forecasting is different. In the above example, even if it is known that $c \neq 0$, the forecaster may be better off by using the simpler model A as a

* I thank three anonymous referees for helpful comments.

[1] Throughout this article, the words 'forecast' and 'prediction' are used interchangeably. Both always refer to out-of-sample prediction, with parameter estimates based on a sample that does not include the observation to be predicted.

prediction tool, as the uncertainty due to the poorly estimated parameter c outweighs the disadvantage of the incorrect restriction $c = 0$, i.e. model A. Again, many forecasters may use restriction tests, but they should be aware of the fact that the significance level of a test that optimizes predictive performance is implied directly by the task of prediction, such that it may be around 10% in a small sample and 1% in a larger one. Although some results can be derived analytically, many may prefer running parametric or non-parametric simulations under the assumption that the estimated parameters are the true ones, for models A and B, and comparing the outcomes. In small samples, the model A of our example will usually predict more precisely than B, and at a specific sample size that depends on the value of c, model B will start outperforming A.

In this regard, a path-breaking result is due to [18] who shows that asymptotically the most precise prediction is performed by the model selected by minimizing the A Information Criterion (AIC) of [1], at least in autoregressive models and under certain conditions. Note that the wording 'efficient selection' used by [18] refers to the optimization of forecasting. Unfortunately, however, this is only an asymptotic result, and small-sample properties can deviate considerably (see, for example, the many carefully designed simulations reported by [15]). For this reason, forecasting research often tends to rely on simulation evidence. For example, [7] compare the performance of forecast model selection by BIC (not AIC) and by straightforward accuracy comparisons on out-of-sample forecasts. They use Monte Carlo and find a slight advantage for the information criterion. There is a catch here, however, that tends to be overlooked. In typical empirical situations, all forecast models are misspecified, that is, they are less rich than the generating structure. Then, forecast accuracy toward the end of the sample is crucial for the relative performance in the future, whereas information criteria and likelihood-based methods give equal weights to performance in the beginning and in the end of the sample.

Our procedure builds on the simulation-based method introduced above. Consider for simplicity the case of two arbitrary—i.e., not the simple autoregressive structures described above—models A and B.[2] Both are fitted to the data, and free parameters are estimated. We note that, apart from the parametric structures fitted to them, the original data are not used any more in the procedure. Then, in a second step, both estimated models are simulated and thus create artificial data, and all simulated structures are then confronted with A and B as tentative forecasting models. This yields 4 cases, A applied to A data, B applied to A data, and so on. Often, although definitely not always, models forecast their own data well, such that predicting A by A and B by B yields satisfactory results. A by B and B by A can be worse, and if A by B is still acceptable but B by A is inferior, the forecaster's choice may be choosing B as a forecast model in summary, even though the question whether data stem from A or B remains unsettled.

[2] Generalizations of the selection problem to more than two models are straightforward and will be considered in future work

A visual representation of the problem is given in Figure 1: Presume there are two parametric model classes competing to forecast data generated from the unknown X. The two model classes are represented by the two curves. In the left graph, one of them contains X, so B is 'misspecified', but the model estimated from data within A is further away from the true model X than the misspecified estimate B. If distance represents forecast accuracy, the 'wrong' choice of B is actually an advantage, as the estimate A is a lousy forecasting tool, even though it is in the generating class. Now presume none of the two entertained model classes contains X. The closest approximation among the A models may be a much better description than the closest approximation among the B models, but there is no guarantee that the same ranking persists among the models estimated using a finite sample of $N = 50$ observations drawn from X.

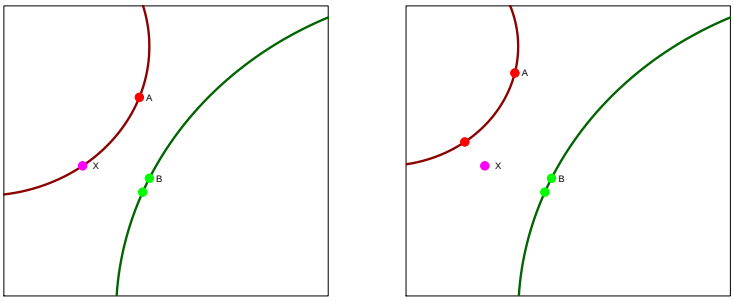


Fig. 1. Forecasting with correct specification (left) and misspecification (right)

This simulation-based forecast-model selection method was advertised by [13], and it was used by [10] as well as, for example, by [6]. The latter contribution made it clear that complex nonlinear models do not always dominate their own generated data, such that the application of the procedure is far from trivial. Paradoxical cases, when incorrect but simple models are better forecast tools than correct but complex models, are sometimes surprising and seen with suspicion. It should be obvious, however, that nonlinear generators, such as for example autoregressions with a small ARCH effect in their errors, are predicted better by pure autoregressions, unless samples are huge. For a related empirical example with a prediction paradox, see [8].

The example described by [11] is a bit simpler. He considers a monthly time series of business economics data with recognizable seasonal behavior. Seasonal dummies are jointly significant at a level of 5%. The model without seasonal dummies, however, forecasts better for all empirically relevant sample sizes. The result is confirmed by some simulations. We note that this is still not a proof

for the deficiency of all statistical method in forecasting, as [2] may opine, but rather a consequence of using the wrong significance level.

We presume that the technique is relatively widespread in the empirical literature, but we are not aware of more good sources. Some thoughts on this and similar procedures are reflected in [14]. For a recent application, see [9], where rival models are structures with and without seasonal constants or modeling in first differences versus modeling in levels. The latter application is of some interest, as classical restriction tests violate usual regularity conditions and the asymptotics become non-standard. By contrast, this is no hurdle for the forecast-model selection, as significance levels adjust automatically.

2 The simulation experiment

In experiment I, the one experiment that is reported in detail in this article, trajectories are generated from the ARMA model

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

for $t = 1, \ldots, N$ and $N = 25, 50, 100$, with a long burn-in to guarantee stationarity. The basic version uses *iid* $N(0, 1)$ noise for ε_t. In a variant, we experiment with a Cauchy error generator. The two fitted classes are the ARMA(1,1) and the AR(1) model. Model A is correctly specified, model B only for $\theta = 0$. The design is nested and thus favors traditional AIC. We report the cases $\phi \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, $\theta \in \{-0.5, 0, 0.5\}$ here.

We now briefly mention a different, classical simulation with much better resolution. In this role-model experiment, trajectories of are generated from ARMA(1,1) models over a comparatively fine grid in ϕ and θ. In Figure 2, these trajectories have length $N + 1 = 51$. The observation at $t = 51$ is then predicted out-of-sample using estimated AR(1), MA(1), and ARMA(1,1) structures estimated from the first $N = 50$ observations. The graph shows which of the three options forecasts best. Obviously, the optimal forecast model does not coincide with the true generators that are schematically presented in the right graph. There is a sizeable area where AR(1) forecasts best, even though $\theta \neq 0$, and there is a much smaller area where MA(1) forecasts best even though $\phi \neq 0$. The graph is informative insofar as it discourages MA models for forecasting in small samples, and also as it clearly shows that setting a non-zero parameter at zero, i.e. imposing an invalid restriction, assists in increasing forecast accuracy.

At this instance, two statistical procedures deserve mention, as they are similar to the suggested simulation-based choice of prediction models. Such statistical procedures may even be seen as counterparts for the aim of choosing true models to the forecasting procedures for the aim of minimizing prediction errors. These statistical procedures are twofold: confirmatory analysis and Bayes testing.

Confirmatory analysis is not popular in the science of statistics, although it is occasionally resuscitated, for example in the cointegration procedures considered by [5]. In a hypothesis testing situation with $H_0 : c = 0$ and $H_A : c \neq 0$, i.e. with classical asymmetry of hypotheses, little is learned from reversing the

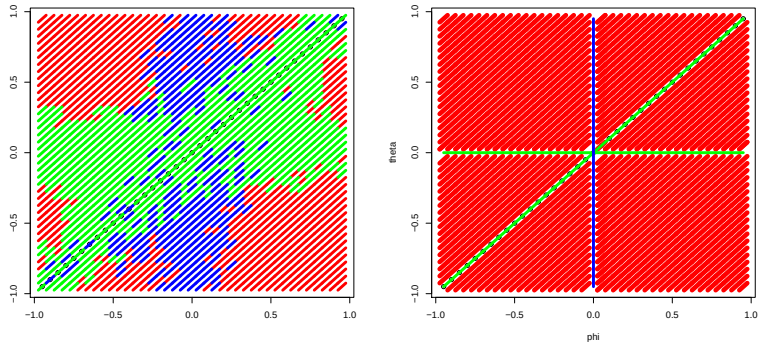


Fig. 2. The choice among AR(1), MA(1), and ARMA(1,1) models that optimizes forecasting properties. Left graph for 50 observations from an ARMA(1,1) model, right graph true models. Preference for AR(1) shown in green (light gray), for MA(1) in blue (dark gray), for ARMA(1,1) in red (black).

construction by setting $c = \hat{c}$ as a new null hypothesis, letting $\hat{c}$ denote the sample estimate of the parameter c, and choosing $c = 0$ as the new alternative. In forecasting, however, the forecast using $c = \hat{c}$ is a valid rival of the one at $c = 0$, and a comparison by direct evaluation and by simulation makes sense. In other words, arguments used against confirmatory analysis are not valid against simulation-based forecast-model selection.

Bayes testing is a standard decision procedure touched upon in all textbooks on Bayesian statistics (such as [17] or [16]). Given—typically uniform—discrete priors on hypotheses and—maybe informative—continuous priors on parameter spaces within each hypothesis, posterior distributions on hypotheses are maximized in the presence of data, and the hypothesis with maximum posterior is selected or preferred. Bayes tests have nice properties such as symmetry among hypotheses and full asymptotic consistency, as both type-I and type-II errors disappear as the sample size increases. In contrast to the increasing enthusiasm for Bayes estimation (see [12]), however, Bayes testing has not become so popular, maybe because researchers tend to feel uneasy in the presence of uniform priors on two competing hypotheses. It is cheap to state that this is the counterpart to the ubiquitous 5% significance in classical testing, rather it should be justified that both hypothesis are *a priori* (equally?) plausible.

The parallel argument may be of some relevance for our forecast selection procedure. In forecasting, a problem may come up if model A is a much better approximation to the data than model B. It may then be that A forecasts badly if B, although fitted to the data, is the generator but the researcher may nonetheless prefer A, as B is an unlikely true model. From experience, this situation is rarely encountered, but it should be carefully analyzed in corresponding simulation designs.

Within the limits of this report, we restrict attention to the experiment I outlined above. In stage one, a true model is simulated. In stage two, the simulation acts as if the true model were unknown, and models A and B are both fit to the pseudo-data. A and B are tentatively simulated and all four pairs (A/A, A/B, B/A, B/B) are evaluated. The thus recommended model (A or B) is determined. In stage three, it is evaluated whether the recommended model really optimizes forecasts in the pseudo-data. Note that M^2 replications are needed in the process, if M is to denote the number of replications in one stage, such that M must be kept in reasonable limits.

The evaluation following step two deserves attention. We consider two methods for selecting the best model. Consider the four (squared, sign-free) error series $u_s(j|k), j, k = A, B$. $u_s(A|B)$ stands for the error committed by using forecast model A on data generated from model B. The first method uses counting cases for comparison, thus we call it the *counting simulator* If $u_s(A|A)$ and $u_s(A|B)$ are smaller than $u_s(B|A)$ and $u_s(B|B)$ most of the time, use model A. Otherwise use model B. The second method uses a quantitative averaging criterion, thus we call it the *MSE simulator*. Consider the sum of $\sum_s u_s(A|A)$ and $\sum_s u_s(A|B)$. If it is smaller than the sum of $\sum_s u_s(B|A)$ and $\sum_s u_s(B|B)$, use model A, otherwise use model B. The simulation methods are compared to an AIC benchmark, a known strong method in forecasting model selection.

Figure 3 provides a visual summary of the ARMA/AR experiment. It appears that the three procedures perform more or less on a par with each other. This is good news for the simulation-based techniques, as AIC is a strong rival. It appears that the simulators even achieve some dominance for small N, lthough they appear to lose ground as N increases. Note that AIC is a strong benchmark and difficult to beat regarding forecasting accuracy. We conjecture that the simulators may show definitive advantages with misspecified frames, where tentative forecasting models do not match the generator, as AIC performs poorly in that situation. We note that also for the Cauchy generator, the simulated data are Gaussian, which is more plausible, as the forecasters are not assumed to know the generating law.

Apart from the basic ARMA/AR experiment, we also considered the following setting for an experiment II. This one is based on economic data summarized in a VAR: U.K. unemployment rate, CPI inflation, and real GDP growth. These three variables are known to be 'borderline' stationary. A VAR is fitted with AIC lag order, and the fitted model serves as a generator. In a basic version, a trivariate *iid* $N(0, \Sigma)$ noise generates the errors, with Σ estimated from the U.K. data. The forecasting experiment focuses on inflation exclusively, i.e. only one of the three generating variables is targeted. The specification with parameter values is taken from [4]. Again, univariate AR(1) and ARMA(1,1) are the rival models. Neither of the two models can capture the informative generator. In other words, reality is more complex than the prediction models, there is 'misspecification'. The setup is nested.

The results of this experiment II cannot be reported in detail here. Generally, the ARMA(1,1) model tends to forecast better than the AR(1) model.

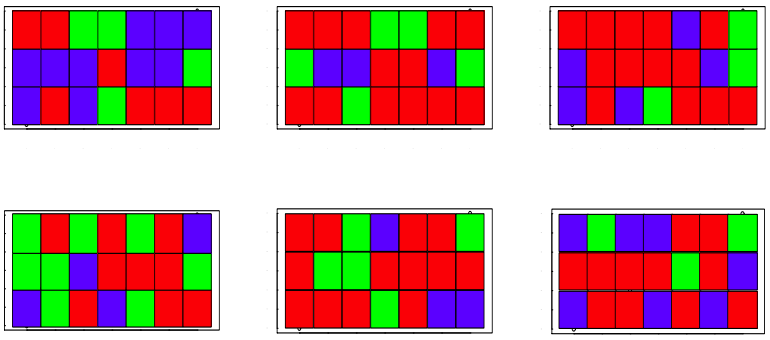


Fig. 3. Top row with N(0,1) error generators, bottom row with Cauchy generators. Left column $N = 25$, then $N = 50$, right $N = 100$. Red prefers AIC, green MSE simulator, blue case counting simulator. $\phi \in \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, $\theta \in \{-0.5, 0, 0.5\}$ on axes.

Convergence problems are encountered often, due to the borderline stationarity of the generators. The cross-simulation experiment is far less favorable for ARMA(1,1) than AIC. The dominant impression is that the simulation-based algorithm cannot compete with AIC, as it is too conservative, and this holds for both simulator variants.

3 An empirical application

We consider a four-variate system of food price, beverage price, crude oil price, and agricultural raw materials price. Monthly observations are available for 1992 to 2015. This section is based on an ongoing joint project ([9]).

The simulation technique is used for deciding modeling questions such as (a) whether a four-dimensional system should be used for forecasting or univariate models; (b) whether the variables should rather be differenced or used in their original (level) values; (c) whether statistically significant evidence on linear long-run structures (cointegration) should be imposed during the forecasting stage. We note that the first of these problems remain in the framework of classical asymptotics with their Central Limit Theorems, whereas the other two decision problems concern unit roots and deviate from the standard asymptotic theory. The forecast simulation has no problems with such asymptotic properties, it fully relies on a finite-sample evaluation. For the problems (a) and (b), we show the simulation evaluation in Figures 4 and 5.

Although some heterogeneity is recognizable across the series, the graphs in Figures 4 and 5 send some clear summary messages. In particular, if the correct

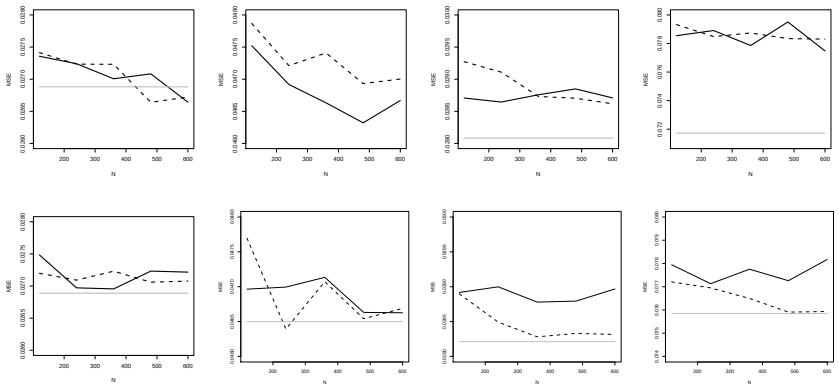


Fig. 4. Mean squared errors for forecasting the four prices (food, beverage, raw materials, crude oil), using univariate (solid) and multivariate (dashed) models. In the top row, univariate models are the generating models, in the bottom row multivariate models. Sample size on x–axis.

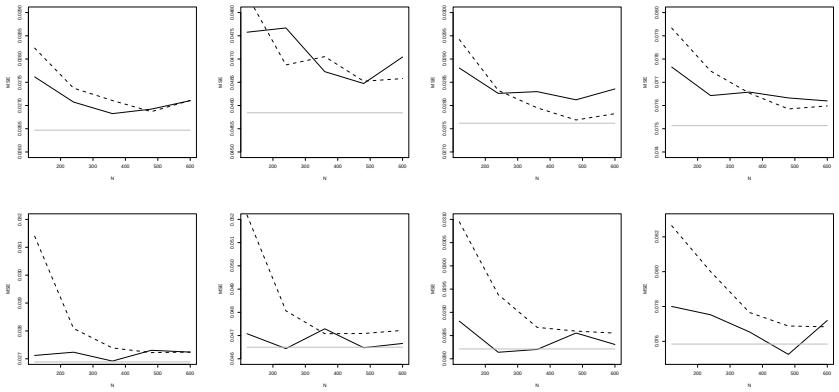


Fig. 5. Mean squared errors for forecasting the four prices (food, beverage, raw materials, crude oil), using level vector autoregressions (dashed) and VAR in first differences (solid). In the top row, level models are the generating models, in the bottom row models in differences. Sample size on x–axis.

model is in first differences, ignoring this fact and forecasting in a 'level' VAR entails inferior forecasts. On the other hand, if the correct model is not in differences, taking differences has a minor effect. This situation implies a general preference for differencing the data. Another feature is that multivariate models tend to outperform univariate models. Other similar but unreported experiments discourage the use of error correction and of seasonal dummy variables. In summary, the simulations permit clear guidelines for forecasting.

4 Summary and conclusion

Through argument and simulations, we have demonstrated that simulation-based choice of prediction models can be a convenient tool. It is only moderately time-consuming and deals with an important empirical concern: what if I have chosen the wrong model? How much do I lose? The technique, however, is no panacea. In simple parametric designs, it performs on a par with AIC, but it can be conservative. The technique may be weak if both forecast models are crude simplifications that miss main characteristics of the generator. The technique may fail if only one model misses the characteristics, as the simulation under the grossly simplified model may become uninformative. The simulation-based technique may be widespread in extant research, it often appears in informal discussion ([11]). We think that a more extensive exploration of designs will reveal further aspects.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 716-723 (1974)
2. Armstrong, J.S.: Significance tests harm progress in forecasting. International Journal of Forecasting 23, 321–327 (2007)
3. Burnham, K.P., and Anderson, D.R.: Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach. Springer-Verlag (2002)
4. Costantinin, M., and Kunst, R.M.: Combining forecasts based on multiple encompassing tests in a macroeconomic core system. Journal of Forecasting 30, 579596. (2011)
5. Dhrymes, P.: Time Series, Unit Roots, and Cointegration. Emerald. (1997)
6. Hauser, M.A., and Kunst, R.M.: Forecasting High-frequency Financial Data with the ARFIMA-ARCH model. Journal of Forecasting 20, 501-518 (2001)
7. Inoue, A., and Kilian, L.: On the selection of forecasting models. Journal of Econometrics 130, 273–306 (2006)
8. Jumah, A., and Kunst, R.M.: Seasonal Prediction of European Cereal Prices: Good Forecasts Using Bad Models? Journal of Forecasting 27, 391–406 (2008)
9. Jumah, A., and Kunst, R.M.: Time-series prediction of agricultural product prices: Are energy prices relevant?. Presented at the International Symposium on Forecasting, Santander. (2016)
10. Jumah, A., and Kunst, R.M.: Optimizing time-series forecasts for inflation and interest rates using simulation and model averaging. Applied Economics 48, 43664378 (2017)

11. Kolassa, S.: Sometimes It's Better to Be Simple than Correct. Foresight 40, 20–26. (2016)
12. Kruschke, J.K., and Liddell, T.M.: The bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review 25, 178–206. (2018)
13. Kunst, R.M.: Cross validation of prediction models for seasonal time series by parametric bootstrapping. Austrian Journal of Statistics 37, 271-284 (2008)
14. Kunst, R.M., and Jumah, A.: Toward a Theory of Evaluating Predictive Accuracy. IHS Economics Series, 162, Institute for Advanced Studies, Vienna (2004)
15. McQuarrie, A.D.R., and Tsai, C.L.: Regression and Time Series Model Selection. World-Scientific (1998)
16. Poirier, D.J.: Intermediate Statistics and Econometrics. MIT Press. (1995)
17. Pratt, J.W., Raiffa, H., Schlaifer, R.: Introduction to Statistical Decision Theory. MIT Press. (1996)
18. Shibata, R.: Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. Annals of Statistics 8, 147–164 (1980)

Robust autocovariance estimation from the frequency domain

Higor Cotta[1,2], Valdério Reisen[1,2], Pascal Bondon[2], and Céline Lévy-Leduc[3]

[1]NuMEs - DEST/PGGEA - Federal University of Espírito Santo - Brazil
[2]Laboratoire des Signaux et Systémes (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Gif sur Yvette, France
[3]UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France
higor.cotta@l2s.centralesupelec.fr
valderio.reisen@ufes.br
pascal.bondon@l2s.centralesupelec.fr
celine.levy-leduc@agroparistech.fr

Abstract. This paper proposes a robust estimation method of the sample autocovariance and autocorrelation functions in the presence of additive outliers. The robustness property is achieved by replacing the standard Fourier transform by its robust version obtained by substituting the least square procedure in the harmonic regression by the non-linear M-regression. Simulation experiments are conducted to assess the performance of the estimators under contaminated and non-contaminated scenarios.

Keywords: Autocovariance function, additive outliers, M-periodogram, estimation

1 Introduction

Atypical observations (outliers) are present in time series of diversified origins. It is well known that outliers significantly destroy the correlation structure of a time series even when only one atypical observation is present, see, for example, [1–3] and the references therein. As a possible approach for solving this problem, [4] proposed a highly robust estimator of the autocovariance function (ACOVF) and autocorrelation function (ACF), denoted by $\hat{\gamma}^{Q_h}$ and $\hat{\rho}^{Q_n}$, respectively. These estimators are based on the Q_n scale estimator proposed by [5], whose asymptotic properties were studied by [6] for univariate time series.

As noticed by [4], their robust ACOVF estimator does not provide a non-negative definite sample covariance matrix. Although this is an undesirable property for an autocovariance function estimator, the highly robust performance of $\hat{\gamma}_Q$ motivated its adoption by [3] to obtain an estimator of the spectral density function which is robust against additive outliers.

In addition, time series analysis in the frequency domain is based on the study of the spectral density function from which the periodogram is an estimator. As demonstrated by [3], the periodogram lacks robustness properties against

outliers. Therefore, robust methods to minimize the effect of outliers on the estimation of periodogram have to be considered. In this direction, different robust periodogram methods have been proposed by the literature, see, for instance, [3, 7–11], among others.

It is known that the periodogram can be obtained directly from a least squares estimates of the Fourier coefficients and is hence sensitive to outliers in data. Thus, to mitigate this problem, one may consider the use of robust regression methods, e.g., a robust M-regression, instead of the standard approach. Recently, this approach has been considered by [10] and [11] providing good results in the estimation of the coefficients of PARMA models and the fractional parameter of ARFIMA models, respectively.

Therefore, this work considers the estimation of the sample autocovariance and autocorrelation functions from the robust M-periodogram. The approach consists in fitting a robust harmonic regression to obtain a robust version of the discrete Fourier transform. That is, at each Fourier frequency, a sine and cosine coefficients are fitted using M-regression. Then, the ACOVF and ACF are obtained from an inverse diagonalization procedure of the matrix containing the estimated spectral density.

The outline of this paper is as follows: Section 2 presents the a briefly review of the influence of additive outliers in the estimation of the ACOVF, ACF and periodogram, Section 3 discusses the estimation of the ACOVF and ACF from the robust M-periodogram. Section 5 summarizes the simulation experiments and the robust performance of the estimators comparing them to [4]'s. Concluding remarks are given in Section 6.

2 Time series model with additive outliers

Let $\{x_t\}_{t\in\mathbb{Z}}$ be a weakly stationary process, and $\{z_t\}_{t\in\mathbb{Z}}$ be the contamination of $\{x_t\}$ by additive outliers,

$$z_t = x_t + \omega\delta_t, \tag{1}$$

where ω is the magnitude of the outliers affecting $\{x_t\}$ and $\{\delta_t\}$ are uncorrelated random variables with $\mathsf{Pr}(\delta_t = -1) = \mathsf{Pr}(\delta_t = 1) = p/2$, and $\mathsf{Pr}(\delta_t = 0) = 1 - p$ with $0 < p < 1$. Notice that δ_t is the product of Bernoulli(p) and Rademacher independent random variables. It follows that $\mathrm{E}[\delta_t] = 0$ and $\mathrm{Var}[\delta_t] = p$. We assume that x_t and δ_t are independent random variables. Related to the Model in (1), the impact of additive outliers on the ACOVF and on the spectral density (when it exists) is the following:

Remark 1. Suppose that $\{z_t\}$ follows (1) and $\{x_t\}$ has spectral density f_x.

i. The autocovariance function (ACOVF) of $\{z_t\}$ is given by

$$\gamma_z(h) = \begin{cases} \gamma_x(0) + \omega^2 p, & \text{if } h = 0, \\ \gamma_x(h), & \text{if } h \neq 0. \end{cases} \tag{2}$$

ii. The spectral density function of $\{z_t\}$ is given by

$$f_z(\lambda) = f_x(\lambda) + \frac{1}{2\pi}\omega^2 p, \quad \lambda \in [-\pi, \pi]. \tag{3}$$

Note that the occurrence of additive outliers increases the variance of the process and reduces the autocorrelation. In the same way, the spectral density of $\{x_t\}$ is increased in the presence of outliers.

In the case of the sample ACOVF and periodogram, the impacts are:

Remark 2. Let $\{z_t\}_{t=1,\dots n}$ be a set of observations generated from (1) and let the outlier occur at time $t = T$ with $h < T < n - h$. It follows that:

i. The impact on the sample ACOVF is given by

$$\begin{aligned}\hat{\gamma}_z(h) = \hat{\gamma}_x(h) \pm \frac{\omega}{n}(x_{_{T-h}} + x_{_{T+h}} - 2\bar{x}) - \frac{\omega^2}{n^2} \\ + \frac{\omega^2}{n}\mathbb{1}_{\{h=0\}} + o_p\left(\frac{1}{n}\right),\end{aligned} \tag{4}$$

where $\hat{\gamma}_x(h) = \dfrac{1}{n}\sum\limits_{t=1}^{n-h}(x_t - \bar{x})(x_{t+h} - \bar{x})$.

ii. The impact on the periodogram is given by

$$I_z(\lambda) = I_x(\lambda) + \frac{1}{2\pi}\Delta(\omega), \tag{5}$$

where

$$\Delta(\omega) = \frac{\omega^2}{2\pi n} \pm \frac{\omega}{\pi n}\left[(x_{_T} - \bar{x}) + \sum_{h=1}^{n-1}(x_{_{T-h}} + x_{_{T+h}} - 2\bar{x})\cos(h\lambda)\right] + o_p\left(\frac{1}{n}\right).$$

3 Robust estimation of autocovariance function

3.1 Robust estimation of autocovariance function from the periodogram

It is known that a given time series $\{x_t\}_{t=1,\dots,n}$ can be represented as the sum of n sines and cosines at the Fourier frequencies $\lambda_k = 2\pi k/n, k = 1, \dots, [n/2]$. The periodogram may be obtained from the least square estimates of the following linear regression

$$x_t = C'_{nt}\boldsymbol{\beta}(\lambda_k) + \varepsilon_t = \beta_1\cos(t\lambda_k) + \beta_2\sin(t\lambda_k) + \varepsilon_t, 1 \le t \le n, \tag{6}$$

where $C'_{nt} = [\cos(t\lambda_k), \sin(t\lambda_k)]$, $\boldsymbol{\beta}(\lambda_k) = [\beta_1(\lambda_k), \beta_2(\lambda_k)]'$ and ε_t is the error term. The least square estimates of C'_r is obtained from

$$\hat{\boldsymbol{\beta}}_n(\lambda_k) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}}\left[\sum_{t=1}^{n}(x_t - C'_{nt}\boldsymbol{\beta}(\lambda_k))^2\right], \tag{7}$$

where $\boldsymbol{\beta}_n(\lambda_k) = [\beta_1(\lambda_k), \beta_2(\lambda_k)]'$.

The solution of (7) is

$$\hat{\boldsymbol{\beta}}_n(\lambda_k) = \frac{2}{n}\left(\sum_{j=1}^{n} x_j \cos(j\lambda_k) \qquad \sum_{j=1}^{n} x_j \sin(j\lambda_k)\right)'. \tag{8}$$

Now, consider the classical periodogram given by

$$I_n(\lambda_k) = \frac{1}{2\pi}|a_k|^2 = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} x_j \exp(ij\lambda_k)\right|^2, \tag{9}$$

where a_k is the discrete Fourier transform of $\{x_t\}$, as given by

$$a_k = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} x_j \exp(ij\lambda_k). \tag{10}$$

Relating (8) and (9), it can be shown that

$$I_n(\lambda_k) = \frac{n}{8\pi}\|\hat{\boldsymbol{\beta}}_n(\lambda_k)\|^2 = \frac{n}{8\pi}\left((\hat{\beta}_1(\lambda_k))^2 + (\hat{\beta}_2(\lambda_k))^2\right), \tag{11}$$

where $\|.\|$ is the Euclidean norm.

The M-estimator $\hat{\boldsymbol{\beta}}_{n,\psi}(\lambda_k)$ is defined as the solution of

$$\sum_{t=1}^{n} C_{nt}\psi(x_t - C'_{nt}\boldsymbol{\beta}_{n,\psi}(\lambda_k)) = 0, \tag{12}$$

where, $\boldsymbol{\beta}_{n,\psi}(\lambda_k) = [\beta_{1,\psi}(\lambda_k), \beta_{2,\psi}(\lambda_k)]$, and following the similar lines as in [11], ψ is the Huber function defined by

$$\psi(x) = \begin{cases} x, & if \ \ |x| \leq k, \\ sign(x)k, & if \ \ |x| > k, \end{cases} \tag{13}$$

where k is some positive constant. In the following, $k = 1.345$ is adopted to ensure an efficiency of 95% for the regression estimator in Gaussian case. Note that ψ satisfies the assumptions

Assumption 1. *ψ is a bounded non decreasing real valued function on $\mathbb{R}$.*

Assumption 2. *$\lambda(x) = \mathrm{E}[\psi(\varepsilon_0 - x)]$ is well defined, strictly decreasing on $\mathbb{R}$ and $\lambda(0) = 0$.*

Therefore, similarly to (9), the robust M-periodogram is defined by

$$I_n^M(\lambda_k) = \frac{n}{8\pi}\|\hat{\boldsymbol{\beta}}_{n,\psi}(\lambda_k)\|^2 = \frac{n}{8\pi}\left((\hat{\beta}_{1,\psi}(\lambda_k))^2 + (\hat{\beta}_{2,\psi}(\lambda_k))^2\right). \tag{14}$$

Now, consider the $n \times n$ circulant covariance matrix $\boldsymbol{\Gamma}^n = [\gamma(i-k)]_{i,k=1}^n$, where γ denotes an absolutely summable autocovariance function. Let $\boldsymbol{Q}$ be the eigenvectors associated to the diagonalization of $\boldsymbol{\Gamma}^n$ defined by

$$\boldsymbol{Q}\prime = [\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{s}_1, \ldots, \boldsymbol{c}_{[n/2]}, \boldsymbol{s}_{[n/2]}] \tag{15}$$

and

$$\boldsymbol{Q}\prime = [\boldsymbol{c}_0, \boldsymbol{c}_{1,1}, \ldots, 2^{-1/2}\,\boldsymbol{c}_{n/2}], \tag{16}$$

for odd and even n, respectively, where, $j = 1, \ldots, [n/2]$,

$$\boldsymbol{c}_0 = \sqrt{1/n}[1, 1, 1, \ldots, 1], \tag{17}$$

$$\boldsymbol{c}_j = \sqrt{2/n}[1, \cos\lambda_j, \cos 2\lambda_j, \ldots, \cos(n-1)\lambda_j], \tag{18}$$

and

$$\boldsymbol{s}_j = \sqrt{2/n}[1, sin\lambda_j, \sin 2\lambda_j, \ldots, \sin(n-1)\lambda_j]. \tag{19}$$

Let $\boldsymbol{D}_n$ be a $n \times n$ matrix defined by

$$\boldsymbol{D}_n = \begin{cases} \text{diag}\{f(0), \ldots, f(\lambda_{[n/2]}), f(\lambda_{[n/2]})\} & \text{if } n \text{ is odd,} \\ \text{diag}\{f(0), \ldots, f(\lambda_{(n-2)/2}), f(\lambda_{(n-2)/2}), f(\lambda_{n/2})\} & \text{if } n \text{ is even,} \end{cases} \tag{20}$$

where $f(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma(h)e^{-ih\lambda}$. It is known that the diagonal matrix $\boldsymbol{Q}'\,\boldsymbol{\Gamma}^n\,\boldsymbol{Q}$ converges to $2\pi\,\boldsymbol{D}_n$ as n increases [12]. Now, replace γ and f by their sample counterparts, i.e, $\hat{\gamma}(h) = n^{-1}\sum_{i=1}^{n-|h|}(x_t - \bar{x})(x_{t+h} - \bar{x})$ and I, respectively. Therefore, one possible approach for the estimation of the ACOVF is

$$\hat{\boldsymbol{\Gamma}}^n = 2\pi\,\boldsymbol{Q}\,\hat{\boldsymbol{D}}_n\,\boldsymbol{Q}\prime, \tag{21}$$

where $\hat{\boldsymbol{D}}_n$ is defined similarly as (20).

In order to robustify the autocovariance matrix we propose the following estimator

$$\hat{\boldsymbol{\Gamma}}_n^M = 2\pi\,\boldsymbol{Q}\,\hat{\boldsymbol{D}}_n^M\,\boldsymbol{Q}\prime, \tag{22}$$

where $\hat{\boldsymbol{D}}_n^M$ is defined similarly as (20) but replacing f with I_M.

Likewise, we define the $n \times n$ autocorrelation matrix $\hat{\boldsymbol{\rho}}$ and its robust version $\hat{\boldsymbol{\rho}}_n^M$ by

$$\hat{\boldsymbol{\rho}}^n = \frac{\hat{\boldsymbol{\Gamma}}^n}{\hat{\gamma}_{11}}, \tag{23}$$

and

$$\hat{\boldsymbol{\rho}}^M = \frac{\hat{\boldsymbol{\Gamma}}^M}{\hat{\gamma}_{11}^M}, \tag{24}$$

where $\hat{\gamma}_{11} = \hat{\gamma}(0)$ and $\hat{\gamma}_{11}^M = \hat{\gamma}^M(0)$.

Finally, the ACOVF and ACF of $\{x_t\}$, $\hat{\gamma}^n$ and $\hat{\rho}^n$, and their robust counterparts, namely $\hat{\gamma}^M$, $\hat{\rho}^M$, are extracted from the rows of the circulant matrices $\hat{\boldsymbol{\Gamma}}_n$, $\hat{\boldsymbol{\rho}}_n$ $\hat{\boldsymbol{\Gamma}}_n^M$ and $\hat{\boldsymbol{\rho}}_n^M$, respectively.

3.2 Robust autocovariance function using Q_n

In this section, for the purpose of comparison between the proposed sample ACF and the Robust ACF of [4], the latter is summarized in the sequel. Based on the Q_n scale estimator proposed by [5], [4] proposed the following highly robust estimator of the ACOVF for univariate time series $\{x_t\}$

$$\hat{\gamma}^{Q_n}(h) = \frac{1}{4}\left[Q^2_{n-h}(U+V) - Q^2_{n-h}(U-V)\right], \tag{25}$$

where U and V are vectors containing the initial $n-h$ and the final $n-h$ observations of $\{x_t\}$, respectively. Then, the autocorrelation function can be obtained from

$$\hat{\rho}^{Q_n}(h) = \frac{Q^2_{n-h}(U+V) - Q^2_{n-h}(U-V)}{Q^2_{n-h}(U+V) + Q^2_{n-h}(U-V)}, \tag{26}$$

where U and V are also vectors containing the initial $n-h$ and the final $n-h$ of $\{x_t\}$. It can be shown that $|\hat{\rho}_{Q_n}(h)| \leq 1$.

The asymptotic results of the above robust autocovariance in time series with short and long memory properties were the motivation for the papers of [6] and [13]. Theorem 4 in [6] presents the central limit theorem for the autocorrelation given by (26).

4 Robust estimators for autoregressive coefficients

Let $\{x_t\}_{t=1,\ldots,n}$ be the zero-mean autoregressive process of order p (AR(p)),

$$x_t - \phi_1 x_{t-1} - \ldots \phi_p x_{t-p} = \varepsilon_t \tag{27}$$

where $|\phi| < 1$ and ε_t is a zero mean Gaussian white noise process with variance σ^2. Here, we are interested in finding estimators of the coefficient vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)\prime$ and of the white noise variance σ^2. Now, consider the Yule-Walker equations,

$$\boldsymbol{\Gamma}_p \, \boldsymbol{\phi} = \gamma_p, \tag{28}$$

and

$$\sigma^2 = \gamma(0) - \boldsymbol{\phi}\prime \boldsymbol{\gamma}_p, \tag{29}$$

where $\boldsymbol{\Gamma}_p = [\gamma(i-k)]_{i,k=1}^{p}$ and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))\prime$. It is well-known that replacing $\boldsymbol{\Gamma}_p$ with $\hat{\boldsymbol{\Gamma}}_p$ in (28) leads to the so-called Yule-Walker estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\sigma}^2$ of $\boldsymbol{\phi}$ and σ^2,

$$\hat{\boldsymbol{\Gamma}}_p \hat{\boldsymbol{\phi}} = \hat{\gamma}_p, \tag{30}$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}\prime \hat{\boldsymbol{\gamma}}_p. \tag{31}$$

Thus, the robustified Yule-Walker estimating procedure consists in replacing the sample ACOVF estimator $\hat{\boldsymbol{\Gamma}}_p$ with $\hat{\boldsymbol{\Gamma}}_p^M$ in (21) and using the Durbin-Levinson algorithm to solve the resulting system of equations.

5 Monte Carlo experiments

This section reports a Monte Carlo simulation study to investigate the performance of the robust sample ACOVF and ACF estimators discussed previously. In the following, (1) is considered as the data generating process, where $\{x_t\}$ is an autoregressive process of order 1 ($AR(1)$) as follows:

$$x_t = \phi x_{t-1} + \varepsilon_t, \tag{32}$$

where $|\phi| < 1$ and ε_t is a zero mean Gaussian white noise process with variance σ^2. For the simulations, $\phi = 0.7$, $\sigma^2 = 1$, $\omega = 15$ and $p = 0.05, 0.10$ and 0.15 are set. The sample size are $n = 200, 500$ and 1000, and each experiment is replicated 1000 times. Two scenarios are considered: (i) the samples are uncontaminated ($p = 0$), and (ii) the samples are contaminated ($p \neq 0$). Under both scenarios, the comparison between the estimators is done by contrasting the plots and the empirical root mean square error (RMSE) of the theoretical $\rho(h) = \phi^h$ with $\hat{\rho}(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$. $\hat{\rho}(h)$ denotes the standard sample estimator of the ACF.

In this direction, Figure 1 displays the plots of $\rho(h)$ and the means of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$ for the uncontaminated scenario. The case of $p = 0.05$ is shown in Figure 2. For the uncontaminated case, we observe similar behavior in all the plots indicating that all the estimators capture the correlation structure of the series. The effects of additive outliers appear by comparing the true correlation to the sample estimates under a contaminated scenario. Not surprisingly, $\hat{\rho}(h)$ and $\hat{\rho}^n(h)$ were completely affect while $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$ provided values much closer to the theoretical ones.

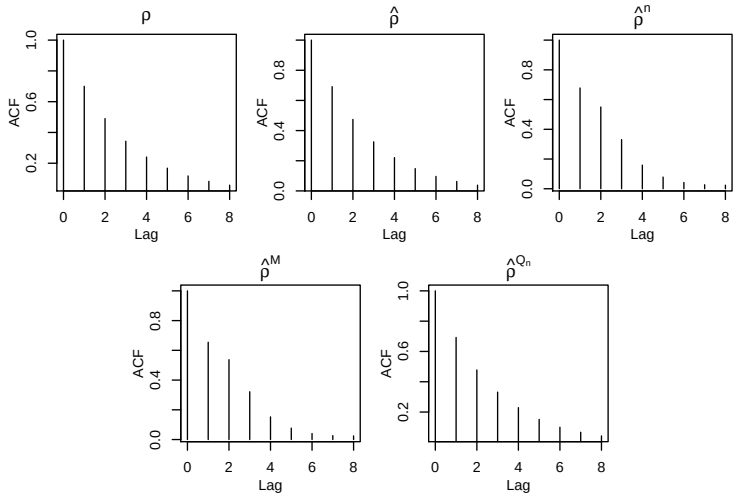


Fig. 1. Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, and $\hat{\rho}^M(h)$, $\hat{\rho}^{Q_n}(h)$ when $p = 0$.

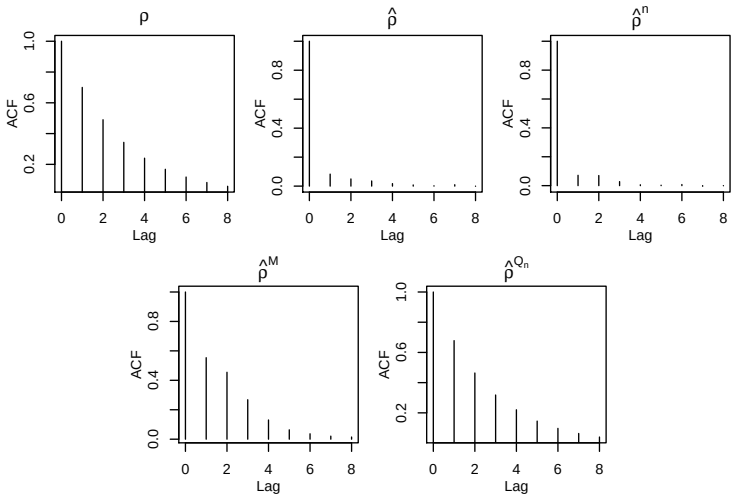


Fig. 2. Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, and $\hat{\rho}^M(h)$, $\hat{\rho}^{Q_n}(h)$ when $p = 0.05$.

Related to Figures 1 and 2, Figure 3 and 4 show the boxplot of the simulated ACF values for both scenarios, respectively.

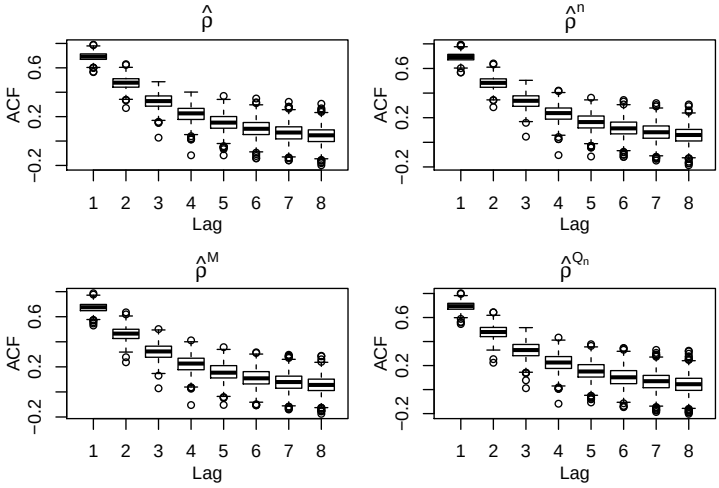


Fig. 3. Boxplots of estimated ACF of z_t when $p = 0$.

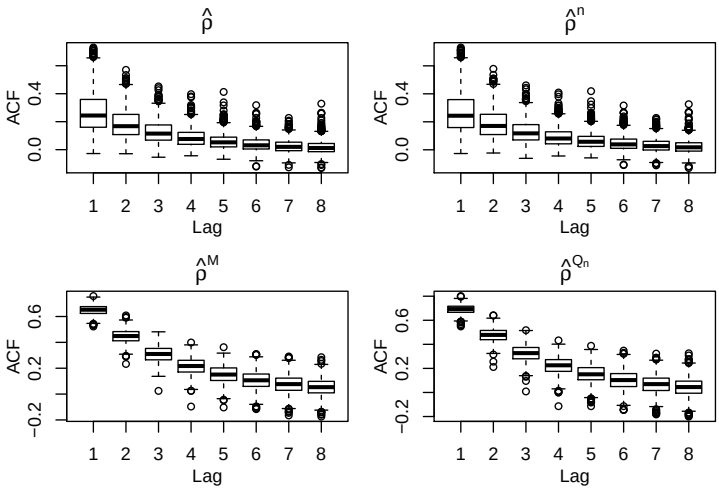


Fig. 4. Boxplots of estimated ACF of z_t when $p = 0.05$.

In Table 1, we present the root mean squared errors (RMSE) of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$ as n increases, for $h = 0, 1, \ldots, 8$ and $p = 0$. The results for the contaminated scenario $p \neq 0$ are displayed in Tables 2, 3 and 4, for $p = 0.05$, $p = 0.10$ and $p = 0.15$, respectively. From Table 1, we observe that the samples and the robust estimators have RMSE close to the each other in the absence of contamination. As expected, $\hat{\rho}(h)$ performed best. Comparing $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, we find that $\hat{\rho}^{Q_n}(h)$ has a slight better performance than $\hat{\rho}^M(h)$ for all lags. Therefore, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$ are useful even in the context which the presence of additive outliers is uncertain.

Table 1. RMSE of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$.

	n	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	200	0.0514	0.0794	0.0962	0.1060	0.1118	0.1137	0.1141	0.1134
	500	0.0347	0.0540	0.0645	0.0696	0.0726	0.0729	0.0742	0.0752
	1000	0.0231	0.0352	0.0425	0.0475	0.0506	0.0521	0.0528	0.0532
$\hat{\rho}^n(h)$	200	0.0510	0.0753	0.0894	0.0983	0.1049	0.1079	0.1085	0.1092
	500	0.0344	0.0526	0.0626	0.0676	0.0706	0.0709	0.0724	0.0738
	1000	0.0230	0.0347	0.0415	0.0463	0.0494	0.0510	0.0517	0.0524
$\hat{\rho}^M(h)$	200	0.0635	0.0832	0.0926	0.0987	0.1030	0.1050	0.1050	0.1060
	500	0.0461	0.0610	0.0670	0.0699	0.0707	0.0701	0.0707	0.0719
	1000	0.0352	0.0437	0.0464	0.0485	0.0502	0.0509	0.0511	0.0518
$\hat{\rho}^{Q_n}(h)$	200	0.0533	0.0835	0.1018	0.1131	0.1218	0.1235	0.1252	0.1270
	500	0.0373	0.0589	0.0702	0.0761	0.0785	0.0777	0.0794	0.0810
	1000	0.0251	0.0383	0.0459	0.0503	0.0538	0.0555	0.0564	0.0576

Table 2. BIAS of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$.

	n	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	200	-0.0193	-0.0279	-0.0322	-0.0340	-0.0316	-0.0305	-0.0298	-0.0292
	500	-0.0078	-0.0126	-0.0136	-0.0150	-0.0161	-0.0167	-0.0167	-0.0164
	1000	-0.0054	-0.0081	-0.0091	-0.0104	-0.0113	-0.0112	-0.0110	-0.0092
$\hat{\rho}^n(h)$	200	-0.0186	-0.0175	-0.0134	-0.0089	-0.0024	0.0013	0.0036	0.0050
	500	-0.0073	-0.0081	-0.0055	-0.0041	-0.0035	-0.0031	-0.0026	-0.0020
	1000	-0.0051	-0.0058	-0.0051	-0.0051	-0.0050	-0.0043	-0.0038	-0.0018
$\hat{\rho}^M(h)$	200	-0.0384	-0.0367	-0.0287	-0.0201	-0.0104	-0.0041	-0.0008	0.0015
	500	-0.0273	-0.0279	-0.0216	-0.0160	-0.0117	-0.0086	-0.0065	-0.0046
	1000	-0.0250	-0.0254	-0.0206	-0.0162	-0.0130	-0.0096	-0.0070	-0.0038
$\hat{\rho}^{Q_n}(h)$	200	-0.0170	-0.0252	-0.0294	-0.0302	-0.0293	-0.0269	-0.0284	-0.0302
	500	-0.0078	-0.0131	-0.0146	-0.0151	-0.0161	-0.0162	-0.0157	-0.0158
	1000	-0.0053	-0.0080	-0.0089	-0.0097	-0.0106	-0.0100	-0.0097	-0.0082

Now, considering the occurrence of outliers, in Table 2, we see that the RMSE of the sample estimate is much larger than the RMSE of the robust estimators when the percentage of contamination is 5%, and, thus, confirming that even a small fraction of contamination can make the sample ACF useless. Moreover, the RMSE of the robust estimators are almost the same in the uncontaminated and the contaminated cases, but, again, $\hat{\rho}^{Q_n}(h)$ performs better than $\hat{\rho}^M(h)$.

Table 3. RMSE of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.05$.

	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6377	0.4510	0.3186	0.2252	0.1617	0.1186	0.0887	0.0693
$\hat{\rho}^n(h)$	0.6365	0.4493	0.3165	0.2228	0.1591	0.1160	0.0866	0.0673
$\hat{\rho}^M(h)$	0.1418	0.1151	0.0940	0.0798	0.0722	0.0677	0.0659	0.0659
$\hat{\rho}^{Q_n}(h)$	0.0432	0.0642	0.0741	0.0768	0.0786	0.0783	0.0797	0.0809

To empirically investigate the breakdown point of the proposed estimator, the RMSE of $\hat{\rho}^M(h)$ as the percentage of outliers in $\{z_t\}$ increases are presented in Tables 3 and 4, respectively. Comparing both tables, note that increasing the percentage of outliers reduces the performance of both estimators. However, not surprisingly, $\hat{\rho}^M(h)$ and $\hat{\rho}_{Q_n}(h)$ are less affect by the outliers, although $\hat{\rho}_{Q_n}(h)$ still performs slightly better than $\hat{\rho}^M(h)$.

Table 4. RMSE of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.10$.

	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6574	0.4645	0.3280	0.2339	0.1668	0.1220	0.0916	0.0713
$\hat{\rho}^n(h)$	0.6558	0.4627	0.3256	0.2317	0.1644	0.1197	0.0894	0.0694
$\hat{\rho}^M(h)$	0.2317	0.1717	0.1268	0.0984	0.0803	0.0711	0.0668	0.0625
$\hat{\rho}^{Q_n}(h)$	0.0549	0.0722	0.0779	0.0787	0.0786	0.0783	0.0799	0.0800

Table 5. RMSE of $\hat{\rho}(h)$, $\hat{\rho}^n(h)$, $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_n}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.15$.

	0	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6587	0.4688	0.3325	0.2363	0.1700	0.1228	0.0922	0.0725	
$\hat{\rho}^n(h)$	0.6572	0.4669	0.3304	0.2339	0.1678	0.1211	0.0904	0.0709	
$\hat{\rho}^M(h)$	0.3091	0.2232	0.1630	0.1195	0.0943	0.0766	0.0677	0.0612	
$\hat{\rho}^{Q_n}(h)$	0.0738	0.0860	0.0861	0.0834	0.0835	0.0804	0.0802	0.0795	

To end this empirical investigation study, we consider the estimation of $\phi = 0.7$ in (32) where $p = 0$ and $p = 0.01$ with $\omega = 10$ for $n = 500$ under 1000 replications for the estimators presented in Section 4. The means, bias and RMSE of $\hat{\phi}$, $\hat{\phi}^M$ and $\hat{\phi}^{Q_n}$ are presented in Table 5. As one may see, in the absence of contamination $\hat{\phi}^M$ and $\hat{\phi}^{Q_n}$ are close to $\hat{\phi}$ which presented the smallest bias and RMSE. When $p = 0.1$, not surprisingly, $\hat{\phi}$ is affected while $\hat{\phi}^M$ and $\hat{\phi}^{Q_n}$ were robust against outliers. Again, $\hat{\phi}^{Q_n}$ presented a better performance than $\hat{\phi}^M$, However, since $\hat{\rho}^M(h)$ is positive semi definite by construction, its adoption is recommended over $\hat{\rho}_{Q_n}(h)$.

	p	Mean	Bias	RMSE
$\hat{\phi}$	0	0.6939	-0.0061	0.1059
$\hat{\phi}^M$		0.6613	-0.0387	0.1703
$\hat{\phi}^{Q_n}$		0.6929	-0.0071	0.1238
$\hat{\phi}$	0.01	0.3731	-0.3269	5.6619
$\hat{\phi}^M$		0.6372	-0.0628	0.2384
$\hat{\phi}^{Q_n}$		0.6890	-0.0110	0.1897

6 Conclusion

This paper presented a new estimation method for the sample autocovariance and the sample autocorrelation functions of stationary univariate processes. The procedure consists in replacing the traditional periodogram by the robust M-periodogram in the inverse diagonalization procedure of the matrix containing the estimated spectral density. The proposed method is also robust to additive

outliers. Therefore, the authors suggest the use of the proposed method in an analysis where the occurrence of additive outliers is possible.

Acknowledgments

The authors would like to thank the support of CNPq, ERASMUS, CAPES and FAPES.

References

1. Chan, W.: A note on time series model specification in the presence of outliers. Journal of Applied Statistics **19**(1) (1992) 117–124
2. Chan, W.: Outliers and financial time series modelling: a cautionary note. Mathematics and Computers in Simulation **39**(3) (1995) 425–430
3. Molinares, F.F., Reisen, V.A., Cribari-Neto, F.: Robust estimation in long-memory processes under additive outliers. Journal of Statistical Planning and Inference **139**(8) (2009) 2511–2525
4. Ma, Y., Genton, M.G.: Highly robust estimation of the autocovariance function. Journal of Time Series Analysis **21** (2000) 663–684
5. Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. Journal of the American Statistical Association **88**(424) (1993) 1273–1283
6. Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M.S., Reisen, V.A.: Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes. Journal of Time Series Analysis **32**(2) (2011) 135–156
7. Zhang, Z., Chan, S.C.: Robust adaptive lomb periodogram for time-frequency analysis of signals with sinusoidal and transient components. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on. Volume 4., IEEE (2005) iv–493
8. Li, T.H.: Laplace periodogram for time series analysis. Journal of the American Statistical Association **103**(482) (2008) 757–768
9. Li, T.H.: A nonlinear method for robust spectral analysis. IEEE Transactions on Signal Processing **58**(5) (2010) 2466–2474
10. Sarnaglia, A.J.Q., Reisen, V.A., Bondon, P., Lévy-Leduc, C.: A robust estimation approach for fitting a PARMA model to real data. In: 2016 IEEE Statistical Signal Processing Workshop (SSP). (2016)
11. 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
12. Fuller, W.A.: Introduction to statistical time series. Volume 428. John Wiley & Sons (1996)
13. Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M.S., Reisen, V.A.: Large sample behavior of some well-known robust estimators under long-range. Statistics **45**(1) (2011) 59–71

Penalty terms for estimation of ARMA models: A Bayesian inspiration

Helgi Tómasson

University of Iceland, Faculty of Economics,
Oddi v/Sturlugötu, IS-101 Reykjavík, Iceland

Abstract. Bayesian methods are based on combining a problem, a model, prior information, and data using Bayes rule. This paper addresses the implementation of a Bayesian approach to stationary ARMA models. The interpretation of the parameters of a ARMA models is somewhat non-intuitive. The interpretation of the spectral function is much clearer. A Bayesian expression of a prior belief in the frequency domain, i.e., stating a preference on the shape of the spectral function, may therefore be more natural than formulating a prior on the time-domain parameters. Stating a prior on a function space is non-trivial. In this paper the fact that the spectral density of an ARMA model is rational is exploited. The use of complex theory residue calculus is used to derive analytic measures of desirable features of the spectral function. The approach is equally suited for discrete- and continuous-time models.

Keywords: Residue calculus, rational spectrum, ARMA

1 Introduction

Bayesian methods are based on combining a problem, a model, prior information and data using Bayes rule. This paper addresses the implementation of a Bayesian approach to stationary ARMA (Auto-Regressive-Moving-Average), and continuous-time-ARMA (CARMA), models. The parameters of ARMA and CARMA models are essentially a parsimonious way of modelling an auto-correlation function. In general correlations, and in particular an auto-correlation function (ACF), and the ARMA-parameters, of a stationary process are hard to interpret. Therefore, defining a sensible a priori opinion about correlations is a non-trivial issue.

In time-series analysis the spectral curve, the Fourier transform of the ACF is much easier to interpret. Expressing a prior opinion in the frequency domain is therefore a more natural approach. However, the spectral density is a continuous function and operating with a probability distribution on a function space is difficult.

A natural Bayesian approach is to state the prior in the frequency domain, i.e., that *a priori*, the spectral density is of a particular type, or is in some sense "close" to a particular spectral density. The concept "close" requires some measure of distance between spectral densities. A conceivable measure of closeness between two curves is the Kullback-Leibler distance.

Calculation of a Bayes estimator can often be implemented as a frequentistic estimator with an added penalty term. The penalty term biases the classical estimator towards a prior. In this case, a prior of a smooth spectrum is illustrated. The technique is based on the fact that the spectrum of the ARMA model is rational. The residue calculus of complex analysis gives exact expressions of some integrals of rational functions. In particular a measure of smoothness, e.g., the integral of the squared second derivative of the spectral density can be calculated directly. A penalty term based on a function of this measure can then be added to an objective function, e.g., a log-likelihood function. Then this improved objective function can be used to shrink the fitted model towards a priori ideas of the spectral shape. This approach can be modified implement other forms of a priori information on the spectral function.

This paper is organized as follows. First a brief review of the continuous-time ARMA model is given. Section 3 shows an intuition of a Bayesian approach and the interpretation of the prior term in the likelihood function as a penalty term in classical estimation. Section 4 reviews mathematical results on partial fractions and residue calculus that are useful for calculation of some penalty terms of interest. In section 5 the computational machinery for a continuous-time ARMA implemented in an R-package is illustrated. Section 6 concludes with a discussion on extending these ideas to discrete-time models, other types of penalty functions, comparison with other types of penalty terms like AIC and BIC.

2 On ARMA models

A continuous-time ARMA, CARMA, process can be defined in terms of a continuous-time innovation process and a stochastic integral. A common choice of innovation process is the Wiener process, $W(t)$. A representation of a CARMA(p,q) process in terms of the differential operator D is:

$$\begin{gathered} Y^{(p)}(t) + \alpha_1 Y^{(p-1)}(t) + \cdots + \alpha_p Y(t) = \\ \sigma d(W(t) + \beta_1 W^1(t) + \cdots + \beta_q W^{(q))}(t)), \\ \text{or } \boldsymbol{\alpha}(D)\, Y(t) = \sigma \boldsymbol{\beta}(D)\, dW(t), \\ \boldsymbol{\alpha}(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p, \\ \boldsymbol{\beta}(z) = 1 + \beta_1 z + \cdots + \beta_q z^q. \end{gathered}$$

Here, $Y^{(p)} = D^p Y(t)$, denotes the p-th derivative of $Y(t)$. The path of a Wiener process is nowhere differentiable so the symbol $D\,W(t)$, and higher derivatives, is of a purely notational nature. The spectral density of $Y(t)$ is a rational function:

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{\boldsymbol{\beta}(i\omega)\boldsymbol{\beta}(-i\omega)}{\boldsymbol{\alpha}(i\omega)\boldsymbol{\alpha}(-i\omega)}.$$

The spectral representation of CARMA is:

$$Y(t) = \int_{-\infty}^{\infty} \exp(i\omega\, t) dZ(\omega),$$
$$E(dZ(\omega)) = 0, \quad E(dZ(\omega)\overline{dZ(\omega)}) = f(\omega)d\omega,$$
$$E(dZ(\omega)\overline{dZ(\lambda)}) = 0, \quad \lambda \neq \omega.$$

In mathematics an univariate linear dynamic system can be expressed as a linear differential equation of a particular order. This can be written as a multidimensional first order differential equation. In the state space form, the AR part of CARMA represents a linear differential equation. Just as in the discrete-time ARMA has several possible state-space representations, the continuous-time CARMA has several possible state-space representations. See, e.g, Tsay (2010) for the discrete-time case, and Bergstrom (1988) and Zadrozny (1988) for the continuous-time case.

The stationarity condition of the CARMA requires the roots of the polynomial $\boldsymbol{\alpha}(z)$ to have negative real-parts and that $p > q$.

3 Intuition

A standard Bayesian approach for the normal model is to assign a normal prior for the mean, e.g,:

$$X|\mu, \sigma \sim N(\mu, \sigma^2),$$
$$\mu|\sigma \sim N(\mu_0, \sigma_0^2), \quad \sigma_0 = \tau\sigma.$$

Given data, $x_1, \ldots, x_n$, and reparameterizing, $v = 1/\sigma$, $v_0 = k_0 v$, the log-posterior is (σ known):

$$log(p(\mu, v|x_1, \ldots, x_n, \mu_0, k_0)) =$$
$$\text{constant} + \underbrace{\frac{n}{2}\log(v) - v\sum_{i=1}^{n}(x_i - \mu)^2/2}_{A} + \underbrace{\frac{1}{2}\log(v) - k_0 v(\mu - \mu_0)^2/2}_{B}.$$

The number k_0 expresses the certainty in the prior. If k_0 is set to zero and the log-posterior (as a function of μ) is maximized the result is the ML estimator and a nonzero k_0 biases the ML-estimate towards the prior (μ_0). The mode of the posterior can serve as a Bayesian point estimate. The objective function is just the likelihood function, A, with an extra "penalty term", B, penalizing for deviations from the central value of the *prior*-distribution. Penalty terms are commonly added to the log-likelihood functions as a model selection tool. Well known examples are AIC and BIC.

The statistical analysis of a ARMA is essentially estimating a correlation structure based on one observations of a particular vector $\boldsymbol{y} = (y(t_1), \ldots, y(t_n))$.

The parameters of the ARMA are, a scale parameter σ, and a set of parameters that decide the auto-correlation function, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. I.e:

$$\boldsymbol{y} \sim N(\boldsymbol{0}, M(\boldsymbol{\alpha}, \boldsymbol{\beta})\sigma^2).$$

The likelihood (with $v = 1/\sigma$) is then:

$$L(v, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}) \propto v^{\frac{n}{2}} |M(\boldsymbol{\alpha}, \boldsymbol{\beta})|^{-1/2} e^{-v\boldsymbol{y}' M(\boldsymbol{\alpha},\boldsymbol{\beta})^{-1}\boldsymbol{y}/2}.$$

Using a gamma prior for v,

$$v \sim gamma(a, b),$$

yields an analytical form of the posterior for v,

$$p(v|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{y}) \propto v^{\frac{n}{2}} |M(\boldsymbol{\alpha}, \boldsymbol{\beta})|^{-1/2} e^{-v\boldsymbol{y}' M(\boldsymbol{\alpha},\boldsymbol{\beta})^{-1}\boldsymbol{y}/2} v^{a-1} e^{-bv},$$

i.e. the posterior is:

$$gamma(a + n/2, b + \boldsymbol{y}' M(\boldsymbol{\alpha}, \boldsymbol{\beta})^{-1}\boldsymbol{y}/2).$$

The parameters, $\boldsymbol{\alpha}, \boldsymbol{\beta}$, define the ACF and the shape of the spectral form. In analogy with the normal mean approach is to add a penalizing term to the likelihood. A natural choice is to multiply the likelihood with a function, h, of the distance between the spectral function, $f(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and a *a priori* spectral function $f(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$. The posterior can be of the form:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}) v^{a-1} e^{-bv} h(KLD(f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma), f^*(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0))).$$

Here KLD is the Kullback-Leibler distance from spectral density f to the function f^*.

$$KLD(f, f^*) = \int_{-\infty}^{\infty} \log(\frac{f(\omega)}{f^*(\omega)}) f(\omega) \mathrm{d}\omega.$$

Here:

$$f(\omega) = \frac{\beta(i\omega)\beta(-i\omega)}{\alpha(i\omega)\alpha(-i\omega)},$$
$$f^*(\omega) = \sigma^* \frac{\beta_0(i\omega)\beta_0(-i\omega)}{\alpha_0(i\omega)\alpha_0(-i\omega)},$$

where σ^* is chosen such that,

$$\int_{-\infty}^{\infty} f(\omega) \mathrm{d}\omega = \int_{-\infty}^{\infty} f^*(\omega) \mathrm{d}\omega.$$

Here the penalty term is based on some kind of a distance measure between the spectral shape defined by the ARMA parameters and an a priory defined spectral shape. In regression it can be based on Ridge-regression or empirical

Bayes approaches such as James-Stein. These penalty terms can be motivated with Bayesian arguments. Technically the ordinary maximum-likelihood is replaced with a penalized maximum-likelihood. The idea is to put a preference on simpler models. The AIC, BIC and the R^2-adjusted all penalize by using a function of the number of estimated parameters. The number of parameters is not always the natural way of grading complexity. In regression it seems reasonable that the model:

$$y = a + bx + e,$$

is simpler than:

$$y = \sin(\cos(ax))^a \exp(-bx)/x^b + e.$$

The ARMA(1,0) model:

$$dY + Y = dW, \quad ,$$

is actually the same as:

$$Y^{(4)} + 4Y^{(3)} + 6Y^{(2)} + 4Y^{(1)} + Y = d(W + 3W^{(1)} + 3W^{(2)} + W^{(3)}.$$

That is the ARMA(1,0) is a special case of (many) ARMA(4,3) models. Estimation of six additional parameters might result in a spectral function with an unreasonable shape. However, it might be of interest to estimate a model which is more complicated than an AR(1). One might, however, want restrict the freedom of the additional parameters.

In time-series analysis, just as in non-parametric regression a smoothness restriction may be enforced on the fitted values. That is the sharp spikes and turns are penalized. In economics a well known procedure of this type is the Hodrick-Prescott filter.

In stationary time-series analysis a natural form of a priori information might consist of a specification of the spectral function or some features of the spectral function. In analogy with the Hodrick-Prescott filter one can introduce a term that penalizes for sharp spikes and turns, e.g., a term proportional to:

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega.$$

For treatment of smoothness prior concepts in time-series analysis, see, e.g., Kitagawa & Gersch (1996).

4 Some computational aspects

A characteristic feature of the spectral density for an ARMA model is that it is a rational function.

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{\boldsymbol{\beta}(i\omega)\boldsymbol{\beta}(-i\omega)}{\boldsymbol{\alpha}(i\omega)\boldsymbol{\alpha}(-i\omega)} = \frac{\sigma^2}{2\pi} \frac{\prod_{j=1}^{q}(1+\mu_j^2\omega^2)}{\prod_{j=1}^{p}(\omega^2+\lambda_j^2)}. \tag{1}$$

Where the λ_j's are the roots of the polynomial $\boldsymbol{\alpha}(z)$, and the μ_j's are the reciprocals of the roots of $\boldsymbol{\beta}(z)$. In the case where the roots of the polynomial $\boldsymbol{\alpha}(z)$ are distinct (and different from the roots of $\boldsymbol{\beta}(z)$, a partial fraction expression of $f(\omega)$ can be given by:

$$f(\omega) = \frac{\sigma^2}{2\pi}(\frac{a_1}{\omega - i\lambda_1} + \cdots \frac{a_p}{w - i\lambda_p} + \frac{b_1}{\omega + i\lambda_1} + \cdots \frac{b_p}{\omega + i\lambda_p}), \tag{2}$$

and another can be given by

$$f(\omega) = \frac{\sigma^2}{2\pi}(\frac{c_1}{\omega^2 + \lambda_1^2} + \cdots + \frac{c_p}{\omega^2 + \lambda_p^2}). \tag{3}$$

In the case of some roots of $\boldsymbol{\alpha}(z)$ being equal, terms of the type $1/(\omega - i\lambda_k)^{m_k}$, where m_k is the multiplicity of the root λ_k, will be present in (2). Both forms of partial fractions are convenient. e.g., the variance due to the frequency interval (ω_1, ω_2) will be given by:

$$\int_{\omega_1}^{\omega_2} f(\omega)d\omega = \frac{\sigma^2}{2\pi}\sum_{j=1}^{p} 1/|\lambda_j|(\arctan(\omega_2/|\lambda_j|) - \arctan(\omega_1/|\lambda_j|)).$$

The residue calculus of complex analysis offers a useful tool for calculating integrals of rational functions. The residue theorem states that

$$\int h(x)dx = 2\pi i \sum Res(h(z)), \quad \text{over a certain path,}$$

where the sum is evaluated over the residues of the function h. For details see e.g., Kreyszig (1999). The theoretical auto-covariance function, $\gamma(\tau) = E(Y(t)Y(t-\tau))$, can be derived by residue calculus:

$$\gamma(\tau) = \int_{-\infty}^{\infty} e^{i\tau\omega} f(\omega)d\omega. \tag{4}$$

As the realpart of λ_j is negative the term $(\omega - i\lambda_j)$ has a pole in the upper negative half-plane, each term in (4) is readily derived:

$$\int_{-\infty}^{\infty} \frac{a_j e^{i\tau\omega}}{\omega - \lambda_j} d\omega = a_j e^{-\lambda_j|\tau|}.$$

Partial fraction of the spectral density can be useful for a variety of descriptive features. E.g. one can define a measure of smoothness:

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega. \tag{5}$$

The expression (5) can be derived directly from (2) by observing that:

$$f''(\omega) = \frac{\sigma^2}{2\pi}(\frac{2a_1}{(\omega - i\lambda_1)^3} + \cdots \frac{2a_p}{(w - i\lambda_p)^3} + \frac{2b_1}{(\omega + i\lambda_1)^3} + \cdots \frac{2b_p}{(\omega + i\lambda_p)^3}),$$

$f''(\omega)^2$ will contain p terms of the type $a_j/(\omega - i\lambda_j)^6$ and p terms $b_k/(\omega + i\lambda_k)^6$ and $p(p-1)$ terms, $k \neq j$, of the type $a_k a_j/((\omega - i\lambda_k)^3)(\omega - i\lambda_j)^3)$ and similarly $p(p-1)$ terms, $k \neq j$, $b_k b_j/((\omega + i\lambda_j)(\omega + i\lambda_j))$. The residues in the upper half-plane of these terms sums to zero. The integral will be the sum of the $2p^2$ residue terms of the type

$$a_k b_j/((\omega - i\lambda_k)^3(\omega + i\lambda_j)^3).$$

The residues of these terms are of the form:

$$3 \cdot 4 a_k b_j/(-(i\lambda_k + i\lambda_j)^5,$$

and the integral therefore,

$$\int_{-\infty}^{\infty} (f''(\omega))^2 d\omega = 2\pi i \cdot 2 \sum_{k=1}^{p} \sum_{j=1}^{p} 3 \cdot 4 \frac{a_k b_j}{-(i\lambda_k + i\lambda_j)^5}.$$

Similarly one can use residue calculus to create of measure of steep hills in the spectrum,

$$\int_{-\infty}^{\infty} (f'(\omega))^2 d\omega,$$

or weighing $(f'(\omega))^2$ or $(f''(\omega))^2$ with a rational function.

The partial fraction trick can also be applied to calculate the Kullback-Leibner (KL) metric, as a measure of the distance between two functions, f and f_0 (e.g. a prior).

$$D(f_1; f_0) = \int f_1(\omega) \log(f_1(\omega)) d\omega - \int f_1(\omega) \log(f_0(\omega)) d\omega. \tag{6}$$

Using a partial fraction formulation of the spectral density the terms in (6) that need to be integrated will be of the form:

$$-\frac{c_{1,k}}{(\omega^2 + \lambda_{1,k}^2)} \log(w^2 + \lambda_{1,j}^2), \text{ and } \frac{c_{1,j}}{(\omega^2 + \lambda_{1,k}^2)} \log(1 + \mu_{1,j}^2 w^2).$$

Gradshteyn & Ryzhik (2007, eq 1, page 560) give the result:

$$\int_0^{\infty} \log(1 + \mu^2 x^2) \frac{dx}{x^2 + \lambda^2} = \frac{\pi}{\sqrt{\lambda^2}} \log(\sqrt{\lambda^2 \mu^2} + 1). \tag{7}$$

Here the square-root is take such that the real part of the square-root is positive. By use of partial fractions the KL distance can be written as:

$$\int \sum_{k=1}^{p_1} \frac{c_{1,k}}{\omega^2 + \lambda_{1,k}^2} (\sum_{j=1}^{q_1} \log(1 + \mu_{1,j}^2 \omega^2)) d\omega - \int \sum_{k=1}^{p_1} \frac{c_{1,k}}{\omega^2 + \lambda_{1,k}^2} (\sum_{j=1}^{p_1} \log(\omega^2 + \lambda_{1,j}^2)) d\omega -$$
$$\int \sum_{k=1}^{p_1} \frac{c_{1,k}}{\omega^2 + \lambda_{1,k}^2} (\sum_{j=1}^{q_0} \log(1 + \mu_{0,j}^2 \omega^2)) d\omega + \int \sum_{k=1}^{p_1} \frac{c_{1,k}}{\omega^2 + \lambda_{1,k}^2} (\sum_{j=1}^{p_0} \log(\omega^2 + \lambda_{0,j}^2)) d\omega.$$

This integral consists of $p_1 \times q_1 + p_1^2 + p_1 \times q_0 + p_1 \times p_0$ terms and each of them can be calculated by the use of (7),

$$\pi(\sum_{k=1}^{p_1}\sum_{j=1}^{q_1}\frac{c_{1,k}\log(1+\lambda_{1,k}\mu_{1,j})}{\lambda_{1,k}} - \sum_{k=1}^{p_1}\sum_{j=1}^{p_1}\frac{c_{1,k}\log(1+\lambda_{1,j}\lambda_{1,k})}{\lambda_{1,j}} + \sum_{k=1}^{p_1}\sum_{j=1}^{q_0}\frac{c_{1,k}\log(1+\lambda_{1,k}\mu_{0,j})}{\lambda_{0,j}} - \sum_{k=1}^{p_1}\sum_{j=1}^{p_0}\frac{c_{1,k}\log(1+\lambda_{1,k}\lambda_{0,j})}{\lambda_{0,j}}).$$

Given the partial fractions of (3) these terms are all known. An analytical result of Kullback-Leibner distance, (6), can be obtained by applying (7) to the $p \times (q + p)$ terms in the formula above. All that is needed are the roots of the polynomials.

5 Implementation in R

A practical partial-fraction algorithm has been implemented in a R-package `ctarmaRcpp`, which is a `Rcpp` version of the `ctarma` packages used for the computations described in Tómasson (2015). For calculation of the partial fractions in (2) and (3) a algorithm based on Chen & Leung (1981) was implemented in the R function `partfrac1`. For each estimated model the roots of the AR and MA part are found and then various measures can be calculated. E.g.

$$1/(6 + 11x + 6x^2 + x^3) = \frac{1}{2(x+3)} - \frac{1}{x+2} + \frac{1}{2(x+1)},$$

here the roots are -1,-2,-3, and the function `partfrac1` gives the coefficients in the partial fraction (all roots distinct).

```
partfrac1(c(6,11,6,1),1,c(-1,-2,-3),1)
[1]  0.5 -1.0  0.5
```

The partial fraction in (3) enables the calculation of the Kullback-Leibler distance between two spectral shapes. A data set on the Earth's temperature for the past 800.000 years is used as an illustration on an unevenly sampled time series. The `ctarmaRcpp` package bundles data and model into a R object. The maximized log-likelihood of a continuous-time ARMA(2,1) is contained in `m2e`. The log-likelihood of `m2e` is calculate by:

```
> ctarma.loglik(m2e)
[1] -5701.584
```

An ARMA(4,3) gives log-likelihood of -5664.627, and an ARMA(6,5) a log-likelihood of -5660.819. The coefficients of the estimated ARMA(2,1), are

```
[1] 1792.32808   13.39429
> m2e$bhat
[1] 1.00000000 0.02315723
> m2e$sigma
[1] 1331.322
```

Similarly the estimated coefficients of the ARMA(4,3) are:

```
> m4e$ahat
[1] 1497.15420 3410.91710 2328.64602   28.11924
> m4e$bhat
[1] 1.0000000 1.2087125 0.3772288 0.0128648
> m4e$sigma
[1] 2239.939
```

The Kullback-Leibler distance is calculated with the function `kullbackDist` (here the implementation is between spectral shapes).

```
> kullbackDist(m4e$ahat,m4e$bhat,m4e$sigma,m2e$ahat,m2e$bhat)
[1] 1.172553
```

and for the ARMA(6,5)

```
> kullbackDist(m6e$ahat,m6e$bhat,m6e$sigma,m2e$ahat,m2e$bhat)
[1] 3.706201
```

The generalization to more complicated models is straightforeward.

6 Discussion

In this paper has shown an application of partial fractions and residue calculus to calculate measures of complexity of the spectral functions. The motivation of this measures is that the number of parameters, such as AIC and BIC, may not have the desired properties. The approach offers a way to measure the distance between functions, and a measure of features such as smoothness of a spectral function. The approach is based on the fact that the spectral function of an ARMA model is a rational function. The approach in this paper is based on continous-time ARMA but the arguments carry directly over to the discrete-time ARMA. The derivation of measures of the spectral features boil down to calculation of roots of polynomials. The ideas described allow the expression of many forms of penalty terms, e.g., a priori formulations of the spectral function.

Bibliography

Bergstrom, A. R. (1988). The history of continuous-time econometric models. *Econometric theory*, *4*, 365–383.

Chen, C. & Leung, K. (1981). A new look at partial fraction expansion from a high-level language viewpoint. *Computers & Mathematics with Applications*, *7*(5), 361 – 367.

Gradshteyn, I. S. & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (Seventh ed.). Elsevier/Academic Press, Amsterdam. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

Kitagawa, G. & Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*. Springer-Verlag New York.

Kreyszig, E. (1999). *Advanced Engineering Mathematics* (8 ed.). John Wiley & Sons. Residue theorem, page 723-724.

Tómasson, H. (2015). Some computational aspects of gaussian CARMA modelling. *Statistics and Computing*, *25*(2), 375–387.

Tsay, R. S. (2010). *Analysis of Financial Time Series* (Third ed.). John Wiley & Sons.

Zadrozny, P. (1988). Gaussian likelihood of continuous-time ARMAX models when data are stocks and flows at different frequencies. *Econometric Theory*, *4*(1), 108–124.

Towards an API for EEG-Based Imagined Speech classification

Luis Alfredo Moctezuma[1], Marta Molinas[1], A. A. Torres García[2], L. Villaseñor Pineda[2], and Maya Carrillo[3]

[1] Department of Engineering Cybernetics, Norwegian University of Science and Technology. Trondheim, Norway
luisalfredomoctezuma@gmail.com, marta.molinas@ntnu.no
[2] Computer Science Department, Instituto Nacional de Astrofísica Optica y Electrónica. Puebla, Mexico
{alejandro.torres, villasen}@ccc.inaoep.mx
[3] Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla. Puebla, Mexico
cmaya@cs.buap.mx

Abstract. In this paper, imagined speech classification is performed with an implementation in Python and using scikit-learn library, to create a toolbox intended for real-time classification. To this aim, the Discrete Wavelet Transform with the mother function Biorthogonal 2.2 is used to then compute the *instantaneous* and *Teager* energy distribution for feature extraction. Then, *random forest* is implemented as a classifier with 10-folds cross-validation. The set of experiments consists of *imagined speech classification, linguistic activity and inactivity classification* and *subjects identification*. The experiments were performed using a dataset of 27 subjects which imagined 33 repetitions of 5 words in Spanish *up, down, left, right and select*. The accuracy obtained with the models were 0.77, 0.78 and 0.98 respectively for each task. The high accuracy rates obtained as a result attest for the feasibility of the proposed method for subject identification.

Keywords: Imagined Speech, Linguistic activity, Subjects identification, Discrete Wavelet Transform, Brain Computer Interfaces, Electroencephalograms, Application Programming Interface

1 Introduction

In the last years, exploration into the identification of various brain activities has considerably increased, motivated by its potential use as a new way of communication, with applications ranging from games to medicine and in general to augmented human capabilities. A brain computer interface (BCI) is a communication system that monitors the brain activity and translates features, corresponding to the user's intention/thoughts/movements into control commands.

When thoughts features are identified and extracted, this can become a new way of human-machine interaction that allows users to employ their thoughts to

control/operate external devices. BCI techniques can be classified into invasive and non-invasive, the first one requiring surgical procedures for which very clear brain signals are obtained because the measurements are not attenuated by the skull and the scalp. Non-invasive BCI, does not require any surgery but signals obtained are weaker. Non-invasive BCI are the most used due to their relatively low cost and easy setup.

Neurophysiological exploration based on the bioelectrical brain activity (Oscillations of brain electric potentials, frequency spectrum in Hz) registered during unconstrained rest, sleep, or with different activation functions are called Electroencephalography (EEG). Electrophysiological sources refer to the neurological mechanisms (also known as neuro-paradigm) used by a BCI user to generate control signals [1]. Wolpaw et al. [2] and Bashati et al. [1] separated these electrophysiological sources into categories based on neuronal mechanisms and the recording technology they use. There are some invasive categories, but for the interest of this work the non-invasive categories are: sensory-motor activity, potential P300, visual evoked potentials (VEP), slow cortical potentials (SCP) and response to mental and multiple tasks neuromechanisms.

Later [3] added the imagined or internal speech; which refers to the internal or imagined pronunciation of words but without uttering sounds and without articulating gestures. Imagined speech as an electrophysiological source has advantages over others, because it needs little training. In this research, imagined speech from EEG signals are employed, the dataset consisting of EEG signals from 27 subjects captured while imagining 33 repetitions of five words in Spanish; *up, down, left, right* and *select*.

The state-of-the-art reports the use of imagined speech as Electrophysiological source, but in the majority of the cases they are limited to reduced vocabularies (yes or no, etc.) or limited to syllables/phonemes. However, the applications can be limited in these cases. If instead of syllables/phonemes are used complete words, the applications can range from medical to biometric systems. The work in this paper is aimed at a new way of communication in real-time between people who cannot produce sounds or with specific diseases like Amyotrophic Lateral Sclerosis [4,5].

Although the reported literature includes research employing imagined speech, the on-line/real-time process is not considered. For a real-time implementation it is first important to identify if the signals correspond to linguistic activity and once the linguistic segment is identified in a signal, another process using a multi-class classifier can determine the imagined word associated to the respective signal.

In this area, there is need for efforts to create a method for *transfer learning* [6], because in real applications people in need of BCI solutions may have some difficulties to train a model anew. The idea of *transfer learning* is to create a classifier with training from a group of subjects and use in a different group of subjects. In [7], the authors reported experiments and propose that a first stage of calibration is needed because the signals are sufficiently different between subjects and between sessions. Other experiments also suggest the use imagined

speech from EEG data as a biometric measurement for subject identification [8], and show experiments limited to syllables. Security systems used by organizations to manage access to facilities, equipment or resources and to protect against theft or espionage by denying unauthorized access, is one of the first applications envisioned by this BCI concept. Different types of safety measures have been proposed and used for a long time, ranging from the use of standard systems (security guards, smart cards, etc) to the use of biometric measurements (fingerprint, palm-print, etc.). A biometric recognition system is able to perform automatic recognition of subjects based on their physiological and/or behavioral features[9]. Any human physiological and/or behavioral characteristic can be used as a biometric characteristic as long as it satisfies the following requirements: *universality, permanence, collectability, performance, acceptability and circumvention*.

Biometric systems are advantageous compared to generic system, because they are more difficult to steal, compromise or duplicate, and can be more convenient for the users since a single biometric trait can be used for the access into several accounts. However, current biometric systems are vulnerable to attacks aimed at undermining the integrity of the authentication process [10]. For example, an intruder may fraudulently obtain the latent fingerprints of a user and later use it to construct a digital or physical artifact of the user's finger [11].

Building on the above existing knowledge and on the need for inviolable methods for subject identification, this paper proposes to advance the research by employing imagined speech from EEG data to identify subjects in real-time. For the experiments, an on-line environment (simulated real-time) was used, as it is explained in the following sections.

The paper is organized as follow; first, the proposed method for different tasks are presented and explained in brief. Next, some experiments to show the application of the proposed method in *imagined speech classification, linguistic distinction* and *subjects identification* are described, to finally discuss the feasibility of the method in real application and possible future improvements.

2 Proposed method

In general, the method can be summarized in 3 fundamental steps: Pre-processing, feature extraction and classification, as described in the flowchart in figure 1. To ensure that instances in training data (off-line) are not used in Test data (on-line), 30% was first separated for *Test Set* and 70% for training and to create the model using 10-folds cross-validation.

To implement the proposed method and used it in real-time, it was created an Application Programming Interface (API) using *Django* with Python 2.7 and to store the models per each task and for all experiments a database in MySQL was used.

In this version of the project the API[4] consist of 2 EndPoints; *dwt/training* and *dwt/{model_id}* with the POST method according to HTTP methods [12].

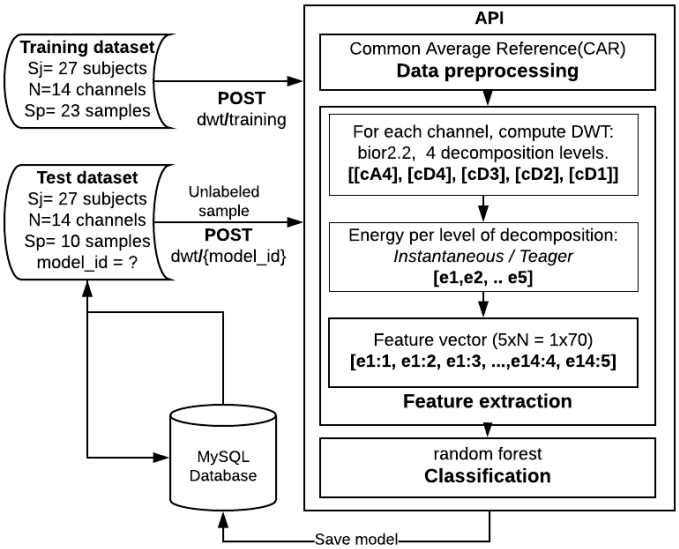


Fig. 1. Flowchart summarizing the steps of the proposed method.

2.1 Dataset

The dataset consist of EEG signals from 27 subjects captured using EMOTIV EPOC while imagining 33 repetitions of five imagined words in Spanish; up, down, left, right and select (Corresponding to *arriba, abajo, izquierda, derecha* and *seleccion*). Each repetition of the imagined words was separated by a state of rest, as shown in figure 2 and described in [13].

EEG signals were recorded from 14 high resolution channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4; see figure 3) with a sample frequency of 128 Hz which were placed according to the 10-20 international system [14].

2.2 Pre-processing

In order to reduce the signal-to-noise ratio, the common average reference (CAR) method was used [15,16]. As we can see in the formula 1, the CAR method remove the common data in all electrodes recorded simultaneously.

[4] For more information about this public project, visit: `https://github.com/wavesresearch/eeg_api_dwt`

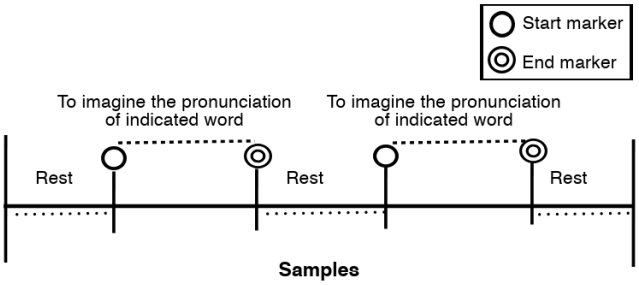


Fig. 2. Protocol designed in [13] for EEG signal acquisition using EMOTIV EPOC.

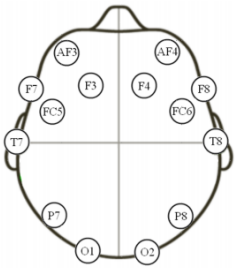


Fig. 3. 10-20 international system for 14 channels [14].

$$V_i^{CAR} = V_i^{ER} - \frac{1}{n}\sum_{j=1}^{n} V_j^{ER} \tag{1}$$

Where V_i^{ER} is the potential between the ith electrode and the reference, and n is the number of electrodes.

2.3 Feature extraction

The EEG signals are usually non-stationary, they change rapidly over time and patterns of brain activity contain information related to specific variations over time. A representation of the signal that considers this behavior is necessary for a proper feature extraction.

When the DWT is applied to a signals S with a decomposition level j=4, it will give a structure with vectors of approximation cA_j and detail cD_j coefficients: $[cA_4, cD_4, cD_3, cD_2, cD_1]$, as shown in the figure 4 and the table 1 shows the related frequencies per level of decomposition.

According to the average size of the dataset, 4 decomposition levels were applied. However, the wavelet coefficients for each decomposition level will vary depending on the signal size(the duration of imagined pronunciation of the word, between imagined words from the same subject and between imagined words from different subjects).

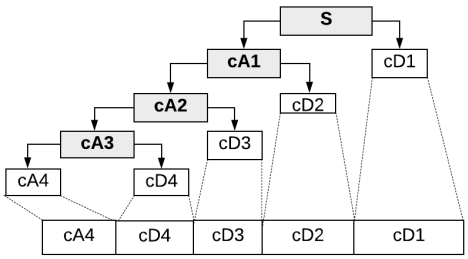


Fig. 4. Coefficient vectors in the 4th decomposition level of DWT for a signal S.

Table 1. DWT with 4 decomposition levels, frequencies ranges and related brain rhythms.

Level	Frequency range	Brain rhythm
cD1	32-64	*Gamma*
cD2	16-32	*Beta (16-30 Hz) and Gamma (30-32 Hz)*
cD3	8-16	*Alpha (8-12 Hz) and Beta (12-16 Hz)*
cD4	4-8	*Theta*
cA4	0-4	*Delta*

To deal with this problem the *instantaneous* and *teager* energy distribution was calculated [17]. These energy distributions were used since they have shown the best results related to imagined speech [18,7]. When energy coefficients are calculated, it is possible to have the same number of features for all instances.

In this work, the feature vector for each instance was represented with energy coefficients that were calculated for each decomposition level of the DWT biorthogonal 2.2(bior2.2) and for each channel which were then concatenated in order to have a single feature vector. The expressions for these energy distributions are shown below:

– Instantaneous: gives the energy distribution in each band [17]:

$$f_j = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}(w_j(r))^2\right) \tag{2}$$

– Teager: This energy operator reflects variations in both amplitude and frequency of the signal and it is a robust parameter for speech recognition as it attenuates auditory noise [19,17].

$$f_j = log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j-1}|(w_j(r))^2 - w_j(r-1) * w_j(r+1)|\right) \tag{3}$$

At this point, instead of having a features vector for each decomposition level we have a single value for each one, and the process is repeated for each channel.

After this process, we have 5 values per channel $(CA_4, CD_4, CD_3, CD_2, CD_1)$ and then all 14 channels are concatenated in order to have a single feature vector with 70 coefficients to represent each instance of the EEG signal.

2.4 Classification

Once feature vectors are compute and obtained for each instance of the EEG signal, *random forest* was used for automatic classification with an implementation in Python 2.7 using the library scikit-learn [20]. For all experiments, the parameters for *random forest* in scikit-learn were: $max\_depth = 5, random\_state = 0, criterion = gini$, that were selected after tested all possibilities. This classifier was selected because of the good results reported in the authors' previous work on imagined speech classification using EEG signals [13,18].

To evaluate the classifier performance with 10-folds cross-validation, an accuracy index was defined and calculated.

3 Experiments and results

In this paper an implementation of the proposed method towards on-line BCI for identification of imagined speech from EEG signals is developed.

The results from several experiments using DWT bior2.2 with Instantaneous and Teager energy distribution are reported here. For all described experiments the models were created off-line and tested on-line, using the API created. The experiments were separated in 3 different groups. First we present the result of using the proposed method for imagined speech classification. Then, we separate the imagined speech into a class named *linguistic activity* and the states of rest into a class named *linguistic inactivity*, in order to check if the proposed method can distinguish the imagined speech from others activities (rest). In addition the signals were separated and tagged with a subject ID (S1, S2, .. S27) to create a *machine learning* model for subjects identification and them use it in real-time. It were made with subject-level analysis and subject-word-level analysis as shown in the next experiments.

3.1 Imagined speech classification

Once the feature extraction was applied to the EEG signals, a classifier for each subject using *random forest* was created. For each subject, the classifier consists of 5 classes corresponding to the 5 imagined words (up, down, left, right and select), and for each imagined word 23 instances were used.

The average accuracy for imagined speech classification task is presented in figure 5, where it can be noted that when the Teager energy distribution was used, the highest accuracy was reached (0.77).

In this case, the model saved in the database to use it in real-time was with instantaneous energy values. Using the trained model in on-line environment with the 30% of the instances, the average accuracy obtained was **0.85**, it can be noted that the accuracy is highly related to the accuracy of the model.

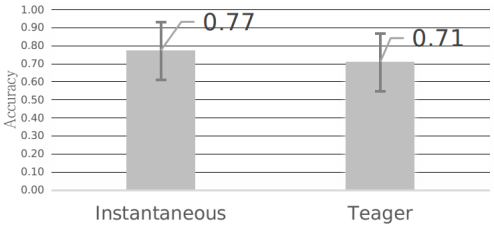


Fig. 5. Average accuracy and standard deviation obtained with 27 subjects for imagined speech classification.

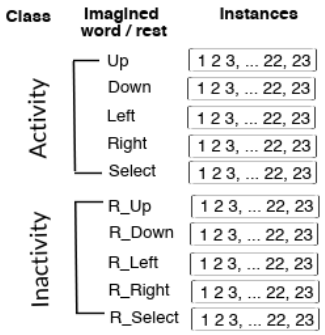


Fig. 6. Procedure to separate instances into linguistic activity and inactivity.

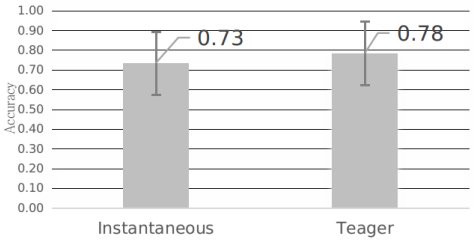


Fig. 7. Average accuracy and standard deviation obtained with 27 subjects for Linguistic activity and inactivity distinction.

3.2 Linguistic activity and linguistic inactivity

In a complete real-time application (In a continuous recording without restrictions) it is necessary to distinguish the brain activity generated by the subject when imagining a word (linguistic activity) from any other brain activity (non-linguistic activity or inactivity). In this part the "complete real-time application" refer to identify the linguistic activity segment and then use another classifier to detect the specific imagined word.

In this experiment the EEG signals were separated into 2 classes, a set of imagined words considering the class of linguistic activity; and states of rest (unconstrained rest) as examples of other brain activity. In this work, the latter are called *linguistic inactivity*. The process to separate the instances into 2 classes is shown in the figure 6, where *R_up, R_Down, R_Left, R_Rigth, R_Select* correspond to rest states between each repetitions.

The model saved in the database was with the instantaneous energy distribution, and in the test stage the average accuracy obtained was **0.78**.

3.3 Subjects identification

This experiment was carried out to check if there are sufficient information in the EEG data for this task and to create a model for subject identification. For this, the 115 instances of imagined words per subject (corresponding to 23 repetitions per each of 5 imagined words) were considered as a single class, tagged it with a subject ID (S1, S2, ..., S27).

This experiment was performed with 27 subjects using Instantaneous and Teager energy distribution based on the DWT bior2.2. The results obtained in the classification step with 10-fold cross-validation with *random forest* are shown in the table 8.

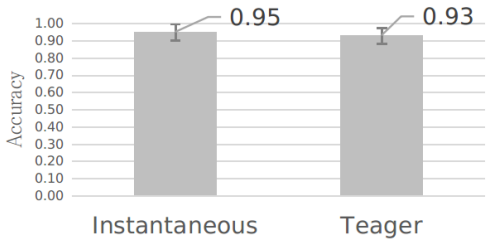


Fig. 8. Accuracy obtained when a classifier with all imagined words as a single class was created with 27 subjects using 10-fold cross-validation.

In figure 8 it can be observed that when using instantaneous energy distribution the best accuracy obtained is 0.95. However using the Teager energy the accuracy is similar and we can use them both. However, the saved model was with instantaneous energy, and after the use of *Test Set* the average accuracy obtained was 0.92.

This result suggests that it is possible to identify subjects regardless of the word that the subject is imagining. This means that a subject can be identified using different words. This brings up the question whether there a specific word that is best suited for subjects identification.

To answer this question, the following experiment was carried out, which consists of testing the classification with *random forest* with all 27 subjects but with a classifier per imagined word. For this, the 23 repetitions of each imagined word were used separately and the experiment was repeated for the 5 words from the dataset of EEG signals. The classification for each of the imagined words was done with the two feature extraction ways used in the previous experiment (Instantaneous and Teager) in order to compare their strengths. The results obtained in the classification step with 10-fold cross-validation using *random forest* are shown in figure 9.

Figure 9 again shows that when using instantaneous energy, the accuracy is higher. The highest accuracy is obtained when using the imagined word *Down*. From this, it can preliminarily be asserted that the most suitable word for the

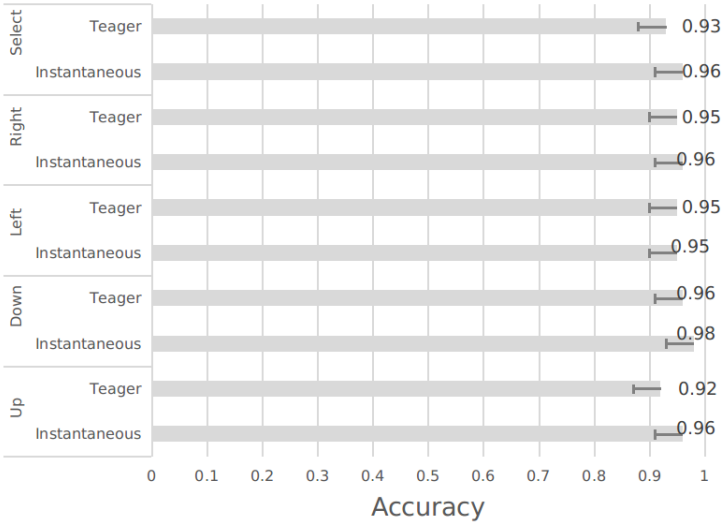


Fig. 9. Accuracies obtained when classifiers for each word separately were created with 27 subjects using 10-fold cross-validation.

task of subjects identification is the imagined word *Down*. However, when using the other imagined words, the results are not much different and in all cases they are above 0.92 of accuracy.

For subjects identification task, the models saved for each imagined word was when the instantaneous energy distribution was used in all cases. As in the proposed was described, for the test stage was used the 30% of the instances (In this case 270 instances per experiment; corresponding to 10 imagined words per each of 27 subjects) for on-line environment. The average accuracy per imagined word was **0.92, 0.91, 0.92, 0.95** and **0.90** respectively.

4 Discussion and Conclusions

I this work, EEG-recorded dataset using the neurological source *imagined speech* were used to create machine learning-based models. To use the method in real-time, an API using Django in Python was designed and created.

The accuracy obtained for on-line environment, is highly related to the model created. In the best-case the average accuracy obtained with the *Test Set* was 0.95. This accuracy level was obtained in the subjects identification task using the imagined word *Down*.

The time for training and saving a machine learning model is not important since it is an off-line process. Using the API for real-time classification, the time to process a new unlabeled entry depends on the size of the signal and the DWT computational complexity ($\mathcal{O}(N \log_2 N)$) [21], however for fast responses it is necessary to implement techniques to distribute the memory/work.

In general, as the experiments in this paper report, EEG signal as a new way of communication can be possible and the work to process new signals can be separated using an API (useful for several predictions at the same time). In addition, EEG signals can be used as a password or as a measure for a biometric security system for several environments. The first experiment for subjects identification shows that subjects can be identified independent of the imagined word. This suggests that it is possible to use a classifier to detect subjects and then use the classified imagined word as a control command (using the API only with other model_id corresponding to the application) in a real application. For example, it can be used as a 2-step verification or to send 2 or more commands at the same time (i.e Give access and call the police, Give access and turn on the lights, etc), in summary, for domestic security applications.

Distributing the work in an API has benefits and will contribute to the use of a single machine learning model for several applications, for example, for subjects identification, the same model could be used to manage the access in 2 or more places and for several users and different tasks. In addition, it could accelerate the process of classification because there are lower restrictions when using computers (I.e High Performance Servers and Supercomputers).

Future research efforts will be dedicated to explore the extent to which specific channels can provide more information for these task in order to reduce the number of channels for real-time applications and decrease the time for a new unlabeled entry. In addition, implementation in a real environment (with additional noise) and with new feature extraction techniques will be tested.

Acknowledgments. This work was done under partial support of CONACYT (scholarship #591475), and the project "David versus Goliath: single-channel EEG unravels its power through adaptive signal analysis - FlexEEG" which is supported by Enabling Technologies - NTNU.

Conflict of interest. The subjects identification experiment is part of a patent.

References

1. Bashashati, Ali, Mehrdad Fatourechi, Rabab K. Ward, and Gary E. Birch: A survey of signal processing algorithms in braincomputer interfaces based on electrical brain signals. Journal of Neural engineering **4**, no. 2 (2007): R32.
2. Wolpaw, Jonathan R., Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan: Brain-computer interfaces for communication and control. Clinical neurophysiology **113**, no. 6 (2002): 767-791.
3. Desain, Peter, Jason Farquhar, Pim Haselager, Christian Hesse, and R. S. Schaefer: What BCI research needs. In Proc. ACM CHI 2008 Conf. on Human Factors in Computing Systems (Venice, Italy) (2008):.
4. Elman, Lauren B., and L. McCluskey: Clinical features of amyotrophic lateral sclerosis and other forms of motor neuron disease. Up-to-date. Waltham: Wolters Kluwer Health (2012): 23.

5. Feller, T. G., R. E. Jones, and M. G. Netsky: Amyotrophic lateral sclerosis and sensory changes. Virginia medical monthly **93** , no. 6 (1966): 328.
6. Jayaram, Vinay, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup: Transfer learning in brain-computer interfaces. IEEE Computational Intelligence Magazine **11**, no. 1 (2016): 20-31
7. Moctezuma, Luis Alfredo: Distinción de estados de actividad e inactividad lingüística para interfaces cerebro computadora. Thesis project of master degree (2017):.
8. Brigham, Katharine, and BVK Vijaya Kumar: Subject identification from electroencephalogram (EEG) signals during imagined speech. In Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on (2010): 1-8
9. Jain, Anil K., Arun Ross, and Salil Prabhakar: An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology **14**, no. 1 (2004): 4-20.
10. Jain, Anil K., Arun Ross, and Umut Uludag: Biometric template security: Challenges and solutions. Signal Processing Conference 13th European IEEE (2005): 1-4.
11. Uludag, Umut, and Anil K. Jain: Attacks on biometric systems: a case study in fingerprints. In Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306, International Society for Optics and Photonics (2004): 622-634
12. Fielding, Roy, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee: Hypertext transfer protocol–HTTP/1.1. No. RFC 2616. (1999):.
13. Torres-García, A. A., C. A. Reyes-García, L. Villaseñor-Pineda, and J. M. Ramírez-Cortís: Análisis de señales electroencefalográficas para la clasificacin de habla imaginada. Revista mexicana de ingeniería biomédica **34**, no. 1 (2013): 23-39.
14. Jasper, Herbert: Report of the committee on methods of clinical examination in electroencephalography. Electroencephalogr Clin Neurophysiol **10** (1958): 370-375.
15. Bertrand, O., F. Perrin, and J. Pernier: A theoretical justification of the average reference in topographic evoked potential studies. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 62, no. 6 (1985): 462-464.
16. Alhaddad, Mohammed J: Common average reference (car) improves p300 speller. International Journal of Engineering and Technology **2**, no. 3 (2012): 21.
17. Didiot, Emmanuel, Irina Illina, Dominique Fohr, and Odile Mella: A wavelet-based parameterization for speechmusic discrimination. Computer Speech & Language **24**, no. 2 (2010): 341-357.
18. Moctezuma, Luis Alfredo, Maya Carrillo, Luis Villaseñor Pineda, and Alejandro A. Torres García: Hacia la clasificación de actividad e inactividad lingüistica a partir de senales de electroencefalogramas (EEG). Research in Computing Science **140** (2017): 135-149.
19. Jabloun, Firas, and A. Enis Cetin: The Teager energy based feature parameters for robust speech recognition in car noise. In Acoustics, Speech, and Signal Processing. 1999 IEEE International Conference on, vol. 1. (1999): 273-276
20. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al.: Scikit-learn: Machine learning in Python. Journal of machine learning research **12**, no. Oct (2011): 2825-2830.
21. Averbuch, Amir Z., and Valery A. Zheludev: Construction of biorthogonal discrete wavelet transforms using interpolatory splines. Applied and Computational Harmonic Analysis **12**, no. 1 (2002): 25-56.

A SIMULATION OF A CUSTOM INSPECTION IN THE AIRPORT

Kalle Saastamoinen*, Petteri Mattila and Antti Rissanen

Department of Military Technology, National Defence University,
P.O. Box 7, FI-00861 Helsinki, Finland
{kalle.saastamoinen,petteri.mattila,antti.rissanen}@mil.fi
http://www.puolustusvoimat.fi/en/

Abstract. Time is an essential part when travelling, since it is a very time-critical process. One main issues when going abroad is a border crossing. Helsinki-Vantaa Airport (HEL) has extensive flight connections and the shortest routes between Europe and Asia, making it a major hub for Northern Europe. The number of external border traffic at Helsinki-Vantaa Airport is increasing steadily.

In this study we will study through simulation if Helsinki-Vantaa Airport is able to answer future challenge with estimated amount of passenger to increase 74 % by 2022 by using cooperation-based data processing in entry and exit checks in the border inspection.

Main result of simulation is that by exchanging traveler's information automatically between different authorities, this future increased amount of traffic can be taken care of without increase of airport capacity, so that airport traffic still stays fluent. This simulation model can be used if we want to measure impacts of renovations and increased traveler's flows to the changes of border inspection time.

Keywords: Airport, Passenger information, Border inspection, Data, Simulation

1 Introduction

Here simulation is used to show the real effects of alternative conditions and courses of action. In order to establish our simulation model, we used queueing theory that is the mathematical study of waiting lines, or queues [1]. This study can be considered to be part of operations research since the simulation results are used to make decisions about the resources needed to provide a service.

In modern traveling, airport operations and the implementation of border checks are very time-critical. Short connecting flight schedules at Helsinki-Vantaa Airport, coupled with high level of the passengers passing through the airport to the next flight, require smoothly running

adfa, p. 1, 2011.

border control process. This demand challenges the border inspection capability to be flexible and able to fill the level required for a fluent border traffic. Passengers' delays from connecting flights can be a significant operational and financial risk for airlines and ultimately passengers themselves. The border inspection authority has no legal right to waive border controls or to change the content of inspections to ensure the smooth running of traffic. The flow of traffic cannot get priority over the border control protocols that are in line with the regulations. However, it is very important that the border inspection process is operated without significant queuing times. [2]

Main research question is to simulate how well can Helsinki-Vantaa Airport answer future challenge with increasing number of passengers and can new demands be handled using cooperation-based data processing with different participating operators? Where main benefits are avoidance of overlapping information and use of automated border checks in entry and exit.

In this article first chapter presents the problem in hand at Helsinki-Vantaa Airport, second chapter goes through possibilities of collaboration-based data-handling, third chapter presents the simulation model and results and fourth chapter gives conclusions and discusses about future directions of this research.

2 Problem description

The development of Helsinki-Vantaa Airport emphasizes the growth of passenger traffic across the Schengen border, which has a direct impact on the Border Guard's operations as well. Estimated growth forecast for traffic will take into account Finavia's (Formerly the Finnish Civil Aviation Administration, that is responsible for maintaining and developing its 21 airports and Finland's air navigation system) and Finnair plans and projections. The largest operator in the HEL airport is Finnair, which invests in traffic between Europe and Asia. Currently Asian traffic accounts for a little over 50 % of Finnair’s traffic. [3] Finavia's future growth strategy underlines the strong growth of transit travelers. In Helsinki-Vantaa this means that Asian passenger traffic will increase, which significantly adds the number of passengers with a visa requirement [4].

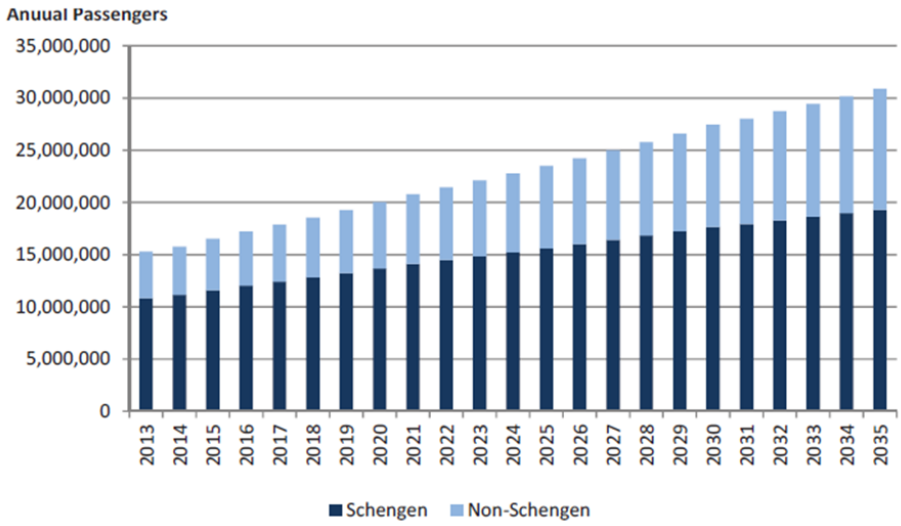


Table 1. Estimated passenger growth between 2013 and 2035. [5]

Helsinki-Vantaa Airport has extensive flight connections and the shortest routes between Europe and Asia, making it a major hub for Northern Europe. When planning for a smooth running of the border control process it is good to know that many of the passengers pass through the airport only to switch the flight. The number of external border traffic at Helsinki-Vantaa Airport is increasing steadily. Table 1 shows estimation for number of passengers up to the year 2035. Estimation forecasts an average growth rate of 4.7 percent per annum. Traffic growth is expected to be particularly high on third-country passengers.

Table 2 shows the realization and development of cross-border traffic in the Helsinki-Vantaa external border traffic 2010-2017. In 2016, Helsinki-Vantaa's total passenger traffic was 17.2 million passengers (Table 2). During the year 2017 the growth was 800 000 passengers. By the year 2022, the number of passengers on external border traffic is estimated to increase approximately by 74 per cent, from 4.6 million (2016) to 8 million passengers. In 2017 more than 5.2 million border checks were carried out. [6]

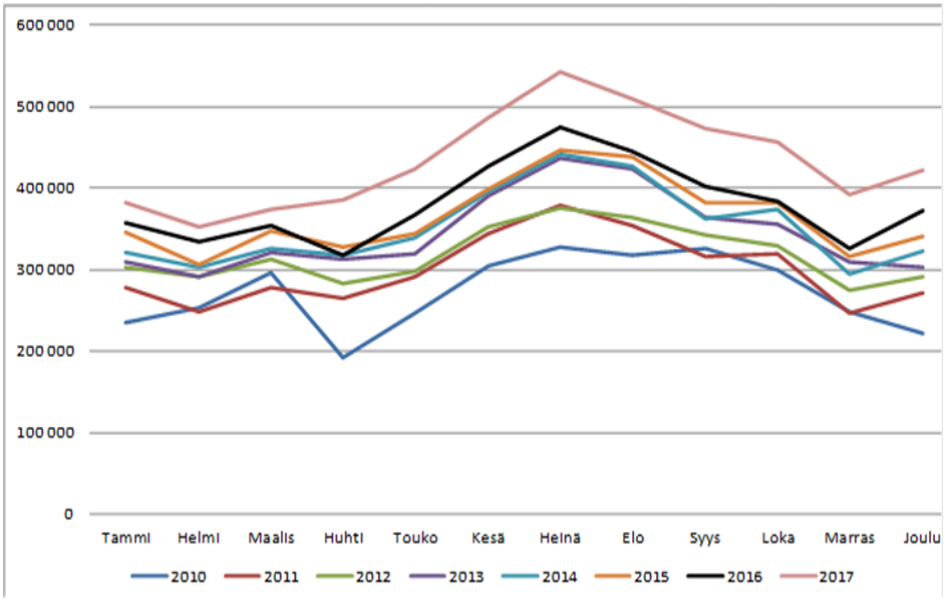


Table 2. Boarder crossing actual values and development in 2010- 2017 [7].

The Head of the Coast Guard of the Gulf of Finland has pointed out that Helsinki-Vantaa Airport's volume and profile of the passenger traffic, especially during the afternoon hours, pose a challenge for the operators, airline carriers and authorities. [8] To respond to the challenge, the Border Guard's operational processes are continuously evaluated and developed. Efforts to streamline and manage passenger traffic have been attempted to meet the needs of growing passenger flows. These actions include the efficient use of personnel and passenger guidance as well as other development measures for the border inspection process. The passenger guidance system has been developed as a co-operation between the airport authority and the Border Guard. [9]

In the actual border checks, processes have been continuously enhanced. As an example, automatic self-service border control desks in the Helsinki-Vantaa Airport have helped to speed up the increase in passenger flows. The major factor in the automated border control utilization is the increase in the proportion of biometric passports. Safety regulation and features for biometrics passports and travel documents are provided at EU level. [10] As such they made an important step towards the development of technical components at border control in Europe.

The share of EU / EEA / CH nationals in the Helsinki-Vantaa external border traffic has been steadily about 70% of the total number of passengers traveling each year. In 2017, EU passengers accounted for around 66% of Helsinki-Vantaa's external border traffic. The number of automatic border checks on total traffic has still been below 40%. [7] However the predicted growth of passenger traffic in the group of third-country nationals means also growth in the group with limited access to the automated system, which demands biometric passport [7,11]. This necessitates the need for the development of audit processes. As an operating environment, the airport requires that the border authority must take into account, in its own activities, the time and activity-related factors determined by the carrier and the airport. [8] These are important for the airport competitiveness. Effective processes are important success factors. Strengths of Helsinki-Vantaa Airport are short distances and at best plane can be changed in 35 minutes. Quick access to connecting flights is a competitive factor and as Asia's transit traffic increases, the smoothness of processes is an essential part to the success of the airport. [12]

The challenge for border checks at Helsinki-Vantaa Airport is the fluctuations in the time distribution of border crossing air traffic. This means number of passengers' variate heavily according to the time. Peaks are related to the structure of Finnair flights. In the afternoon the congestion of border checks for incoming and outgoing flights is mainly due to the so-called "the Asian wave". At that time, Finnair's traffic generates incoming traffic at border checks of up to 3,600 passengers per hour. As an example, if Finnair's long-distance traffic would add two daily flights that would mean that the number of border crossing individuals would increase between 400 to 600 border crossing individuals per day. On an annual level, increase would be up to 200 000 border checks more.

According to Finavia's data, in summer, especially in May, the number of passengers is growing strongly [13]. Figure 1 illustrates the hourly variation for air travelers on May 31st.

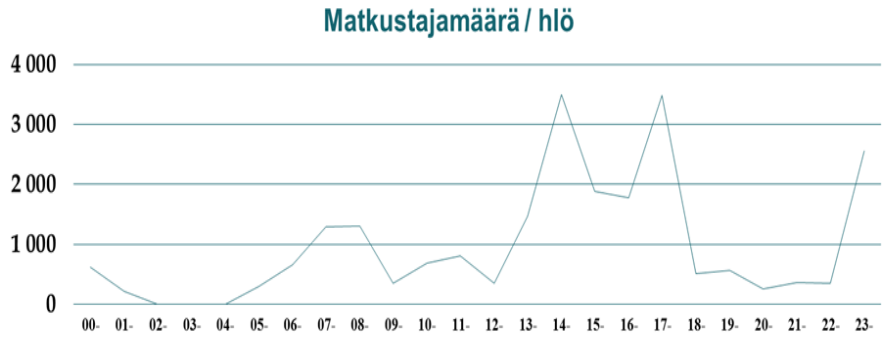


Fig. 1. Time fluctuation in passenger volume for 24h at Helsinki-Vantaa. [13]

When assessing the number of passengers, Finnair's own growth forecasts should not be considered as the only source of information. With regard to Helsinki-Vantaa, it would also be worth noting the growing interest of all airline carriers of transit traffic in Helsinki-Vantaa Airport. The admission of new companies to airlines operating via Helsinki-Vantaa may have a major impact on passenger transport growth. However, making forecasts is challenging because it is hard to anticipate passenger flows. Mobility is affected by various phenomena and decisions made outside Finland's borders. Surprising changes in people's behavior, mobility and travel can be caused by international security situation, immigration or climate change. [14]

Automation can significantly assist in verifying a person. The whole set of personal and travel information that has been identified at various stages of the border check process has become complex. Technical progress has made it possible to carry out checks of documents and persons in parallel. Future factors will be the automation of the audit process and the wider utilization of biometric identifiers. The border inspection process is then advancing towards solutions in which at the border crossing point the passengers themselves sovereignly master the procedures while the authorities are there mainly to supervise and help. Technical solutions can ensure the reliability of inspection activities and reduce queues at border crossing points. The implementation of technical services to facilitate the flow of cross-border traffic requires the pooling of data for different authors. [15]

3 Possibilities of collaboration-based data handling

In order to identify the information needed for border control, a border inspection application is utilized. The necessary data in the application is combined for sources to allow the user to view, process and compare different data. The Border Inspection Application is part of the Border Guard's Operational Information System (RVT) where all data entered at the border check is stored. [11]

Air carriers provide passenger information to the border inspection authority system Ratas using a secure connection using a standardized message format. At reception of passenger information, an aero-based MQ-based network is used. The communication follows the standards defined by the International Civil Aviation Organization (ICAO), the World Customs Organization (WCO) and the International Air Transport Association (IATA). The message format is the PAXLST (passenger list) message standard that passes within the Type B frame. The Border Guard has its own address on the network, where the sent messages are sent to the Ratas system of the Frontier Guard. [2]

Passenger information and regulation on border checks revealed the convergence of entries in border checks in relation to carriers' data in air traffic. It is clear that the current passenger data collected by the air carrier contains data which will also be surveyed and recorded in the entry and exit checks. Passenger information can be used to find out the necessary information without requiring the verification of documentary evidence presented in support of this information.

Unlike Advance Passenger Information (API) data, Passenger Name Record (PNR) data cannot be used to support border checks in accordance with the directives [16]. The data contained in the PNR data set is related to the information needed to determine the cross-border motive in order to establish the purpose and conditions of the intended stay. The duration of the planned stay PNR data supports API data in order for officer to determine the real destination.

As a whole, these "pre-requisite border controls" can be used, for example, in entry checks. An air carrier collects the following information that is consistent with the information available at the border inspection:

1. Country of Departure (API)
2. Transport, ID (API)
3. Destination (API)

4. Planed (Scheduled) Travel Date (s) (PNR).

To make better use of passenger data, automation is needed to process the information for acting border inspector. API data is designed for originally arriving passengers, so utilization for boarding is a challenge. Wider use of this data does not seem possible because of the delivery time specified in the law. This change would require both a review of the legal basis and new technical arrangements. [2]
Uniform operating models for collecting data may also contribute to collaborative information processing. Co-operation based data processing is not only related to the use of passenger data. Automation is not just a check-in automation. It is expected that technological solutions will not only be limited to checking-in or making reservations, but that the services will be further extended in the future [2]. At the same time, it is necessary to ponder the question whether or not independent controls are needed for border controls or would they be more efficient when the activities are linked together into the entire travel process? From the point of view of border checks, identification data from travel and personal documents can be obtained from the check-in desk. Primarily, the use of check-in machines would be airport-specific and related to the start-up inspection. The use of check-in machines for entry checks is technically complicated both in collecting and delivery of data. [2]

4 Modelling and Results

Border inspection process can be described by the following flowchart:

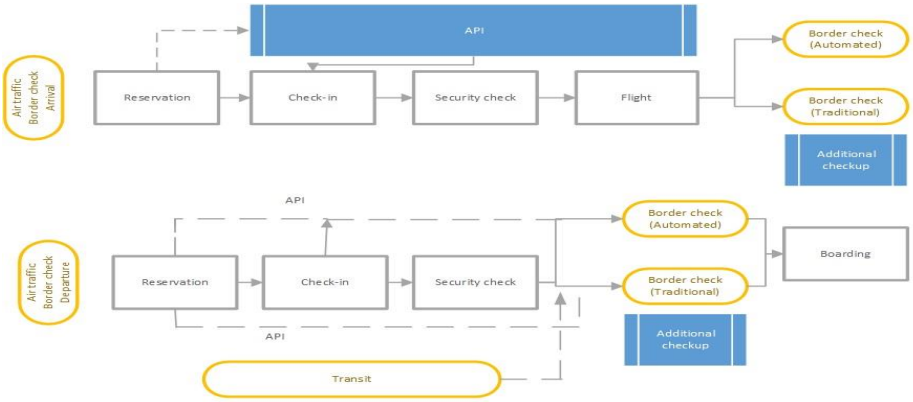


Fig. 2. Process flowchart of the passengers border crossing

In order to define correct parameters for the simulation model we did empirical testing about inspection times at Helsinki-Vantaa Airport. Test was carried out in 19.2. – 21.2.2018 and concentrated to the customers coming outside Europe. Sample size was 50 and samples were selected randomly.

Simulation purposes we defined measures that were able to give us information we needed from the boarding process. Here we were only interested how long time it takes to go through border inspection. Simulation model helped to find critical points from the process and gave valuable insight how time critical border checkup time is with respect to resources available. Simulation here models multiphase process, from where one can see how passengers incoming and exiting to the process affects to the queue length. Using this simulator, we tested how much sharing information between officials and use of automated border checkup could boost up passenger's flowrate.

Border checkup is queue process. Phases are arrival, waiting, service and exit. Bottle neck in practice in this process is waiting time in queue. In HEL border checkup is organized as follows:

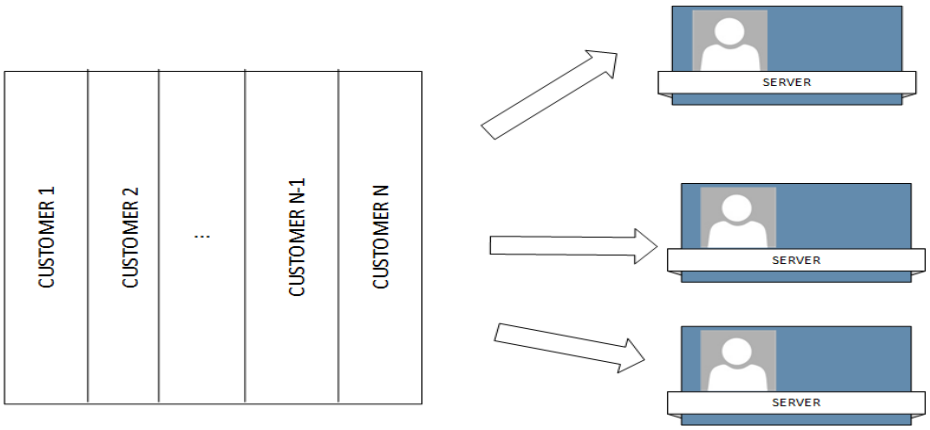


Fig. 3. One line many servers

Here we have one line with many servers. In order to make border check as effective as possible we can affect mainly to the effectiveness of servers. Service time distribution is a key factor in the fluent working of the queue. In the simulation model the following principles form the working model: 1) Number of customers in the queue 2) Number of customers in the whole system 3) Average time customers wait in the

queue 4) Average time of the border inspection 5) Average time in the system.

The model simulated the timing of the queue length by calculating passenger arrivals and exits at border inspection. Flight Passenger Parameters could be determined per flight. The user interface included setting the flight fill rate. The average waiting time for the first incoming flight queue was assumed to be empty at first. The simulation was continued until the last flight had arrived and the passengers had passed the inspection. In this case, the observation from a single simulation time was the average of all wait and wait times for all the passengers in that simulation run.

ARR	Cap max	Täyttöaste %	Manuaali tarkastettavat	ABC %	Tarkastajien määrä
0:25	100	100	30	70	10
0:30	209	100	62,7	70	10
0:55	165	100	49,5	70	10
0:55	144	100	43,2	70	10
1:40	212	100	63,6	70	10
5:45	297	100	89,1	70	10
6:25	297	100	89,1	70	10
6:55	100	100	30	70	10
6:55	100	99	29,7	70	10
7:00	100	100	30	70	10
7:00	100	100	30	70	10
7:10	306	100	91,8	70	10
7:55	254	100	76,2	70	10
8:50	289	100	86,7	70	10
10:20	87	100	26,1	70	10
10:50	289	100	86,7	70	10
11:40	130	100	39	70	10
11:55	289	100	86,7	70	10
12:20	209	100	62,7	70	10
12:35	140	100	42	70	10
13:10	100	100	30	70	10
13:30	100	100	30	70	10
13:45	209	100	62,7	70	10
13:50	297	100	89,1	70	10
13:55	336	100	100,8	70	10

Fig. 4. The user interface occupancy of the flights and to determine the auditees and auditors.

A total of 22 manual and 20 automated inspection lines are available for border checks of incoming passengers. Manual lines 8 are located in the first floor and the rest 14 in the second-floor spaces. Border checks on outbound passenger traffic are carried out on one floor, where 16 manual and 15 automated control lines are available. [6] In order to shorten the waiting times, there are separate routes for passengers belonging under the European Union's free movement right [17]. As a result of these arrangements, a total of 20 inspection points is available

for third-country nationals and 15 manual check points on outbound checks. With the terminal extension, the capacity of external border controls will also increase in the future. Exact line numbers have not yet been established [6], so existing border control capacity was used in simulation and analysis of results.

The number of flights varies daily. Through the simulation it was essential to illustrate the duration of the checks in relation to the resources used. To illustrate the situation as close as possible to reality, it was important to make use of reality-based timetables, which saw flight times and flight-specific maximum passenger capacity. According to the statistics from Finavia [13], passenger numbers will grow strongly, especially in May. In the simulation model we wanted to note the effects of the increase in the number of passengers in the summer timetable, so the preliminary data on the busiest day of May that was 30.5.2018 see Fig.1 was used as basic of simulation model.

Table 3 illustrates the results of the simulation with the current total time control (169.70 seconds) versus in case of automatically changing the passenger data (64.01 seconds). The simulation showed that with the selected parameters, the number of inspectors decreased considerably if the data already collected by the air carrier were automatically transferred to the simulation between the information systems. The time spent on border checks may also be increased within the limits permitted by the airport's border inspection capability, then the number of border inspectors may be a variable factor.

ARR	Number Inpectors (169,70)	Number Inpectors (64,01)	Exchange time min	ARR	Number Inpectors (169,70)	Number Inpectors (64,01)	Exchange time min
0:25	10	4	0:27	14:40	49	17	0:29
0:30	23	9	0:24	14:40	50	18	0:26
0:55	24	8	0:30	14:45	50	20	0:26
0:55	23	10	0:23	14:50	55	20	0:28
1:40	17	7	0:25	14:50	54	20	0:26
5:45	23	9	0:25	14:50	42	20	0:23
6:25	23	9	0:25	15:30	33	20	0:32
6:55	13	7	0:30	15:30	39	20	0:30
6:55	23	8	0:26	15:35	45	20	0:30
7:00	23	9	0:27	15:35	60	20	0:25
7:00	29	9	0:24	15:20	60	20	0:26
7:10	28	9	0:24	15:20	56	20	0:22
7:55	20	8	0:25	15:25	52	16	0:24
8:50	24	9	0:24	15:50	4	5	0:33
10:20	7	3	0:25	15:50	22	11	0:23
10:50	23	9	0:24	16:20	14	5	0:25
11:40	11	4	0:25	17:40	17	7	0:25
11:55	23	9	0:24	18:05	14	6	0:24
12:20	17	7	0:25	20:05	13	5	0:26
12:35	11	5	0:24	20:25	8	3	0:27
18:10	8	3	0:27	21:00	17	7	0:25
18:30	8	3	0:29	22:05	12	5	0:24
18:45	24	10	0:29	22:15	8	3	0:26
18:50	40	14	0:27	22:30	8	3	0:27
18:55	40	14	0:24	23:00	18	8	0:25
14:10	35	10	0:27	23:05	18	6	0:24
14:10	30	15	0:23	23:25	15	6	0:24
14:15	24	10	0:25	23:35	23	9	0:29
14:20	35	15	0:24	23:35	36	11	0:24
14:25	37	16	0:24	23:40	37	12	0:23
14:30	48	17	0:24	23:55	12	6	0:26

Table 3. Simulation result of automatic transmission of data on the flow rates of passengers

Table 4 shows the effects of the usage of advanced passengers information(API) to the inspection-time. We see that result is significant.

	Present (s)	**Automated (s) (change %)**
In-depth arrival inspection - Phase 4 - Departure (API), conveyance, identifier (API) - Phase 6 - Destination (API), Travel plan (PNR).	169,70	64.01 (- 62.28 %)
In-depth boarding inspection - Phase 5 - Destination (API), conveyance, identifier (API)	43.16	41.05 (- 4.89 %)

Table 4. Effects of the use of passenger information to the inspection-time

Conclusions and Future

The simulation showed that the current border control total time (169.70 seconds) cannot achieve the current maximum time set for inspection, i.e. the waiting time for a border check can be 10 minutes (Table 3). The issue was solved in this simulation model so that the sum of the waiting time and the inspection time was a total of 10 minutes. It depicted the time taken to inspect one machine for passengers so that even the las one in the queue does not have to wait longer than the target time.

The key result of the simulation was that automatic exchange of passenger data can be used to carry out the necessary flight checks with the current border inspection capacity so that the effectiveness objective is met. In exit checks benefit was significant. The effects of the utilization of passenger data on the inspection times are presented in Table 4.

The result of the study was that automatic procedures for handling current passenger data should be developed in the future so that border inspectors will be able to access the necessary information electronically. The question is mainly about the new technical implementation and the new sharing of information among the different authors. Passenger data sharing speeds up the required checks to a significant extent in connection with the border crossing process. The impact was particularly evident in a thorough entry check, which only covered one third of the current time.

The simulation model built during the research gave grounds for evaluating performance, explaining the flow efficiency of the process and its impact from the point of view of border authorities: how to allocate resources and how to use them to achieve the overall quality of services sought. Performance was measured in order to see what kind of results were achieved with current activity. The most important elements of the simulation were the waiting time and the time of inspection. It was critical to keep the times within the allowed limits. When attention was paid to the most important variables in the course of the process, it was possible to identify the factors in which the process can be enhanced. The simulation model used was particularly suitable for modeling incoming air traffic. In the simulation model, the percentage of EU / EEA / CH nationals on the external borders was based on the calculated averages of total number of passengers. With a view to monitoring the impacts of a credible growth, the relationship between third-country nationals of each flight should be monitored separately. This is also possible if we want to extend the simulation. This study was based on a simplified model that provides sufficient initial data for the examination of the border inspection process for third-country nationals with selected variables.

References

1. Sundarapandian V.: Probability, statistics and queuing theory. PHI Learning Pvt. Ltd. (2009)
2. Mattila P., Saastamoinen K. & Rissanen A: DESCRIPTION OF A MODERN DIGITALIZED AND CO-OPERATED BORDER CONTROL ENVIRONMENT AT AIRPORT, Sent ISMS Conference 2018.
3. Finnair Strategy https://investors.finnair.com/en/finnair-as-an-investment/strategy (Accessed 27th Jun, 2018)
4. Finavia annual Report 2017, https://www.finavia.fi/en/newsroom/2018/five-reasons-behind-record-breaking-passenger-volumes-finnish-airports (Accessed 27th Jun, 2018)
5. Haapasaari V., HELSINKI VANTAAN KEHITYSOHJELMAN ETENEMINEN KOHTI ENTISTÄ PAREMPAA LENTOASEMAA, Presentation, available in http://docplayer.fi/816453-Helsinki-vantaan-kehitysohjelman-eteneminen-kohti-entista-parempaa-lentoasemaa.html (Accessed 27th Jun, 2018)
6. Internet article, Lentoposti, http://www.lentoposti.fi/uutiset/helsinki_vantaan_lentoaseman_rajatarkastuslaajennus_k_ynnistyy_rajavartiolaitos_saa_lis_tiloja (Accessed 27th Jun, 2018)
7. RATTI 2018 Boarder Guard internal system, report from information system.
8. Border Guard Resource Plan 2018 and Action Plan and Economic Plan 2019-2022. (Accessed 18th Jan 2017)
9. Boarder Guard (Finland) Internal filesystem, Document: Rajatarkastushenkilöstötarpeet osana Helsinki-Vantaan lentoaseman rajanylityspaikan kehittämistä, RVLDno/2016/1262, (2016)
10. Council Regulation (EC) No 2252/2004, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32004R2252, (Accessed 27th Jun, 2018)
11. Finnish Law 579/2005 Act on the Processing of Personal Data by the Border Guard, https://www.finlex.fi/en/laki/kaannokset/2005/en20050579 accessed 27th Jun, 2018
12. Finavia2017, Finavia invests in developing Helsinki Airport, https://www.finavia.fi/en/about-finavia/development-airports/invests-in-helsinki accessed 27th Jun, 2018
13. Finavia. Statistics 2017, https://www.finavia.fi/en/about-finavia/about-air-traffic/traffic-statistics/traffic-statistics-year accessed 27th Jun, 2018
14. Ministry of Internal Affairs (Finland) Publication 15/2017, http://julkaisut.valtioneuvosto.fi/handle/10024/80782 accessed 27th Jun, 2018
15. European Commission, On establishing a framework for interoperability between EU information systems, http://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-794-F1-EN-MAIN-PART-1.PDF accessed 27th Jun, 2018
16. Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime, https://publications.europa.eu/en/publication-detail/-/publication/2ba036c2-11bd-11e6-ba9a-01aa75ed71a1/language-en accessed 27th Jun, 2018
17. Regulation (EU) 2016/399 of the European Parliament and of the Council of 9 March 2016 on a Union Code on the rules governing the movement of persons across borders (Schengen Borders Code), https://publications.europa.eu/en/publication-detail/-/publication/42fba6c3-f0c5-11e5-8529-01aa75ed71a1/language-en accessed 29th Jun, 2018

Complex networks of scalar time series using a data compression algorithm

Débora Corrêa*, David Walker, and Michael Small

Department of Mathematics and Statistics, University of Western Australia,
Nedlands, Perth, WA 6009 Australia
{debora.correa,david.walker,michael.small}@uwa.edu.au

Keywords: time series; complex networks; compression algorithms

1 Abstract

A representation of scalar time series, which has proved to be useful to derive insights about the underlying process generating the data, is to transform the time series into a complex network. This allows data analysis via the rigor of mathematical graph theory, and characterization through application of a range of complex network metrics. Indeed the success of network representations as compact summaries of time series data has promulgated diverse techniques for such time series to network transformations [1–4].

For most of these techniques, however, the size of the network (given by the number of nodes) will scale with the size (length) of the time series, enabling only a statistical description of the network properties. This can be problematic if a detailed study of the fine structure of the network evolution is required, e.g., calculation of proxies for estimates of dynamical invariants of the systems. Longer time series should help improve estimates, but if longer time series translate to larger, more cumbersome to analyze, networks, then large-scale data becomes a hindrance. To permit a full network analysis of large-scale time series, this work proposes an intermediate step in the transformation process.

We first apply a compression algorithm to the time series and represent its output as a complex network. The data compression algorithm exploits recurrences in the time series leading to a more compact description of the underlying dynamics. In this compression-based approach, nodes of the network correspond to symbolic codewords derived from the data compression algorithm, while the network edges indicate successive codewords within the compressed time series. We refer to this network as a compression network.

In previous work, we demonstrated that properties of compression networks are capable of distinguishing different dynamical behaviours, even in the presence of noisy data [1]. In this talk, we discuss extensions of this network transformation scheme and provide some initial results for two different exploratory applications. In the first, we show that with a suitable windowing strategy, tracking changes in compression network sub-graphs appears capable to detect concept

* corresponding author. Email: debora.correa@uwa.edu.au

drift [5] in data streams. The smaller scale of the compression networks makes such fine-grained analysis feasible. In the second, we consider a classification problem whereby music can be grouped according to how similar their compression network representations are. A useful similarity score between two refrains, regarded as a complexity measure, is their respective compression ratios. We use this complexity measure to construct a similarity matrix and classification proceeds by examining the community structure of networks induced by this matrix. We show that (broad) musical styles can be classified and that the classification is robust with respect to a thresholding parameter.

References

1. Walker D. M., Correa D. C., Small M.: On system behaviour using complex networks of a compression algorithm, Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 28(013101), pp. 1–9, (2018).
2. Lacasa L., Luque B., Ballesteros F., Luque J., Nuno J. C.: From time series to complex networks: The visibility graph, Proceedings of the National Academy of Sciences, 105(13), pp. 4972–4975, (2008).
3. Marwan N., Donges J. F., Zou Y., Donner R. V., Kurths J.: Complex network approach for recurrence analysis of time series, Physics Letters A, 373(46), pp. 4246–4254, (2009) .
4. Small M., McCullough M., Sakellariou K.: Ordinal network measures – quantifying determinism in data, Proceedings of ISCAS, pp. 1–5, (2018).
5. da Costa F. G., Rios R. A., de Mello R. F.: Using dynamical systems tools to detect concept drift in data streams, Expert Systems with Applications, vol. 60, pp. 39–50, (2016).

Computation and validation of wind and solar time series based on global reanalysis

Marta Victoria, Gorm B. Andresen, and Martin Greiner

Department of Engineering, Aarhus University,
Inge Lehmanns Gade 10, 8000 Aarhus C, Denmark
*mvp@eng.au.dk

Future low-carbon power systems will most probably rely on Variable Renewable Energy Sources (VRES), *i.e.*, onshore and offshore wind and solar photovoltaics (PV). Under the constraint of 95% CO_2 reduction, relative to 1990, the cost-optimal European power system includes large shares of wind energy in northern countries and PV in southern countries [1, 2]. Increasing transmission capacities among neighbor countries ease the grid integration of VRES. Alternatively, coupling the electricity to other sectors such as heating and transportation can provide additional storage to counterbalance the inherent variability of VRES, as well as help to decarbonize those sectors [3, 4]. The outcome of energy models such as [1–4] are significantly impacted by the time series representing VRES used as inputs [5]. Hence, we must ensure that they properly capture the temporal characteristics of renewable generation.

This presentation will provide details on the generation and validation of time series representing the hourly capacity factors for PV, onshore, and offshore wind energy at a national scale for every country in Europe. The global Renewable Energy atlas (REatlas) from Aarhus University [6] has been used to generate the time series. The REatlas uses as input reanalysis data, that is, the output of global atmospheric simulations, together with simplified models of renewable technologies, *i.e.*, power curves for wind turbines and efficiency curves and orientation for solar panels. The Climate Forecast System Reanalysis (CFSR) from the National Center for Environmental Prediction (NCEP) [7] is used as input for the model. It comprises a 38 years-long global high-resolution dataset (hourly time resolution and approximately 40 x 40 km^2 space resolution).

Using global reanalysis datasets to generate VRES time series has significant advantages. Firstly, since they cover the entire globe, the methodology can be replicated to obtain hourly capacity factors in every country. Secondly, reanalysis datasets usually expand several decades enabling the generation of VRES capacity factors for long-time periods. Thirdly, reanalysis dataset can be used to validate and bias correct climate models enabling the assessment of climate change impacts on energy system as in [8]. Fourthly, they can also be used to generate time series representing hydroelectricity production and, hence, a consistent set of renewable generation time series to feed-in a certain energy model

can be obtained. Finally, reanalysis data is usually freely available making it suitable for scientific analysis and the replicability of results.

Onshore and offshore wind time series

The location, height, and capacity of every onshore and offshore wind turbine in Europe is retrieved from a comprehensive database [9]. The turbines are classified into several 250 kW-wide categories (see Table 1 in reference [6]) and associated to the point in the CFSR grid that is closer to its actual position. A logarithm curve is used to estimate the wind speed at hub height from the speed at 10 m height provided in the CFSR dataset, and the power curve of the turbine is used to compute the electricity output. As shown in [6], adding some correction to the wind power curves to represent the fact that the aggregation of wind power plants shows a smoother curve than one single turbine significantly improves the matching with real wind production. A Gaussian smoothing function is used and their parameters are determined, for every country, by minimizing the Kullback-Leibler divergence between the modeled time series for 2015 and the historical reference. The latter is obtained by processing historical data on cumulative installed capacity and hourly wind generation retrieved from national TSOs. The method is proven to outperform other alternative approaches [10, 11].

Photovoltaic time series

Generating PV time series present some additional challenges. Reanalysis is not able to properly capture clouds dynamics which heavily impact the irradiance reaching PV system. Furthermore, the location and configuration (orientation and inclination) of the thousands of PV installations in every country is unknown. Similarly to wind time series, irradiance from CFSR reanalysis dataset is converted into electricity generation using a simplified model and aggregated at country level to obtain PV time series. Prior to conversion, reanalysis irradiance is bias corrected using satellite-based SARAH dataset and a globally-applicable methodology. For every, country twelve correction factors, one per month, are computed to include the effect of local atmosphere. Moreover, a novel procedure is proposed to infer the orientation and inclination angles representative for PV panels based on the historical PV output throughout the days around summer and winter solstices. For every country, four different PV configurations, *i.e.* rooftop, optimum tilt, tracking, and delta, have been investigated. These are shown to have a strong influence in the hourly difference between electricity demand and PV generation [12]. 38-year long hourly time series for every configuration have been made available under open-license.

References

1. E. H. Eriksen, L. J. Schwenk-Nebbe, B. Tranberg, T. Brown, M. Greiner, Optimal heterogeneity in a simplified highly renewable european

electricity system, Energy 133 (Supplement C) (2017) 913 – 928. doi:https://doi.org/10.1016/j.energy.2017.05.170.
URL http://www.sciencedirect.com/science/article/pii/S0360544217309593

2. D. P. Schlachtberger, T. Brown, S. Schramm, M. Greiner, The benefits of cooperation in a highly renewable European electricity network, Energy 134 (Supplement C) (2017) 469–481. doi:10.1016/j.energy.2017.06.004.
URL http://www.sciencedirect.com/science/article/pii/S0360544217309969
3. T. Brown, D. P. Schlachtberger, A. Kies, S. Schramm, M. Greiner, Synergies of sector coupling and transmission extension in a cost-optimised, highly renewable european energy system, arXiv:1801.05290.
URL https://arxiv.org/abs/1801.05290
4. K. Zhu, M. Victoria, T. Brown, M. Andresen, Gorm B. Greiner, Impact of CO_2 prices on a highly decarbonised coupled electricity and heating system in europe, to be published.
5. D. P. Schlachtberger, T. Brown, M. Schäfer, S. Schramm, M. Greiner, Cost optimal scenarios of a future highly renewable european electricity system: Exploring the influence of weather data, cost parameters and policy constraints, arXiv preprint arXiv:1803.09711.
6. G. B. Andresen, A. A. Sndergaard, M. Greiner, Validation of Danish wind time series from a new global renewable energy atlas for energy system analysis, Energy 93 (Part 1) (2015) 1074–1088. doi:10.1016/j.energy.2015.09.071.
URL http://www.sciencedirect.com/science/article/pii/S0360544215012815
7. S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H. ya Chuang, M. Iredell, M. Ek, J. Meng, R. Yang, M. P. Mendez, H. van den Dool, Q. Zhang, W. Wang, M. Chen, E. Becker, NCEP climate forecast system version 2 (CFSv2) selected hourly time-series products (2011).
URL https://doi.org/10.5065/D6N877VB
8. S. Kozarcanin, H. Liu, G. B. Andresen, Climate change impacts on large-scale electricity system design decisions for the 21st Century, arXiv:1805.01364 [eess]ArXiv: 1805.01364.
URL http://arxiv.org/abs/1805.01364
9. Wind energy database.
URL https://www.thewindpower.net/
10. EMHIRES dataset . Part I, Wind power generation., Tech. rep., Joint Research Center (Nov. 2016).
URL https://publications.europa.eu/en/publication-detail/-/publication/85b2dc7f-aa61-11e6-aab7-01aa75ed71a1/language-en
11. I. Staffell, S. Pfenninger, Using bias-corrected reanalysis to simulate current and future wind power output, Energy 114 (Supplement C) (2016) 1224–1239. doi:10.1016/j.energy.2016.08.068.
URL http://www.sciencedirect.com/science/article/pii/S0360544216311811
12. M. Victoria, G. B. Andresen, Using validated reanalysis data to investigate the impact of the PV system configurations at high penetration levels in european countries, to be published.

A STUDY WITH NDVI TIME SERIES OF THE BRAZILIAN CAATINGA

Claudionor R. Silva[1], *Aracy A. Araújo*[1], *Sérgio L. D. Machado*[1]
[1] Federal University of Uberlândia, Postgraduate Program in Environmental Quality, BRAZIL

ABSTRACT

The Brazilian Caatinga represents 9.92% of the territorial area. Is it exclusively Brazilian and very rich in biodiversity. Because it has low resistance to human interference is necessary to know the important factors in monitoring the biome. Normalized Difference Vegetation Index (NDVI) has been recently explored in the description of tree phenology. In this study we sought to create NDVI/Landsat series to describe the phenological behavior of the Caatinga. A time series analysis was carried out, describing the intrinsic parameters of the series (Seasonality and Trends), the forecast of NDVI values using. The results showed that the NDVI series presents seasonally, but does not exhibit a trend, and the forecasting process presenting relatively low error at a 95% confidence interval.

Keywords: SARIMA model, Caatinga, NDVI.

1. INTRODUTION

Brazil has an enormous land area, approximately 8,500,000 km^2, where you can verify very diverse weather, soils and vegetation coverage. Throughout the length of the country there are portions with similar characteristics and continuous biodiversity, which defines the so-called "biomes". There are six well-defined biomes in Brazil (% in area): Amazon (49,29%), Cerrado (23,92%), Atlantic Forest (13,04%), Caatinga (9,92%), Pampa (2,07%) and Pantanal (1,76%) [6]. The Caatinga is an exclusively Brazilian biome. This name stems from the whitened landscape, characterized by the appearance of the tree trunks in the dry season due to the almost total loss of foliage.

Studies indicate show that the Caatinga is heterogeneous, being very rich in biodiversity and endemic species. Caatinga vegetation is classified as steppe savanna and the landscape is quite different due to the sharp variation of rainfall, fertility and types of soil and topography. Despite its great wealth in biodiversity, the Caatinga is the most fragile Brazilian biome and the indiscriminate use of soil and natural resources, during the occupation/exploitation process, has degraded the Caatinga (soil end vegetation coverage). The uniqueness, fragility, unsustainable use, particular wealth and small extent express the need for studies which seek to preserve the biome [9].

An extensively explored parameter for the analysis of vegetation coverage is the Normalized Difference Vegetation Index (NDVI) calculated with the use of digital images, more specifically, the red and infrared bands of a multi-spectral sensor. This index is a numerical indicator of the intensity of the presence of vegetation, ranging from [-1 +1], indicating with -1 (minus one) the absence of vegetation (low or nil chlorophyll content) and 1 (one), the presence of vegetation in its chlorophyll plenitude. NDVI has shown its efficiency in recording the phenological changes in vegetation [11]. Recently some studies have been made in order to show the applicability of NDVI values to describe the behavior of different plant vegetation types [7, 8, 10].

In this context, this study proposes a reliable approach for Caatinga phenology mapping in protection areas by means of NDVI time series. The intrinsic parameters of the series (trend, cycle, seasonality) were analyzed, as well as a prediction of their behavior being modeled. Knowledge of time series elements allows for the identification and understanding of standards and ecological and evolutionary processes that operate in the Caatinga. Given that the Caatinga is the biome which is proportionately less studied and less protected in the country, the results of this research become important contributions to the elaboration of the management plan of the species in the region, avoiding problems such as the loss of unique species and the formation of desertification centers.

2. MATERIAL AND METHOD

2.1. Study Area

Study area is the Serra da Capivara National Park (PARNASC), which has 91,848.88 ha of protected area in the Southeast of Piauí State, in northeastern Brazil (Figure 1). It has a semi-arid climate and covers part of the sandstone plateau, dominated by tree-shrub Caatinga (location chosen for sampling-pixels) [2, 4].

2.2. Datasets

Due to the long period in operation, we chose to use the images of the Landsat 5 satellite (Thematic Mapper sensor - TM), for generation of NDVI time series. However, due to a problem which occurred with the supply of images from Landsat 5, in 2002, images from Landsat 7 ETM + were used for that year. Although both the TM sensor and the ETM + generate 7 spectral bands (electromagnetic spectrum), this study used only 3 (red: 0.63-0.69µm) and 4 (near infrared: 0.76-0.90µm). This bands (spatial resolution of 30 m) was collected monthly during the period from 1984 to

2010. All Landsat images were obtained free on the site/web United States Geological Survey [12].

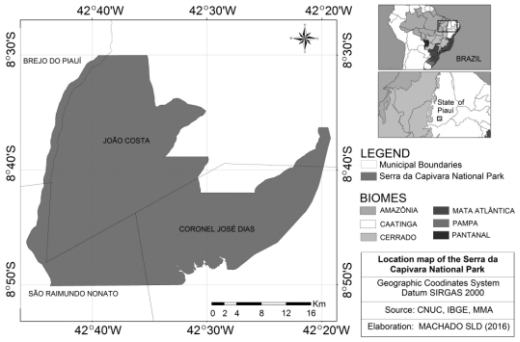


Figure 1. Serra da Capivara National Park.

The ENVI 4.8 software image processing (Exelis Visual Information Solutions) is used to images preprocessing and to determine NDVI. The processing of the temporal series was carried out using free Gretl-2016c software.

2.3. Generation of NDVI Time Series

To make Landsat7/ETM+ equivalent to Landsat5/TM data was used the linear model proposed by [13]. All images were previously recorded (first degree polynomial model) so as to ensure the geometric consistency between the corresponding pixels of images. For this, the image of 1984 was used as reference. Still in the preprocessing phase, all Landsat images were corrected for atmospheric effects. These preprocessing procedures were performed in the ENVI 4.8 software.

In the first step of the processing, the NDVI was calculated in all images used. The method is widely recognized in literature and uses simple arithmetic calculations of spectral bands (Equation 1).

$$\mathrm{NDVI} = \frac{\mathrm{IR}-\mathrm{R}}{\mathrm{IR}+\mathrm{R}} \qquad (1)$$

where R and IR are the spectral red and infrared bands, respectively.

Subsequently, in each NDVI image, the digital values of the pixels in Caatinga feature were collected and their mean stored in the form of a column vector forming the monthly NDVI time series. In this vector, the first twelve cells store the NDVI values of the first year (1984), in the order from January to December. The next twelve values refer to the subsequent year (1985), and so on. Storing in a vector shape is to facilitate entry of the time series in the Gretl software.

Although the region presents characteristics of prolonged drought, in some years there are periods with heavy rainfall, resulting in images covered by many clouds. This fact occurs in specific months, usually from December to March. This problem was resolved in two ways: a) when the month before or after the month with clouds had two scenes, one was used for this problematic month; and b) when the previous case did not occurred, the NDVI values were interpolated. The time series is monthly and corresponding to a period of 27 years, 1984-2010.

2.4. Basic elements of the NDVI time series

A time series is a set of observations ordered in time in a continuous or discrete form. A time series (Y_t) is made up of four basic elements: trend (T), seasonality (S), cycle (C) and irregular variations (I), as shown in Equations 2 (multiplicative) and 3 (additive). These elements can be separately analyzed so as to identify particularities of the series. An important data obtained from the time series is that of forecasting variables, events or phenomena in an interval of time after that observed. Before the calculation of the forecast it is necessary to carry out some tests in the series so as to verify the applicability as well as identifying features such as the stationarity of this series.

$$Y_t = T \cdot S \cdot C \cdot I \qquad (2)$$
$$Y_t = T + S + C + I \qquad (3)$$

The trend expresses the long-term behavior and shows changes that occur in the average level of the series which may be deterministic (mathematical model) or stochastic (random process). Irregular variations are changes which were not foreseen in relation to the trend line, resulting from unexpected events such as a fire or deforestation in the case of NDVI series, commonly referred to as white noise.

Cyclical or cycle variations are fluctuations in the form of long waves that occur around the trend line, referring to the values of variables, events or phenomena, and which repeat over time. Seasonal variations are cyclical movements that are completed during a period and reproduce themselves in other similar periods, defining regular patterns of the series. In the study of a time series it is important that the seasonal component be identified and isolated so that it can work with a deseasonalized series (free or seasonally adjusted).

In a stochastic time series it is necessary to know whether the randomness of the variables do not change in time, which characterizes a stationary process. In this way, a stochastic time series is said to be stationary when presenting average and constant variance in time and the covariance between two periods depends only on the distance or the gap between the two periods, and not on the effective time period in which the covariance is calculated [1, 3, 5].

The stationarity of a time series can be obtained by unit root tests. A test for analysis as to the existence of a unitary root is the Augmented Dickey-Fuller (ADF). In this test, if the null hypothesis $H_o: \delta = 0$ is rejected, the data series does not have a unitary root, in other words it is stationary.

2.5. Forecast analysis of NDVI series

The autoregressive models which are integrated with a moving average (ARIMA) are widely used to forecast data using time series, especially because they are also used in non-stationary series. Such models are based on the theory that a variable Y_t can be explained by the historical values of this variable itself, added to the stochastic error terms of the series [3].

In forecast for the NDVI series the year 2010 was removed from the series for the 12-month forecast error analysis.

3. RESULTS AND CONCLUSIONS

3.1. Trend, Seasonality and Stationarity

The time serie NDVI have 324 observations, covering the time period from January 1984 to December 2010. It was observed that there are no outliers and not even a definite trend. On the other hand, a possible seasonality is revealed on the NDVI series because of the uniform behavior in the form of waves. To confirm, we proceeded to the trend and seasonality tests. The NDVI series, as shown in Table 1, did not really present a trendy. The test was carried out by adding a time trend variable and making an regression. It was verified that the time variable added to the model was not statistically significant. In addition to this test two other experiments were also made to check the trend: one with the square of the time variable, previously created and another with the logarithm of the NDVI variable. In both cases it was also found that the series showed no trend.

Table 1. Analysis of the trend of the NDVI series.

	Model 1: NDVI	
	Coefficient	Standard Error
Constante	0,4436***	0,0140509
Time	7,99414e-05	7,49401e-05
	Model 2: Time²	
Constante	0,44746***	0,0105203
Sq_Time	2,60425e-07	2,23229e-07
	Model 3: ln(NDVI)	
Constante	−0,870683***	0,0340675
Time	0,000266657	0,000181699

*** Statistically significant at 1%.

Based on the graph/periodogram the presence of seasonality was observed. Then we tried to model this seasonality by seasonal lag. The seasonal effect was modeled with only one seasonal lag. After the identification of the seasonal component, the correlogram test to identify the presence of a unit root and the Augmented Dickey-Fuller test (ADF) in order to identify if the number in question is stationary were applied, in other words, if it does not have root unit. The results showed that the NDVI series is not stationary, as the p-value calculated was > 5%, so it was not possible to reject the null hypothesis (the series has a unit root).

In order to make the series "stationary", differentiations were made in series and once again the tests were carried out. For the ADF test, after the second difference, a p-value <5% was obtained, demonstrating that the series was stationary. In addition to the ADF test, the correlogram analysis also confirmed that the series has become stationary.

Based on the definition of the stationarity order of the NDVI series the identification of the autocorrelation filter (RA) and the median filter (MA) were made. Accordingly, after analyzing the autocorrelation the estimated model is a SARIMA (1,2,2) x (0,1,2), whose result of the estimation is shown in Table 2.

Table 2. SARIMA Model (1,2,2)x(0,1,2).

	Coefficient	Standard Error
ϕ	0,3587***	0,0598
θ_1	-1,9305***	0,0250
θ_2	0,9308***	0,0250
Θ_1	-0,8571***	0,0351
Θ_2	-0,0481***	0,0183

*** Statistically significant at 1%.

It is observed that all the variables of the estimated model (Table 2) were significant at 1% in the formulation of the SARIMA model (1,2,2)x(0,1,2).

3.2. Forecast of the NDVI Series

The good specification of the model, together with the fact that the series became stationary, guarantees the forecast of the series in a safe manner. The forecast was made for the year 2010, with the series reduced for the 1984-2009 period (Figure 2). The objective is to compare the forecasted values with the actual values of the NDVI series, recorded in the year 2010.

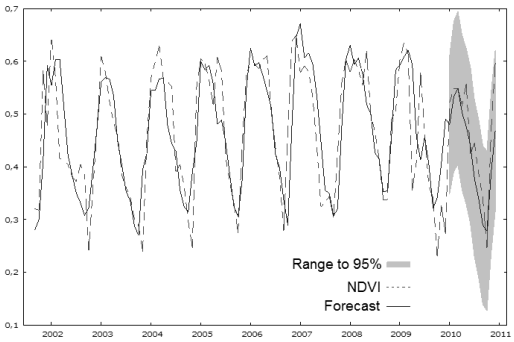


Figure 2. Forecast and real values of the NDVI series.

Based on Figure 2 it can be seen that the forecasted values are within the significance areas and are very close to the actual values of 2010. The reliability of the forecasting method used is guaranteed by the low value of the mean square error (0,00285) and by the modest value of the standard deviation error (0,05036).

4. ACKNOWLEDGMENT

The authors thank the Minas Gerais Research Support Foundation (FAPEMIG) for the financial support given to the research project.

REFERÊNCIAS

[1] BOX, G.E.P.; JENKINS, G.M.; REINSEL, G.C. **Time Series Analysis – Forecasting and Control.** 5th ed.; John Wiley & Sons: New Jersey, US. 2015. p.712. ISBN 111867491X.

[2] EMPERAIRE, L. **Végetation et gestion des resources naturelles dans la caatinga du sud-est du Piauí (Brésil).** 1939. Doctorat d'Etat ès Sciences Naturelles. Université Pierre et Marie Curie, Paris.

[3] ENDERS, W. **Applied econometric time series**. John Wiley & Sons: New York, US. 1995. p.433.

[4] FUMDHAM - **THE AMERICAN MAN MUSEUM FOUNDATION**. Available at: http://www.fumdham.org.br/. Accessed in: 02 july 2016.

[5] GUJARATI, D.M. **Basic Econometrics**. 6nd. McGraw-Hill Irwin: Boston, Mass, US; 2009. p.922. ISBN 0071276254

[6] IBGE - **BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS**. Available at: http://www.ibge.gov.br/. Accessed in: 25 june 2016.

[7] JIA, K.; LIANG, S.; WEI, X.; YAO, Y.; SU, Y.; JIANG, B.; WANG, X. Land Cover Classification of Landsat Data with Phenological Features Extracted from Time Series MODIS NDVI Data. **Remote sensing**. v.6, p.11518-11532, 2014. doi:10.3390/rs61111518.

[8] LAMBERT, J.; DENUX, J.P.; VERBESSELT, J.; BALENT, G.; CHERET, V. Detecting Clear-Cuts and Decreases in Forest Vitality Using MODIS NDVI Time Series. **Remote Sensing**. v.7, p.3588-3612, 2015. doi:10.3390/rs70403588.

[9] LEAL, I.R.; TABARELLI, M.; SILVA, J.M.C.; LARCHER, T.E. Changing the course of biodiversity conservation in the Caatinga of Northeastern Brazil. **Conservation Biology**. v.19, p.701-706, 2005. doi:10.1111/j.1523-1739.2005.00703.x.

[10] PINZON, J.E.; TUCKER, C.J. A Non-Stationary 1981–2012 AVHRR NDVI3g Time Series. **Remote sensing**. v.6, p.6929-6960, 2014. doi:10.3390/rs6086929.

[11] RAO, Y.; ZHU, X.; CHEN, J.; WANG, J. An Improved Method for Producing High Spatial-Resolution NDVI Time Series Datasets with Multi-Temporal MODIS NDVI Data and Landsat TM/ETM+ Images. **Remote Sensing**. v.7, p.7865-7891, 2015. doi:10.3390/rs70607865.

[12] USGS - **United States Geological Survey**. Available in: http://landsat.usgs.gov/landsat8.php. Accesses: 02 feb. 2017.

[13] VOGELMANN, J. E.; HOWARD, S. M.; YANG, L.; LARSON, C. R.; WYLIE, B. K.; VAN DRIEL, J. N. Completion of the 1990's National Land Cover Data Set for the conterminous United States. **Photogrammetric Engineering and Remote Sensing**, 67, 650-662. 2001.

Characterizing Market Behavior through Risk Forecasts: a Powerful VaR Backtesting

Marta Malecka

Department of Statistical Methods, University of Lodz,
41/43 Rewolucji 1905 r., 90-214 Lodz, Poland
marta.malecka@uni.lodz.pl

Abstract. We refer to the current Basel framework and investigate the issue of risk model evaluation through backtesting VaR forecasts. In this context we compare the likelihood ratio tests, which has become the industry standard for backtesting VaR, to other econometric methods applicable to the binary VaR failure process. We apply the binomial and the normal statistics to testing VaR unconditional coverage and for conditional coverage we employ the Ljung-Box and the Pearson's χ^2 statistics. Via simulations we compare the statistical properties of the considered methods. Through the empirical example, based on S&P500 data, we demonstrate how the statistical properties translate into practical differences in risk analysis. In particular we show that, by replacing the standard approach with the proposed methods, we achieve a more insightful conclusions about S&P500 dynamics during the market crash and in the aftermath of the crisis.

Keywords: Value-at-Risk, backtesting VaR, Kupiec Test, Markov Test, Test Size, Test Power

1 Introduction

The current international regulatory framework, as indicated by the Basel III and the Basel IV agreements [1,2], imposes *Value-at-Risk* as a risk measure used for evaluating market risk models. Thus backtesting VaR remains a topical issue in academic research, attracting also attention of market practitioners and regulators. The VaR testing framework is based on a binary failure process, which, under the correct risk model, is required to satisfy the iid Bernoulli property. This property is commonly split into into the postulates of unconditional and conditional coverage property. The first postulate refers to the overall VaR failure rate and means that the number of violations should match the assumed VaR tolerance level, while the conditional coverage property requires independence of violations.

Though the plethora of methods have been proposed in the literature for backtesting one or two of the VaR failure process properties, only two tests, developed specifically for the purposes of risk management by Kupiec [14] and

Christoffersen [6], have won a wide recognition in the industry[1]. Both tests use the likelihood ratio (LR) framework and both are frequency-based in the sense that they check the number of VaR failures in the sample. The Kupiec test, as dedicated to the unconditional coverage property, uses the overall frequency of failures. The conditional coverage Christoffersen's test treats the failures in relation to the previous observation and, referring to the framework of the binary Markov chain, checks their time dependence. Offering advantage of convenient application to real VaR failure sequences, these mainstream tests are repeatedly criticized for their unsatisfactory power properties. This boosts a discussion about other possibilities of developing a VaR-based risk model evaluation procedures.

Without resorting to the use of procedures dedicated specifically to risk management, we explore possibilities of backtesting VaR through standard econometric methods, applicable to the Bernoulli sequence and independence testing. To avoid using asymptotic theorems we suggest testing the unconditional coverage property through the direct reference to the binomial distribution. Within a continuous-distribution-based procedures we replace the LR χ^2 statistic with a normal test, using the asymptotic convergence of the binomial distribution to the normal one. According to the conditional coverage property we suggest VaR backtesting through two standard econometric tests, which jointly cover more general forms of dependence than the first-order Markov chain. We employ the Ljung-Box autocorrelation test, which has the potential to detect high order linear dependencies, and the χ^2 Pearson's test of independence, which has the power against nonlinear alternatives.

The aim of the paper is to assess the ability of the proposed standard econometric methods to detect incorrect VaR models. Their assessment is performed in relation to the popular LR VaR tests. We first compare statistical properties of the examined tests through the Monte Carlo study, designed to reflect typical VaR failure setting. Basing on the results from the Monte Carlo study, we provide the empirical example, which illustrates advantages of the proposed methods. To expose differences among the backtesting procedures, we use all considered methods to evaluate S&P500 VaR forecasts produced by leading market risk models. The results demonstrate that the examined tests differ markedly in forecast evaluation and hence provide a different picture of market behavior. Contrary to the LR procedures, the proposed tests allow for distinguishing market models suitable for describing S&P500 dynamics under various market conditions. In particular they provide a more insightful conclusions about S&P500 dynamics during the market crash and in the aftermath of the crisis.

The paper proceeds as follows. Section 2 introduces the VaR model evaluation framework and details the testing procedures. Section 3 is dedicated to the Monte Carlo comparison of test properties. Section 4 gives the empirical example

[1] Apart from the mainstream approach, other trends developed in literature, in the context of backtesting VaR, include autoregression of VaR failures [10], the duration-based approach [7,5], the spectral theory [4,11] or multi-level VaR testing [9,3,12,8].

of backtesting S&P500 VaR forecasts by all examined methods and discusses differences in test outcomes.

2 VaR Test Construction

2.1 Unconditional Coverage Testing through Binomial and Normal Statistics

The VaR model evaluation framework is based on a binary failure process $I_t = \mathbf{1}_{\{R_t<VaR_p(R_t)\}}$, where R_t is the random return form a portfolio, with the continuous distribution function F_{R_t}, and VaR, at the p-tolerance level, is the p-quantile of R_t, i.e. $VaR_p(R_t) = F^{-1}_{R_t}(p)$. Under the correct VaR model, the I_t process is required to be the iid Bernoulli process with the parameter p, i.e. $I_t \overset{iid}{\sim} B(p)$. The iid Bernoulli condition consists of the postulate of unconditional coverage, referring to the unconditional probability of failure p, and the postulate of conditional coverage, requiring independence of failures.

The industry standard to verify the unconditional coverage property is the Kupiec test [14], which assumes the identical, independent Bernoulli distribution $I_t \overset{iid}{\sim} B(\pi_1)$ and checks the π_1 parameter: $H_0 : \pi_1 = p$. The parameter value is estimated through the empirical rate of violations $\hat{\pi}_1 = \frac{T_1}{T}$, where T_1 is the number of violations and T is the number of observations. The H_0 restriction is checked through the likelihood ratio statistic:

$$LR_{uc} = -2\log\frac{p^{T_1}(1-p)^{T_0}}{\hat{\pi}_1^{T_1}(1-\hat{\pi}_1)^{T_0}}, \tag{1}$$

where $T_0 = T - T_1$. With one parameter restriction, the likelihood ratio LR_{uc}, under the null, has the asymptotic $\chi^2_{(1)}$ distribution.

Under the above assumptions, the T_1 statistic, as the sum o iid Bernoulli variables: $T_1 = \sum_{t=1}^{T} I_t$, has the binomial distribution $T_1 \sim B(T, \pi_1)$. If the H_0 restriction is satisfied this distribution becomes $B(T, p)$:

$$T_1 = \sum_{t=1}^{T} I_t \sim B(T, p). \tag{2}$$

Using (2), the H_0 test may be based on the T_1 statistic and conducted through a direct reference to the binomial distribution. With the T_1 statistic the one-sided and the two-sided approaches are possible. If the sum of violations exceeds the expected value $EB(T, p) = Tp$, the VaR model is expected to underestimate the risk. Than the one-sided approach requires a right-sided $B(T, p)$ critical region. The number of violations below Tp, in turn, suggests risk overestimation and requires a left-sided critical region.

By the Central Limit Theorem, the binomial distribution converges to the normal one. Thus, for large samples, the unconditional coverage VaR test may

also by conducted with the use of the continuous normal distribution. The test statistics Z takes the form:

$$Z = \frac{T_1 - Tp}{\sqrt{Tp(1-p)}} \tag{3}$$

and, under the null, has the asymptotic standard normal distribution $N(0,1)$.

2.2 Conditional Coverage Testing through Sample Autocorrelations and Independence Property

According to the conditional coverage postulate, on top of backtesting overall failure rate, VaR model evaluation requires checking independence of failures. The standard approach to verify the conditional coverage property is the Markov test [6], which embeds the failure process within the binary first-order Markov chain. The test is formulated in terms of single-step transition probabilities. The independence condition implies that the transition probabilities π_{01} and π_{11} are equal, where π_{ij} denotes the probability of the transition of I_t from state i to state j. The null $H_0 : \pi_{01} = \pi_{11}$ is verified through the likelihood ratio statistic of the form

$$LR_{cc} = -2\log \frac{\hat{\pi}_1^{t_1}(1-\hat{\pi}_1)^{t_0}}{\hat{\pi}_{01}^{t_{01}}(1-\hat{\pi}_{01})^{t_{00}}\hat{\pi}_{11}^{t_{11}}(1-\hat{\pi}_{11})^{t_{10}}}, \tag{4}$$

where $\hat{\pi}_1 = \frac{T_1}{T_0+T_1}$, $\hat{\pi}_{01} = \frac{T_{01}}{T_0}$, $\hat{\pi}_{11} = \frac{T_{11}}{T_1}$ and T_{ij} is the empirical number of transitions form state i to state j. The likelihood ratio LR_{cc}, under the null, has the asymptotic $\chi^2_{(1)}$ distribution.

Relying on the single-step transition probabilities, the Markov test has only the potential to detect first-order dependencies. This deficiency may be made up for by employing the well known econometric Ljung-Box test, which has the power against linear alternatives of any order [15]. Application of the Ljung-Box test to a VaR failure process implies the null formulated in terms of correlation coefficients between VaR violations, $H_0 : \rho_h = 0$, $h = 1, ..., H$, $H < T$. Then the test statistic has the following form:

$$LB_H = T(T+2)\sum_{h=1}^{H}\frac{\hat{\rho}_h^2}{T-h}, \tag{5}$$

where $\hat{\rho}_h$ are sample autocorrelations of order h in the I_t process. Under the null, the LB_H statistic has the χ^2_H distribution.

Treating the failure process I_t and the lagged process I_{t-1} as two random variables observed in time, the conditional coverage postulate may be verified through checking independence of the two variables. This is equivalent to checking the null condition of the Markov test $H_0 : \pi_{01} = \pi_{11}$. The VaR testing procedure based on I_t and I_{t-1} variables, contrary to the Ljung-Box test, is limited to the first-order dependencies, however it is more general in the sense that is has the power against nonlinear alternatives. The independence property may by verified by using the χ^2 statistic in the standard independence Pearson's test

[16]. The sample of T observations of the failure process generates $T-1$ values for each of the variables I_t and I_{t-1}. Based on these observations displayed in the contingency table of the form (1), the test statistic is computed as

$$Q_{ind}^{\chi^2} = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{(T_{ij} - \hat{T_{ij}})^2}{\hat{T_{ij}}}, \tag{6}$$

where $\hat{T_{ij}} = \frac{T_{i\bullet}T_{\bullet j}}{T-1}$, $i,j = 0,1$. Under the independence condition this statistic has the asymptotic $\chi^2_{(k)}$ distribution. The number of degrees of freedom k may be computed from the dimensions of the contingency table. If r denotes the number of rows and c – the number of columns, then k is equal to the number of cells, given by rc, minus the reduction in degrees of freedom $p = r + c - 1$ which results from fitting the independent model parameters. Eventually the number of degrees of freedom reduces to $k = (r-1)(c-1)$, which for the contingency table (1) is $k = 1$.

Table 1. Contingency table for VaR independence test based on $Q_{ind}^{\chi^2}$ statistic

I_{t-1}	I_t		$T_{i\bullet}$
	0	1	
0	T_{00}	T_{01}	$T_{0\bullet} = T_{00} + T_{01}$
1	T_{10}	T_{11}	$T_{1\bullet} = T_{10} + T_{11}$
$T_{\bullet j}$	$T_{\bullet 1} = T_{00} + T_{10}$	$T_{\bullet j} = T_{01} + T_{11}$	$T-1$

3 Monte Carlo Study

In the first step of the comparative analysis of the VaR backtesting methods, we use the Monte Carlo study to evaluate their statistical properties. The study includes the size and the power of the tests, evaluated at the typical 0.05 level of significance, for sample sizes 250, 500, 750, 1000. With reference to the unconditional coverage property we compare the the binomial T_1 and the normal Z statistics, applied to a VaR failure series, to testing VaR models through the Kupiec LR_{uc} test. In accordance to the conditional coverage property we evaluate the properties of the Ljung-Box LB_5 test and the Pearson's independence $Q_{ind}^{\chi^2}$ test in relation to the properties of the Markov-chain-based LR_{cc} procedure.

The size assessment is done through generating iid Bernoulli samples with the probability of success $p = 0.05$, equal to the chosen VaR tolerance level. The binomial T_1 test is excluded from the size comparison as it is, by definition, exact for finite samples. The rejection frequencies obtained for other unconditional

coverage tests show remarkable accuracy, understood as a compliance between the estimated size and the nominal test level of 0.05 (Tab. 2). However the normal Z statistic seems superior to the standard Kupiec procedure, since the LR_{uc} rejection frequencies signal that this test tends to be oversized for all series lengths.

Table 2. Size estimates of VaR tests

		Series length			
Property	Test	250	500	750	1000
Unconditional coverage	LR_{uc}	0.0564	0.0575	0.0569	0.0576
	Z	0.0386	0.0547	0.0461	0.0545
Conditional coverage	LR_{cc}	0.0682	0.0855	0.1199	0.1329
	LB_5	0.0580	0.0553	0.0511	0.0515
	$Q_{ind}^{\chi^2}$	0.0451	0.0382	0.0340	0.0398

Large differences in test accuracy are observed within the group of the conditional coverage tests. The LB_5 statistics seems most accurate, with rejection frequencies stable over the sample sizes and close to the nominal test size. The Markov test performs worst overall. Especially for large samples it tends to be much oversized, with rejection frequencies exceeding the nominal tests size more than twice. Thus, according to our results, this standard procedure is likely to reject correct risk models.

The power comparison is conducted in a two-stage procedure, where the first stage involves the false unconditional coverage and the second stage, while assuming the correct overall failure rate, violates the conditional coverage property. The false unconditional coverage is implemented through generating random Bernoulli numbers with the parameter p set subsequently to 0.01, 0.03, 0.07 and 0.09, different than the VaR tolerance of 0.05. In the second stage of the power comparison, violation of the conditional coverage property implies that the VaR failures are serially dependent. To stay in line with real data, this is achieved by using a constant VaR level and return data from a $GARCH$ model with controlled volatility clustering. The $GARCH$ model assumes the mean equation of the form $R_t = \sqrt{h_t}Z_t$, $Z_t \sim N(0,1)$ and the following variance specification: $h_t = \omega + \alpha R_{t-1}^2 + \beta h_{t-1}$. The volatility clustering is measured by the correlation coefficient of the squared returns ρ, set subsequently to 0.1, 0.3 and 0.5. The parameters ω and β are fixed at levels 0.01 and 0.6, respectively, while the α parameter is set to such a value that ensures the desired level of ρ.

The results show that the three unconditional coverage testing procedures are comparable in terms of the test power (Tab. 3). However the observed differences in power estimates indicate that the binomial T_1 and the normal Z tests tend to outperform the standard Kupiec LR_{uc} approach.

Table 3. Power estimates of the unconditional coverage VaR tests

Test	ρ	Series length			
		250	500	750	1000
LR_{uc}	0.1	0.905	0.993	0.999	1.000
	0.3	0.372	0.663	0.811	0.916
	0.7	0.308	0.457	0.659	0.789
	0.9	0.743	0.933	0.992	0.999
T_1	0.1	0.985	1.000	1.000	1.000
	0.3	0.373	0.663	0.811	0.916
	0.7	0.308	0.529	0.659	0.789
	0.9	0.743	0.952	0.992	0.999
Z	0.1	0.985	1.000	1.000	1.000
	0.3	0.373	0.663	0.811	0.916
	0.7	0.392	0.529	0.659	0.789
	0.9	0.815	0.952	0.992	0.999

Unlike for unconditional coverage, the results for conditional coverage show remarkable differences in test quality, understood as the power to detect serially dependent VaR failures (Tab. 4). The LB_5 autocorrelation test is most powerful in detecting $GARCH$-model-generated dependencies. It exhibits higher power than the other approaches for all levels of correlation and for all considered sample sizes. Its superiority is most visible close to the null – for 0.1 correlation experiment, where it doubles the results of the standard Markov LR_{cc} statistic. This suggests that this procedure, being generally more effective than the usual approach, may be especially useful for detecting low-scale correlation in VaR failures. The rejection frequencies of $Q_{ind}^{\chi^2}$ suggest that this procedure also outperforms the Markov LR_{cc} test. Therefore, given that the LB_H and $Q_{ind}^{\chi^2}$ exhibit power against different alternatives, the backtesting based on the two tests instead of the standard LR_{cc}, seems most effective in detecting incorrect VaR models.

4 Empirical Example

The empirical study uses S&P500 data to compare the proposed methods, applied to testing properties of VaR failures, in relation to the standard VaR backtesting procedure. The study is conducted with the view on showing differences in practical conclusions from risk analysis that occur as a result of a test choice. To this end eleven market models, used to forecast daily VaR for S&P500, are evaluated subsequently by all considered tests. The range of models used to generate VaR forecasts covers both parametric and non-parametric methods. Within the parametric framework, the basic constant variance models are followed by conditional variance models with various error term specifications. We employ the normal distribution, the Student-t distribution as well as

Table 4. Power estimates of the conditional coverage VaR tests

Test	ρ	Series length			
		250	500	750	1000
LR_{cc}	0.1	0.1128	0.1565	0.1926	0.2398
	0.3	0.3682	0.5830	0.7410	0.8405
	0.5	0.4996	0.7518	0.8828	0.9423
LB_5	0.1	0.2033	0.2996	0.3815	0.4514
	0.3	0.6180	0.8497	0.9418	0.9779
	0.5	0.7695	0.9474	0.9867	0.9958
$Q_{ind}^{\chi^2}$	0.1	0.2106	0.3175	0.3785	0.4497
	0.3	0.5472	0.7732	0.8803	0.9402
	0.5	0.6688	0.8828	0.9526	0.9786

the Picks over Thresholds (POT) method, which, through the Extreme Value Theory, uses the Generalized Pareto Distribution. The conditional variance is modeled through the $GARCH$-class processes, to ensure representation of the volatility clustering phenomenon. To allow also for an asymmetric volatility response, relating to upward and downward market trends, we use the asymmetric GJR-$GARCH$ models. Within nonparametric methods we employ the historical simulation model and the filtered historical simulation technique, with filtering based on $GARCH$ model residuals.

The empirical analysis is based on S&P500 daily close-to-close log returns dating back to 2006. The 2006-2017 sample is divided into three subperiods: 2006-2009, 2010-2013, 2014-2017. In this way we obtain three testing samples of around 1000 observations each, which corresponds to the longest sample size examined in the simulation study. The samples are characterized by distinct volatility regimes, as shown by the descriptive statistics (Tab. 5). This allows us to compare the examined procedures in various market conditions.

Table 5. Descriptive statistics of daily S&P500 returns

	Statistic					
Period	Mean	Std deviation	Minimum	Maximum	Skewness	Kurtosis
2006-2009	-0.0001	0.0167	-0.0947	0.1096	-0.22	11.2
2010-2013	0.0005	0.0107	-0.0690	0.0463	-0.47	7.3
2014-2017	0.0004	0.0077	-0.0402	0.0383	-0.41	6.1

The values of the descriptive statistics for the first sample indicate crisis-driven behavior, which manifests itself through the high volatility and the excess kurtosis. The market crash results also in a low mean, an extremely low minimum and a relatively high maximum. The 2010-2013 and the 2014-2017 samples are

closer to each other in terms of the descriptive statistics. However the post-crisis sample of 2010-2013 still shows signs of an increased volatility regime, with larger volatility, higher kurtosis and a wider range of return values.

The results of the empirical analysis show how statistical properties translate into differences in test outcomes (Tab. 6-8). The results are particularly distinct among the tests when verifying the conditional coverage property. While the differences in p-values of the unconditional coverage tests do not result in different decisions about model choice, the conditional coverage tests imply diverse practical conclusions.

Table 6. $P-$values in VaR tests for daily S&P500 returns, 2006-2009

	Test					
VaR model	LR_{uc}	T_1	Z	LR_{cc}	LB_5	$Q_{ind}^{\chi^2}$
HS	0.007*	0.003*	0.002*	0.145	0.000*	0.128
FHS	0.597	0.267	0.296	0.736	0.028*	0.951
Normal	0.000*	0.000*	0.000*	0.361	0.000*	0.391
GARCH-normal	0.010*	0.004*	0.003*	0.120	0.004*	0.177
GJR-GARCH-normal	0.005*	0.002*	0.001*	0.095	0.161	0.147
Student-t	0.000*	0.000*	0.000*	0.002*	0.000*	0.002*
GARCH-Student-t	0.007*	0.003*	0.002*	0.107	0.005*	0.162
GJR-GARCH-Student-t	0.001*	0.000*	0.000*	0.057	0.109	0.100
POT	0.000*	0.000*	0.000*	0.466	0.000*	0.576
GARCH-POT	0.001*	0.000*	0.000*	0.198	0.095	0.257
GJR-GARCH-POT	0.005*	0.002*	0.001*	0.095	0.224	0.147

Notes: Outcomes significant at 5% level are marked with *.

The Kupiec LR_{uc}, the binomial T_1 and the normal Z tests all indicate that the vast majority of models are not reliable at times of the market crash, for the number of VaR failures does not correspond to the assumed level of tolerance (Tab. 6). The only model suitable for the high volatility period, in accordance to the unconditional coverage property, is the filtered historical simulation model. This model reflects the volatility dynamics in the form of the $GARCH$ process, while being non-parametric in the sense that it does not assume any parametric distribution of the $GARCH$ residuals. This shows that distributional assumptions used in forecasting VaR for highly volatile data tend to produce excessive number of VaR failures. This result is specific to the market crash sample. The same tests applied to the samples of 2010-2012, 2014-2017 admit most of the considered models (Tab. 7-8). Especially all $GARCH$-class models are classified as correct with respect to the overall failure rate.

The conditional coverage tests complement the VaR backtesting procedure, indicating models that produce serially dependent failures. Within this group of methods, the results strongly depend on a test choice. The Markov LR_{cc} test p-values admit nearly all the models in all periods, not distinguishing models

that best represent the market dynamics. In particular, this test fails to reject most of the constant-variance models. Contrary to this result, the Ljung-Box LB_5 test indicates models that are appropriate in terms of the conditional coverage property as well as it shows differences among the samples. For the market crash sample it admits only the narrow class of $GJR\text{-}GARCH$ models, which account for volatility asymmetry. The p-values of this test suggest also that, under a high volatility regime, the heavy-tailed POT models, which use the Extreme Value Theory and the Generalized Pareto Distribution, outperform other models. For the other samples, which do not include extremal price movements, all $GARCH$-class models are suitable, in the sense that they do not produce correlated VaR failures. The LB_5 test indicates therefore that the asymmetry volatility component is vital for describing S&P500 dynamics only for the crisis time. The above conclusions about risk forecasts provide an insight into investors' behavior connected with going through a financial crisis. They suggest that the high, crisis-driven volatility, influences trading patterns, stimulating more violent reactions for dropping prices.

Table 7. $P-$values in VaR tests for daily S&P500 returns, 2010-2013

	Test					
VaR model	LR_{uc}	T_1	Z	LR_{cc}	LB_5	$Q^{\chi^2}_{ind}$
HS	0.352	0.202	0.181	0.158	0.001*	0.118
FHS	0.965	0.520	0.483	0.642	0.799	0.745
Normal	0.850	0.462	0.425	0.291	0.001*	0.273
GARCH-normal	0.850	0.462	0.425	0.676	0.457	0.791
GJR-GARCH-normal	0.629	0.349	0.317	0.733	0.953	0.889
Student-t	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
GARCH-Student-t	0.597	0.267	0.296	0.498	0.440	0.576
GJR-GARCH-Student-t	0.276	0.119	0.133	0.015*	0.595	0.678
POT	0.000*	0.000*	0.000*	0.281	0.011*	0.177
GARCH-POT	0.165	0.099	0.089	0.515	0.512	0.588
GJR-GARCH-POT	0.123	0.075	0.068	0.551	0.822	0.625

Notes: Outcomes significant at 5% level are marked with *.

The Pearson's $Q^{\chi^2}_{ind}$ independence test generally admits most of the models, apart from the least volatile period of 2014-2017. In this period, in line with the LB_5 test, it classifies only the $GARCH$-class models as suitable in terms of independence of VaR failures. Since this test, unlike the LB_5 statistics, has the power only against the first-order dependence, the combined results of the Pearson's $Q^{\chi^2}_{ind}$ and the LB_5 tests suggest that the first-order dependence is typical for calm periods in the market. Under high volatility regimes, in turn, VaR violations depend on higher-order lagged values, rather than being influenced only by the last observation. This result confirms that it is essential that the VaR

backtesting procedures be based on more informative measures of dependence that the first order Markov chain parameters, which are used in the popular Markov LR_{cc} test.

Table 8. $P-$values in VaR tests for daily S&P500 returns, 2014-2017

	Test					
VaR model	LR_{uc}	T_1	Z	LR_{cc}	LB_5	$Q_{ind}^{\chi^2}$
HS	0.591	0.330	0.298	0.071	0.000*	0.04*
FHS	0.591	0.330	0.298	0.746	0.079*	0.924
Normal	0.959	0.442	0.480	0.043*	0.000*	0.021*
GARCH-normal	0.253	0.148	0.133	0.734	0.231	0.862
GJR-GARCH-normal	0.809	0.442	0.405	0.697	0.775	0.824
Student-t	0.003	0.001*	0.001*	0.010*	0.000*	0.004*
GARCH-Student-t	0.591	0.330	0.298	0.746	0.509	0.924
GJR-GARCH-Student-t	0.236	0.101	0.112	0.692	0.415	0.847
POT	0.108	0.066	0.060	0.076	0.007*	0.038*
GARCH-POT	0.492	0.278	0.249	0.759	0.450	0.976
GJR-GARCH-POT	0.809	0.442	0.405	0.584	0.775	0.640

Notes: Outcomes significant at 5% level are marked with *.

5 Conclusion

The study tackled the issue of powerful risk model validation. We referred to the Basel framework, which recommends testing risk models based on the VaR measure, and investigated statistical properties of the VaR tests. In this context we compared standard likelihood ratio approaches to other econometric methods applicable to a binary VaR failure process. We applied the binomial and the normal statistics to testing unconditional coverage and for conditional coverage we employed the Ljung-Box and the Pearson's χ^2 statistics. Using simulations we confirmed superiority of the proposed procedures, it terms of their power, over the industry standards. Through the empirical example based on S&P500, we demonstrated differences in risk analysis attributable to the choice of the backtesting method. We showed that the results diverged markedly among the tests. The more powerful tests, in comparison to the standard procedures, allowed for distinguishing models that best characterized index dynamics under various market conditions. Contrary to the LR approach, the proposed procedures pointed out the $GARCH$-class models as appropriate for S&P500, in the sense that they prevented clustering of VaR violations. For the market crash period only the narrow class of GJR-$GARCH$-class models turned out acceptable. This suggested that the high, crisis-driven volatility influenced investors' behavior, stimulating more violent reactions for dropping prices. Comparison of these

results to the outcomes of the standard VaR testing procedure demonstrated that the proposed methods allowed for a more effective detection of incorrect risk models and hence gave a more insightful description of investors' behavior.

Acknowledgments. The research was supported by the Polish National Science Centre grant DEC-2013/11/N/HS4/03354.

References

1. Basel Committee on Banking Supervision: Minimum Capital Requirements for Market Risk (2016). `http://www.bis.org/bcbs/publ/d352.pdf`
2. Basel Committee on Banking Supervision: High-level summary of Basel III Reforms (2017) `https://www.bis.org/bcbs/publ/d424_hlsummary.pdf`
3. Berkowitz, J.: Testing Density Forecasts with Applications to Risk Management, J. Bus. Econ. Stat. **19(4)**, 465–474 (2001). doi: 10.1198/07350010152596718
4. Berkowitz, J., Christoffersen P., Pelletier D.: Evaluating Value-at-Risk Models with Desk-Level Data. Manage. Sci. **12(57)**, 2213–2227 (2011). doi: 10.1287/mnsc.1080.0964
5. Candelon, B., Colletaz, G., Hurlin, C., Tokpavi S.: Backtesting Value-at-Risk: a GMM duration-based test. J. Financ. Economet. **9(2)**, 314–343 (2011). doi: 10.1093/jjfinec/nbq025
6. Christoffersen, P.: Evaluating Interval Forecasts. Int. Econ. Rev. **39(4)**, 841–862 (1998). doi: 10.2307/2527341
7. Christoffersen, P., Pelletier, D.: Backtesting Value-at-Risk: A Duration-Based Approach. J. Financ. Economet. **1(2)**, 84–108 (2004). doi: 10.1093/jjfinec/nbh004
8. Colletaz, G., Hurlin, C., Perignon, C.: The Risk Map: a New Tool for Risk Management. J. Bank. Financ. **37(10)**, 3843–3854 (2013). doi: 10.1016/j.jbankfin.2013.06.006
9. Crnkovic, C., Drachman, J.: Quality Control in VaR. In: Grayling, S. (eds.) VaR: Understanding and Applying Value-at-Risk. Risk Publications, London (1997)
10. Engle, R.F., Manganelli, A.: CAViaR: Conditional Autoregressive Value-at-Risk by Regression Quantiles. J. Bus. Econ. Stat. **22(4)**, 367–381 (2004). doi: 10.1198/073500104000000370
11. Gordy, M.B., McNeil, A.J.: Spectral Backtests of Forecast Distributions with Application to Risk Management. In: Finance and Economics Discussion Series 2018-021. Board of Governors of the Federal Reserve System, Washington (2018). doi: 10.17016/FEDS.2018.021
12. Hurlin, Ch., Tokpavi, S.: Backtesting value-at-risk accuracy: a simple new test. J. Risk **9(2)**, 19–37, (2007). doi: 10.21314/JOR.2007.148
13. JP Morgan: Riskmetrics. Technical document, Morgan Guaranty Trust Company (1994)
14. Kupiec, P.: Techniques for Verifying the Accuracy of Risk Measurement Models. J. Deriv. **3(2)**, 174–184 (1995). doi: 10.3905/jod.1995.407942.
15. Ljung, G.M., Box, G.E.P.: On a Measure of Lack of Fit in Time Series Models. Biometrika **65(2)**, 297–303 (1978). doi: 10.1093/biomet/65.2.297
16. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **50(302)**, 157–175 (1900). doi: 10.1080/14786440009463897

The Long-term memory effects of the Baltic Dry Index

J. Ramon San Cristóbal

Nautical Sciences Department. University of Cantabria, Spain.
Jose.sancristobal@unican.es

Abstract. Even the most traditional approaches to financial time series analysis begin by assuming that the underlying dynamics are primarily random. Under this hypothesis, financial systems were usually modeled by a random walk type process with no long-term memory. Subsequent researchers indicated the even the most competitive and complex markets are not exclusive random suggesting the occurrence of memory in the data and the presence of characteristics that influence the signal trend over long periods. Processes that we might naively assume purely white noise sometimes turn out to exhibit long memory processes, sometimes referred to as colored noise in contrast to white noise. In this paper we raise the question of whether the Baltic Dry Index (BDI) is a persistent process with long-run memory effects and then consistent with the random walk assumption. With this aim, we apply the Hurst exponent to the BDI series from January 2011 to January 2016. The Hurst exponent is a classical test to detect long memory in time series and a measure of stock market efficiency and predictability. Our results show that the BDI is not totally random. The large Hurst exponent indicates a strong trend structure which can be used by forecasting models to forecast more accurately than series with Hurst exponent values close to random series.

Keywords: Baltic Dry Index; Hurst Exponent; R/S analysis

1 Introduction

Efficient market hypothesis has attracted a considerable number of studies in empirical finance, particularly in determining the market efficiency of the capital market. The purpose of the fractal market hypothesis is to investigate whether the selected market abides by a particular evolution pattern of the random walk hypothesis. The Baltic Dry Index (BDI) is one of the leading indicators of world irregular bulk shipping market. The index is a composite of three sub-indices, the Baltic Capesize Index, Baltic Panamax Index, and Baltic Handymax Index, which reflect merchant shipping tonnage and routing using information gathered by the Baltic Exchange.

Owing to the magnitude of investment required by the shipping industry, understanding the volatility and long memory of BDI is crucial. Volatility is not only closed to the characters of uncertainty and risk, which is important for the companies to manage risk, volatility is also the key variable in assets pricing. Volatility-related decisions vary from what kind of contract, time/voyage charter, to engage in to whether to enter/exit the business [1]. Freight rate fluctuations create extreme operational risk and uncertainty for the dry bulk shipping industry [2]. Thus, better understanding of these fluctuations and influences on freight rates can reduce the operational risks to shipping companies.

The BDI and the dry bulk market have attracted the attention of many authors. Cullinane, Mason, and Cape [3] made a comparison of forecasting models for the BDI. Wong, Tu, and Fan [4] investigated the influence of the BDI on the bulk shipping industry integrating grey relational analysis with the entropy method to construct a composite index combination. Bulut, Duru, and Yoshida [5] developed a fuzzy integrated logical forecasting model of time charter rates in dry bulk shipping. Chen, Meersman, and van de Voorde [6] employed time series models to forecast spot rates at main routes for three types of dry bulk vessels. Chung and Weon [7] employed two versions of bivariate asymmetric mixed normal GARCH models to capture the skewness and kurtosis detected in both the conditional and unconditional return distributions of dry bulk freight rates. Chang, Chou and Wu [8] applied value-at-risk models to evaluate the risk of dry bulk freight rates and Zeng, Qu, Ng, and Zhao [9] developed a new approach for BDI forecasting based on empirical mode decomposition and neural networks.

Studies of long-term memory of BDI are, however, scarce. Zhang, Wei, and Zhang [10] used value-at-risk of spot freight rates with exponential autoregressive conditional heteroskedastic to investigate fluctuations of the BDI in order to model long memory of the index. Gong and Jing [11] applied a long memory stochastic volatility model to study the volatility and long memory in the shipping industry and Leonov and Nikolov [1] studied the fluctuations in the freight rates of two Baltic Panamax routes using a hybrid model of wavelets and neural networks.

Apart from the cited papers there is a scarcity of literature regarding volatility and long memory of BDI. In this paper we raise the question of whether the BDI is efficient

or it is a persistent process with long-run memory effects and then consistent with the random-walk assumption. Answering these questions can be of prime importance in the understanding of interactions among brokerage, international events, and free-market forces. With this aim, we apply the Hurst exponent to the BDI series from January 2011 to January 2016 [12].

2 The Hurst exponent

Even in the most traditional approaches to financial time series analysis there has been a huge discussion about whether random walk process and the like represent accurately evolution of financial data [13], [14]. Almost all statistical analysis of economic and financial systems begin by assuming that the dynamics are primarily random. That is, there is an underlying process that has many component parts and the interactions of those components is so complex that a deterministic explanation is not possible. In the early studies, the system was usually modeled by a random-walk-type process with no long-term memory, which implies that the events measured are independent and identically distributed. Physically, this means that the events must not influence one another and they must be equally likely to occur [14].

The hypothesis was first questioned by Mandelbrot [15] who proposed that financial return did indeed possess a long memory that may well be described by Fractional Brownian Motion [16]. Subsequent research showed that even the most competitive and complex markets are not exclusive random, and moreover that simple random processes cannot describe the dynamics of economic and financial systems [14]. Since then, the study of long memory, the presence of characteristics that influence the signal trend over long periods, caught the interest of researchers during the seventies [17].

The estimation of long-term memory in a data set is a key feature of nonlinear signal processing. It is also of prime importance in the understanding of interactions among brokerage, international events, free-market forces, and so on, [14]. The Hurst exponent, which is a classical test to detect long memory in time series and a measure of stock market efficiency and predictability [18], can be estimated using three methods: average rescaled range (*R*/*S*) analysis, detrended fluctuation analysis (DFA); and periodogram regression [17]

R/*S* analysis is the eldest and best-known method to estimate the Hurst exponent. The method was proposed by Mandelbrot and Wallis [19] based on the previous work of Hurst [20] in the mid-20th century when he worked on a flood control project in the River Nile. The two factors used in the *R*/*S* analysis are: (*i*) the difference between the maximum and the minimum cumulative values; and (*ii*) the standard deviation from the observed values. The procedure to apply the *R*/*S* analysis over multiple regions of the data is as follows:

1. Assume a time series $X = X_1, X_2, \cdots . X_N$.
2. Calculate the mean $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$.

3. Calculate the mean adjusted series $Y_t = X_1 - \mu$, $t = 1, 2, \cdots . N$.
4. Calculate the cumulative deviate series $Z_t = \sum_{t=1}^{N} Y_t$.
5. Calculate the range series, $R_t = max(Z_1, Z_2, \cdots, Z_N) - min(Z_1, Z_2, \cdots, Z_N)$, $t = 1, 2, \cdots . N$, where u is the mean value from X_1 to X_t..
6. Calculate the standard deviation series, $S_t = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(X_i - u)^2}$, $t = 1, 2, \cdots . N$.
7. Calculate the rescaled range series $(R/S)_t = \frac{R_t}{S_t}$, $t = 1, 2, \cdots . N$.

Note that $(R/S)_t$ is averaged over the regions $[X_1, X_t]$, $[X_{t+1}, X_{2t}]$, until $[X_{(m-1)t+1}, X_{mt}]$ where $m = N/t$. To use all data for calculation, a value of t is chosen that is divisible by N. The relationship between $(R/S)_t$ and the Hurst exponent is given by:

$$(R/S)_t = Ct^H \qquad (1)$$

where C is a constant. The R/S value scales as we increase the time increment by a power law value which equals H, the Hurst exponent which can be estimated by:

$$\log(R/S)_t \approx \log C + H \log t \qquad (2)$$

$$H = \log(R/S)/\log(t) \qquad (3)$$

To estimate the Hurst exponent, following the method developed by Mandelbrot et al., (1968; 1979), a linear regression line through a set of points, composed of the log of the size of the areas on which the average rescaled range is calculated and the log of the average rescaled range over a set of regions of size n, is calculated. The slope of the regression line, obtained using the ordinary least squares, is the estimate of the Hurst exponent.

The values of the Hurst exponent lie between 0 and 1. For a white noise, $H = 0$ while an exactly self-similar process would be characterized by $H = 1$. A value of $H = 0.5$ indicates a true random process (sometimes referred to as fractional brownian motion) where the observations are uncorrelated and there is no correlation between any element and a future element. A Hurst exponent $0.5 < H < 1$ indicates "persistent behavior" with positive correlations, which is characterized by long-run memory and trend reinforcing *series* affecting all time scales. The current value of the series is affected not only by its most recent values but also by its long-term history. For this reason, the process is often referred to as having "long-memory" or "long-range dependence". This means that all index fluctuations are correlated with all future index fluctuations. Persistence implies that if the index has been up or down, then the chances are that it will continue to be up or down in the future respectively. If there is an increase from time step t_{i-1} to t_i, there will probably be an increase from t_i to t_{i+1}. The same is true of decreases, where a decrease will tend to follow a decrease. The strength of persistence increases as H approaches 1. A Hurst exponent value of $0 < H < 0.5$ will exist for a

process with “anti-persistent behavior”, or negative autocorrelation. The process has short range dependence, where the correlations sum up to zero. Here an increase will tend to be followed by a decrease. Whenever the index has been up, it is more likely to be down in the close future or viceversa. This behavior is sometimes called “mean reversion”. In this sense, an antipersistent process will display wilder oscillations than a pure random walk. It is important to note that persistent stochastic processes have little noise whereas antipersistent processes show presence of high-frequency noise.

3 Application

In this section we apply the rescaled range analysis to the BDI series from January 2011 to January 2016 shown in Fig. 1. First, the rescaled range is calculated for the entire BDI time series. Then, the rescaled range is calculated for the two halves of the series, resulting in R/S_0 and R/S_1. These two halves are averaged, resulting in R/S_{Ave1}. The process continues by dividing each of the previous sections in half and calculating the rescaled range for each new section. The rescaled range values for each section are then averaged.

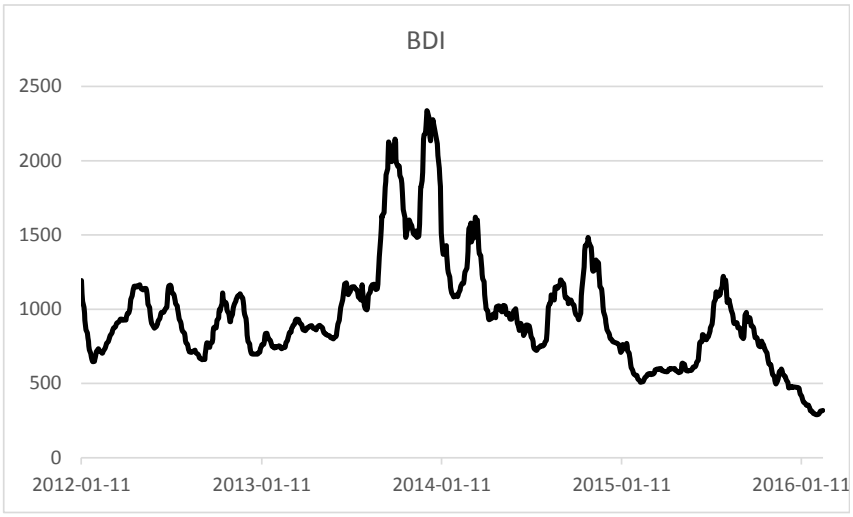


Fig. 1. Baltic Dry Index

To estimate the Hurst exponent, a vector of points is created, where x_i is the log of the size of the data region used to calculate R/S_{Avei} and y_i is the log of the R/S_{Avei} value. Table 1 shows these x_i and y_i values used to create the vector of points which is also

represented in Fig. 2. Next, the Hurst exponent is estimated through a linear regression line of the form $y = a + bx$ where the term a is the y-intercept and b is the slope of the line. For the data in Table 1 the linear regression line results in $y = -0.9284 + 0.93x$. In this case the Hurst exponent is $H = 0.93$.

Table 1. x_i and y_i values

Region size (N)	R/S_{Ave}	log(N)	log(R/S_{Ave})
1024	314.24	10	8.2957
512	177.42	9	7.4710
256	85.16	8	6.4121
128	51.84	7	5.6961
64	27.31	6	4.7716
32	13.36	5	3.7403
16	6.82	4	2.7692
8	2.74	3	1.7724

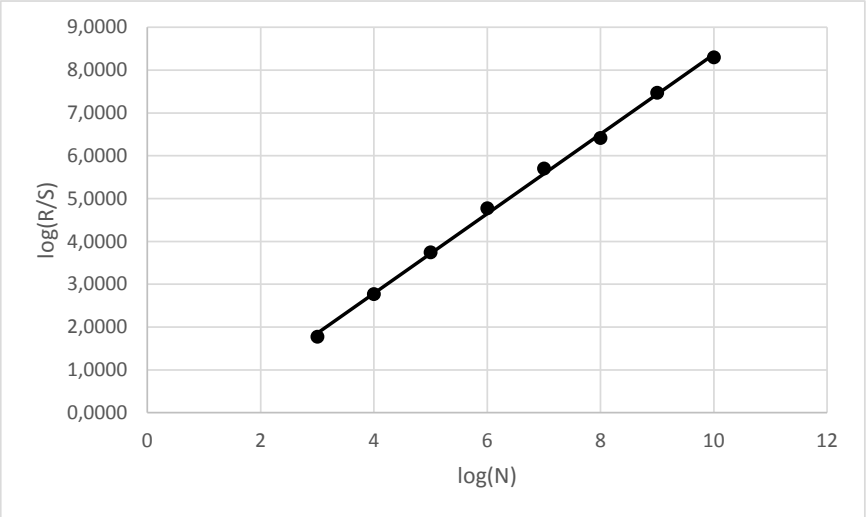


Fig. 2. Regression Line

4 Conclusions

In this paper we raise the question of whether the BDI is a persistent process with long-run memory effects and then consistent with the random walk assumption. Our results show that the BDI is not totally random suggesting the occurrence of memory in the data and the presence of characteristics that influence the signal trend over long periods. The large Hurst exponent, $H = 0{,}93$, indicates a long range dependence of the series and a strong trend structure. The fact that Baltic stock indices display long memory has important consequences. The main implication of this circumstance is that an efficient market hypothesis can be rejected because the BDI does not follow a random walk. If the BDI shows long memory means that predictability is a possibility. Since predictability exists at periods during which the process is less noisy, our results can be used to build models for prediction and forecast more accurately than series with *H* values close to random values.

References

1. Leonov Y, Nikolov V (2012) A wavelet and neural network model for the prediction of dry bulk shipping indices. Marit Econ Logist 14(3): 319-333.
2. Veenstra AW, Franses PH (1997) A Co-integration Approach to Forecasting Freight Rates in the Dry Bulk Shipping Sector. Transport Res A 31(6): 447-458.
3. Wong, HS, Tu, Hand Y, Fan SH (2010) Integrated GRA with Entropy method to evaluate factors affecting Baltic Dry Index. 4th International Conference on Genetic and Evolutionary Computing. IEE Computer Society, 493:496.
4. Cullinane KPB, Mason KJ, Cape M (1999) A comparison of models for forecasting the Baltic Freight Index: Box-Jenkins revisited. Int J Marit Econ 1: 15-39.
5. Bulut E, Duru O, Yoshida S (2012) A fuzzy integrated logical forecasting (FILF) model of time charter rates in dry bulk shipping: A vector autoregressive design of fuzzy time series with fuzzy c-means clustering," Marit Econ Logist 14: 300-318.
6. Chen S, Meersman H, van de Voorde E (2012) "Forecasting spot rates at main routes in the dry bulk market. Marit Econ Logist 14: 498-537.
7. Chung SK, Weon JH (2013) Long memory and asymmetric time varying spillover effects in dry bulk freight markets. Marit Econ Logist 15: 494-522.
8. Chang CH, Chou HC, Wu CC (2014) Value-at-risk analysis of the asymmetric long-memory volatility process of dry bulk freight rates. Marit Econ Logist 16: 298-320.
9. Zeng Q, Qu C, Ng AKY, Zhao X, (2015) A new approach for Baltic Dry Index forecasting based on empirical mode decomposition and neural networks. Marit Econ Logist doi:10.1057/mel.2015.2.
10. Zhang H, Wei F, Zhang Z (2008) Modeling volatility of Baltic Dry Bulk Freight rates. Proceedings of the IEEE International Conference on Automation and Logistics, Qingdao, China, September.

11. Gong XX, Jing LV (2009) Long memory on dry bulk volatility in the shipping market. International Conference on Service Operations. Logistics and Informatics.
12. Quandl Financial and Economic data. www.quandl.com. March, 2016.
13. Lo AW, MacKinlay AC (1999) A Non-random Walk down Wall Street. Princeton, Princeton University Press.
14. Álvarez-Ramirez J, Cisneros M, Ibarra-Valdez C, Soriano A (2002) Multifractal Hurst analysis of crude oil prices. Physica A 313: 651-670.
15. Mandelbrot B (1963) The variation of certain speculative prices. J. Business 36: 394–419.
16. Couillard M, Davison M (2005) A comment on measuring the Hurst exponent of financial time series. Physica A 348: 404-418.
17. Sánchez Granero MA, Trinidad Segovia JE, García Pérez J (2008) Some comments on Hurst exponent and the long memory processes on capital markets. Physica A 387: 5543-5551.
18. Tzouras S, Anagnostopoulos C, McCoy E (2015) Financial time series modeling using the Hurst exponent. Physica A 425: 50-68.
19. Mandelbrot B, Wallis JR (1969) Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. Water Resour. Res 5: 967–988.
20. Hurst H (1951) Long term storage capacity of reservoirs," *Trans. Amer. Soc. Civ. Eng.*, vol. 6, pp. 770–799.

Forecasting Peak Period of Travel Time

Béla Paláncz[1], Jianhong (Cecilia) Xia[2*], and Yuchen Liu[2]

[1]Department of Geoinformatics, Budapest Technical University, HU
palancz@epito.bme.hu
[2]School of Earth and Planetary, Curtin University, Perth, Australia
*c.xia@curtin.edu.au, yuchen.liu@postgrad.curtin.edu.au

Abstract. In this contribution, recurrent neural network is applied to forecast the duration of Peak Period (PP) indicating road congestion in case of a traffic flow. The method employs time series data of PP as well as its Starting Time point (ST). The recurrent network has one hidden layer with 3 nodes having different activation functions, like sigmoid, Hardy spline, Weinland, Laguerre-Gauss and Thin Plate Spline functions. The network is able to estimate the length of the PP of a specific day as output, based on the length of the peak period of its last two days, the ST of the peak of that specific day and its previous day as inputs. The error of the network of employing different activation functions is demonstrated. The method is illustrated by numerical example based on real field data collected from the Perth Metropolitan Area in Western Australia.

Keywords: Road congestion, Peak period, Recurrent neural networks, Forecasting

1 Introduction

Traffic congestion has been commonly recognized as a major urban transport problem, which causes a variety of economic, social and environmental impacts such as lot of productivity, road rage, vehicle crashes, and noise and air pollution [1]. A rush hour or peak period is a period of time, when traffic congestion on road reaches to the highest value [2]. It usually occurs in the early morning when people go to work, so called morning peak, or in the late afternoon when people go back home, afternoon peak. Fig. 1 shows that to avoid heavy congestion during a peak period (PP), the travel demand may shift from a critical peak time to its shoulders, which will make PP longer [3].

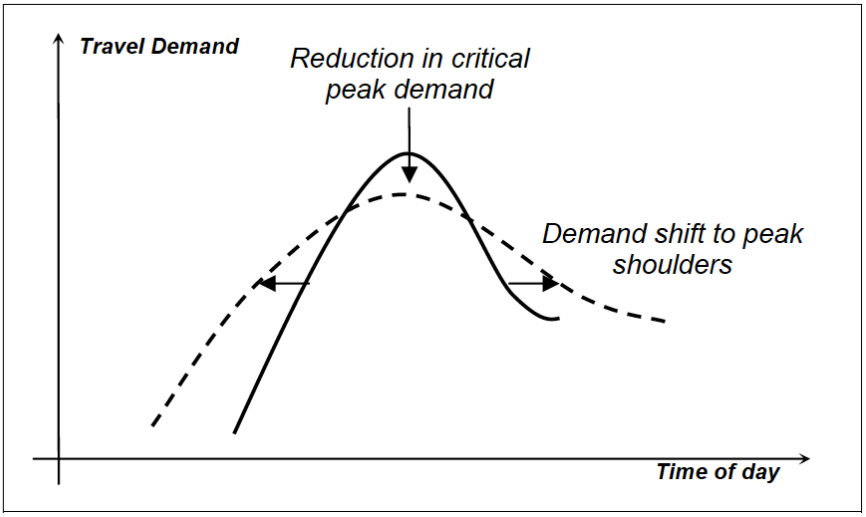


Fig. 1. Representative effect of peak spreading on AM travel demand profile (Reference: [4])

Many methods have been developed for measuring a peak period or spreading, for example, the Peak Period Efficiency Ratio (PSER) [5], which is the ratio of total traffic to the maximum possible flow in period, such as 7:00 am - 9:00 am for measuring the range of peak spreading and how it changes over time. Some studies applied volume to capacity ratio to understand peak spreading [6]. Various modeling efforts have been made to address the problem of short-term traffic forecasting, which focus on the estimation of the traffic characteristics, like PP, on daily horizon or shorter. Due to stochastic nature of traffic flow and highly nonlinear characteristics for short term prediction, time series models, artificial intelligence techniques have received much attention, see i.e. [7,8], besides other techniques, as local regression i.e. [9] and support vectors machine, see [10]. Recently even image processing view has been proved to be very effective [11]. The aim of this study is to develop a simple measure of PP and employ dynamical neural network to forecast the PP.

In the next two sections we describe data collection methods and the model for their processing. This is followed by derivation of the structure of the recurrent neural network and its implementation in Mathematica system. Then the results of the investigations with different activation functions are presented and discussed how to use network forecasting in practice. At the end we summarize the main features of our study and provide hints for the future researches.

2 Study Area

The research data was collected from the link 90-76 of Canning Hwy (eastbound) in Perth Metropolitan Area, Western Australia, between 2017.08.21 and 2017.09.13. The road link starts from the signalized intersection of North Lake Rd, site90, to the signalized intersection of Riseley St, site76, with a length of 2.17 km and a posted speed limit of 70 km/h. Each of these two intersections has a Bluetooth scanner for probe data collection. Fig. 2 shows the study area.

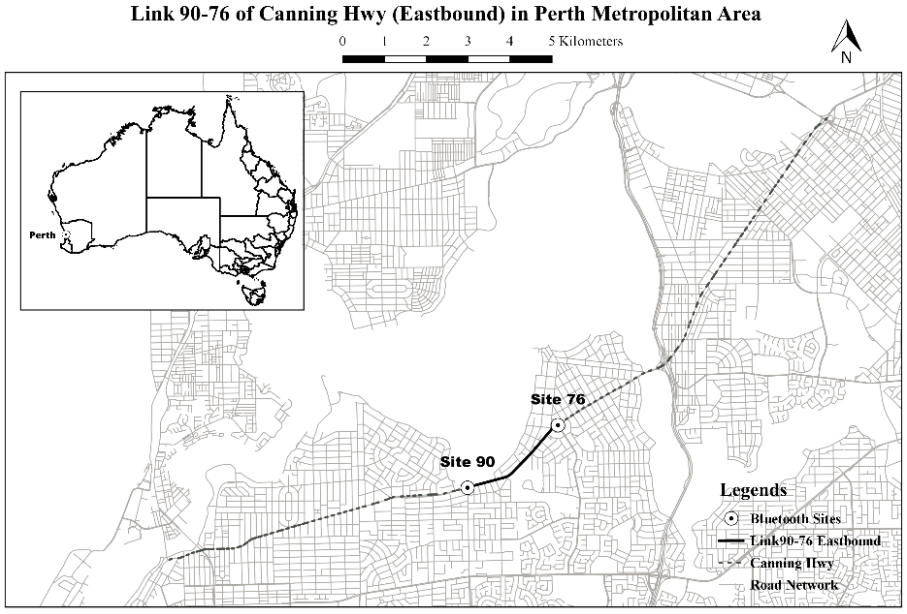


Fig. 2. Study area

3 Data Collection Methods

Bluetooth technology has been recognized as a cost-effective and feasible method to estimate real-time travel time or traffic speed and monitor the traffic condition [12]. Bluetooth scanner can record the media access control (MAC) address identifier of Bluetooth-enabled devices passing by, such as car stereos and GPS navigation devices that are in discovery mode, with the timestamp information. Therefore, if the same MAC address is recorded in two adjacent Bluetooth scanner sites, the difference in the two timestamps can be used to calculate the individual travel time and travel speed of this MAC address between these two sites. Individual travel times from various vehicles

or MAC addresses can then be used to derive the mean travel time or travel speed to represent the traffic condition of that road link between those two sites in the studying time period. In this study, the mean travel times (in seconds) of link 90-76 were calculated at 30-minute intervals, from 6 am to 12 pm.

4 Peak Period Measure

In this study, we developed a simple measure of the PP of travel time at a road segment level. The PP depicts the start time, end time and duration of traffic congestion. A typical travel-time vs. day time curve for the morning period of a day (a morning peak) can be seen on Fig. 3.

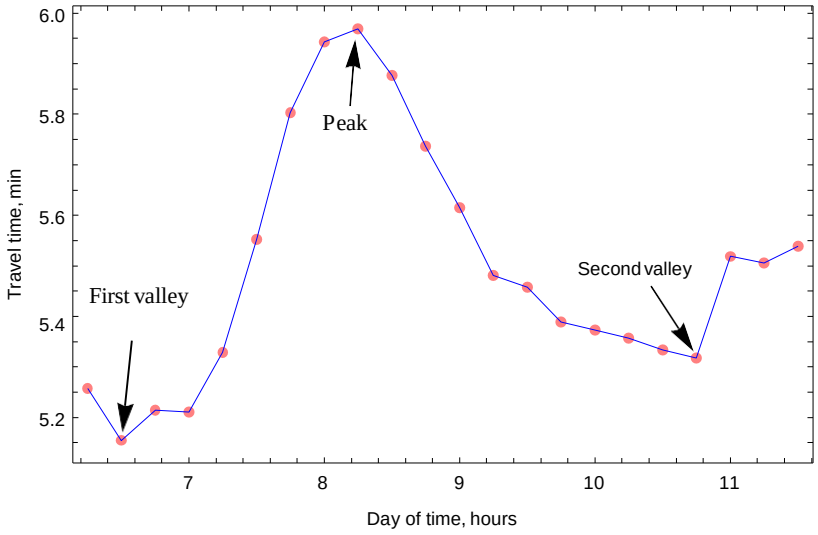


Fig. 3. Discrete travel-time values (minute) versus day-of-time (hour)

For example, a morning peak starts the half way between the first valley and the peak (starting time ST_m) and ends the half way between the peak and the second valley (ending time ET_m), see Fig. 3. The peak is the longest travel time of the morning peak travel time profile. Then the length of morning PP(PP_m) can be computed as [2],

$$ST_m = \frac{1}{2}\left(First\_valley\_time + Peak\_time\right) \quad (1)$$

$$ET_m = \frac{1}{2}\left(Peak\_time + \mathrm{Sec}ond\_valley\_time\right) \quad (2)$$

$$PP_m = ET_m - ST_m \tag{3}$$

Using the Bluetooth travel time data and the proposed PP measure, 24 consecutive days' PPs and STs have been calculated, and they are shown on Fig. 4 and Fig. 5 respectively.

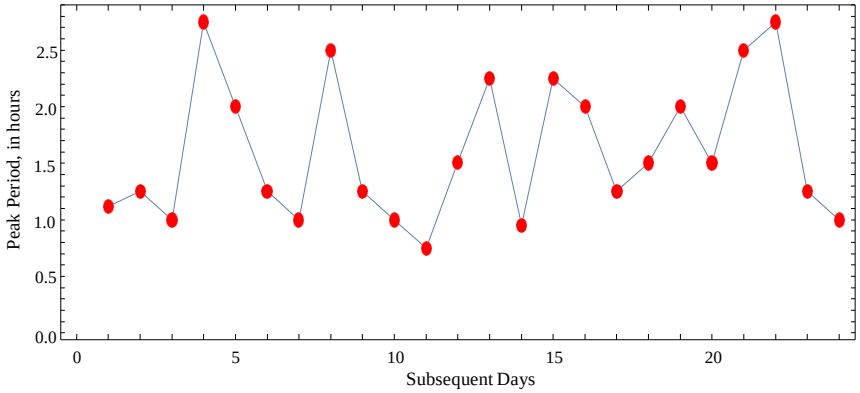


Fig. 4. Length of the Peak Period

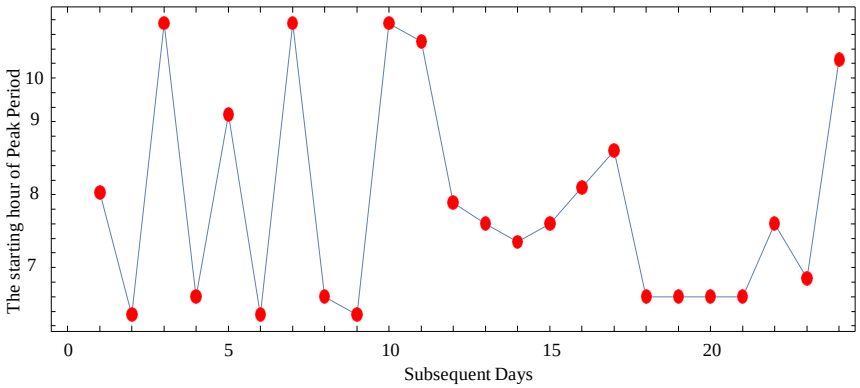


Fig. 5. Starting times of the Peak Period

5 Neural Network Model

In actual case, let us suppose the following model [13],

$$y_e(t) = \theta\left(y_e(t-1), y_e(t-2), u(t), u(t-1)\right) \tag{4}$$

where

$y_e(t)$ - the estimated length of the peak period today

$y_e(t-1), y_e(t-2)$ - the length of the peak period in the last two days

$u(t)$ - the start time of the peak period today

$u(t-1)$ - the start time of the peak period yesterday.

These relatively few data allow a modest recurrent neural network with one

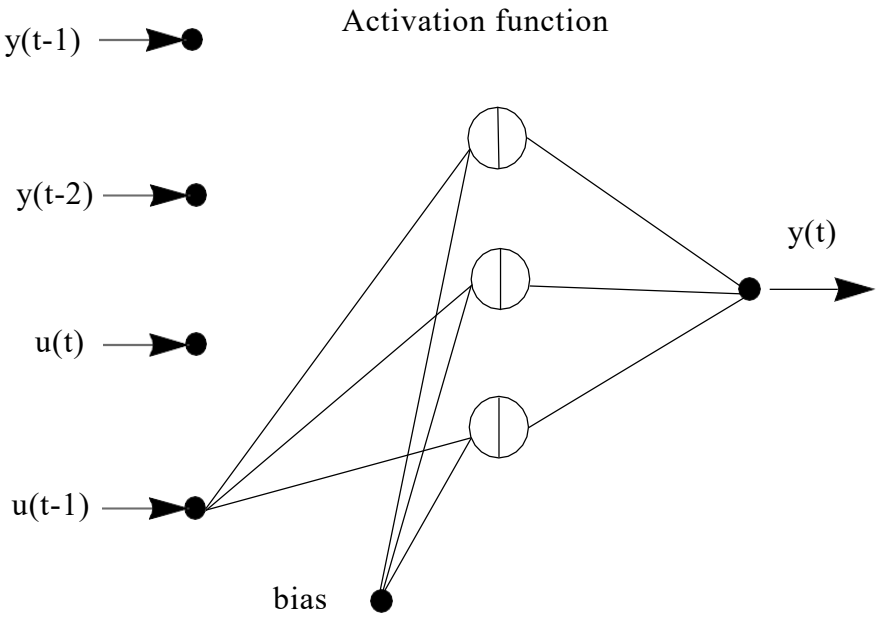


Fig. 6. Structure of the dynamical network

hidden layer containing 3 nodes, where signal transformation takes place accordingly their activation functions. Basically, we have a multiply input-single output (MISO) model.

The general structure of our network can be seen in Fig. 6.

6 Implementation of the Model

The model was implemented in Mathematica and the computations were carried out using built-in functions. The elements of the training set `trainD` have the following structure,

```
trainD[[i]]=(y[t-1],y[t-2],m[t],m[t-1])→{y(t)}.
```

The actual activation function was defined as an `ElementwiseLayer`, for example in case of Gauss- Laguerre function,

```
myGL = ElementwiseLayer[Exp[-#^2](1-3#^2+1/2#^4)&].
```

Fig. 7 shows this activation function.

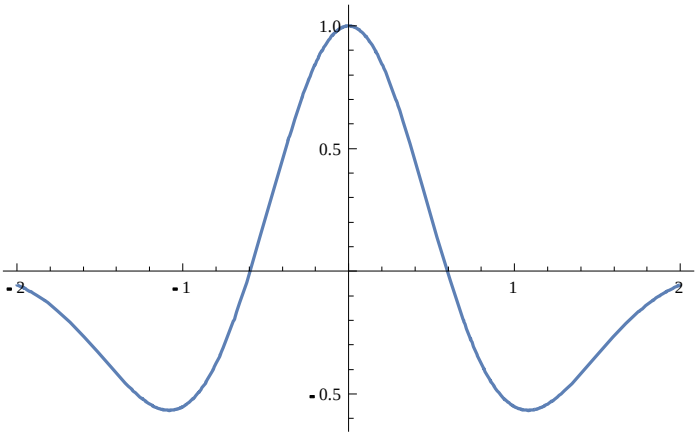
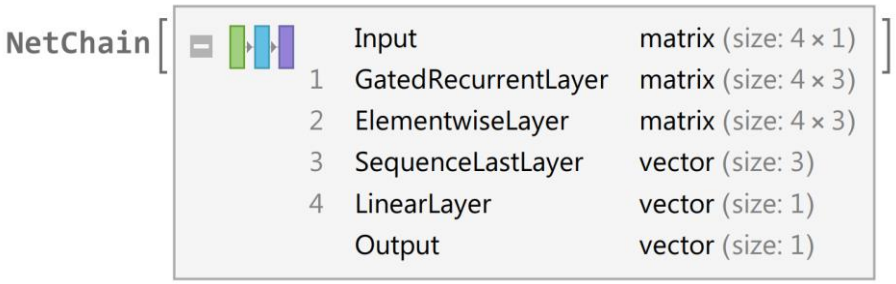


Fig. 7. The Gauss-Laguerre activation function

Let us employ the following recurrent network,

```
net = NetInitialize@NetChain[{GatedRecurrentLayer[3],
      myGL,SequenceLastLayer[], LinearLayer[1]},
      "Input"→{4,1},"Output"→1].
```



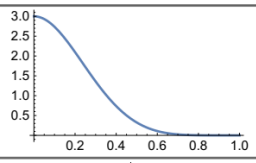
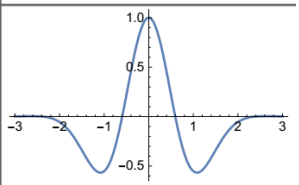
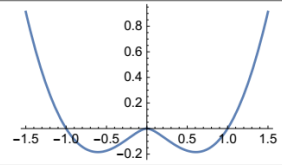
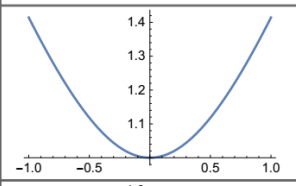
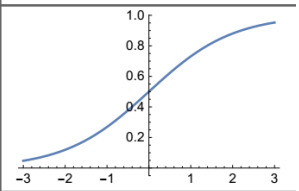
The training can be carried out with stochastic gradient descent using an adaptive learning rate derived from exponentially smoothed average of gradient magnitude method, and using the global loss associated with the L2 norm

of all learned tensors running on GPU,

```
trained = NetTrain[net, trainD, Method→{"RMSProp",
          "L2Regularization"→0.00001}, TimeGoal→30,
          BatchSize→22, TargetDevice→"GPU"]
```

This network was trained with different activation functions, and the errors (RMSE) of the resulted expression of Eq. (4) can be seen in Table 1.

Table 1. Activation functions and results of the corresponding networks

Type	Function	RMSE	Visualization
Weinland	$(1-x)^6\left(35x^2+18x+3\right)$	0.18	
Gauss – Laguerre	$\mathrm{Exp}\left(-x^2\right)\left(1-3x^2+\frac{x^4}{2}\right)$	0.04	
Thin Plate Spline	$\mathrm{If}\left(x\neq 0,\ x^2\,\mathrm{Log}\,(\mathrm{Abs}\,(x)),\ 0\right)$	0.06	
Hardy Spline	$\sqrt{1+x^2}$	0.12	
Sigmoid	$1/(1+\mathrm{Exp}\,(-x))$	0.10	

Further improvement can be achieved employing Long-Short-Term-Memory Layer instead of Gated-Recurrent Layer with 30-50% for the prediction.

7 Forecasting

Now having this model, it can be employed for forecasting the PP. We used one day as the prediction horizon, which means the forecasted value is based on always known (correct) values and not on the previously predicted values. We can see in Fig. 8 that the model gives a good estimation after a short transition time. The network is very simple, therefore it can be trained very fast, so we can use it practically online.

In this way, at the morning when the peak period starts, we can forecast the length of the period employing this starting time, the starting time which was yesterday as well as the lengths of peak period which were two days before. Obviously, this is only one possible combination of the input data, and other characteristic data from the past can be also employed as a different model.

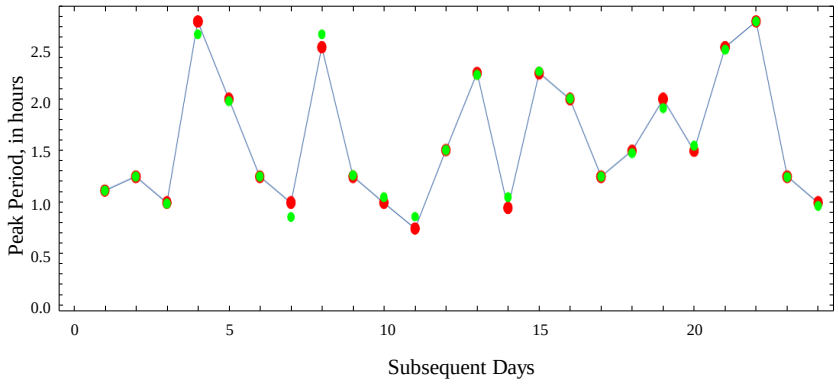


Fig. 8. The forecasting values of the PP by the network are green points in case of Gauss-Laguerre activation function.

8 Discussion and future work

In this study, we developed a simple measure of PP and a method to forecast PP using neural network model. The PP has been defined in various ways in the literature [5,6]. This study defines the starting and ending time of the PP using a local optimization (minimum) method. The advantage of this method is to derive the PP based on the nature of travel time. In the future, we will further develop PP measure in two ways: 1) Instead of using average measure, we may calculate the starting or ending point using inflection point of travel time profile when the slope of a travel time line changes, e.g., from decrease to increase; 2) We may also consider the velocity of traffic on a highway using the typical highway Level-Of-Service (LOS) ratings as a benchmark to

validate the measure [14].

In addition, we systematically compared the results of forecasting the PP using five different activation functions of neural networks and found Gaus-Laguerre function has the smallest prediction error. This is a generalization of a Gaussian and it is globally supported, strictly positive definite oscillatory function. It plays important role in the meshless approximation as RBF basis function [15].

To test these methods further and apply other machine learning techniques in this investigation, considerable more data should be collected. Our aim is to develop a method that can carry out data processing and estimating the parameters of the network online progressively, which means that as the time goes by, more and more data are at our disposal, so the network size can be optimized, and the estimation can be more precise and reliable.

9 Conclusion

We have seen that using this modest recurrent network structure we can get a quite good estimation. However, to get more reliable result, much more measured data would be necessary. In that case training and validation data can be used. The Mathematica Neural Networks package provides symbolic form of the network model; therefore, the resulted code of the network can be easily transported into another software environment.

References

1. Department of Infrastructure, Energy and Energy and Resources 2011. “Congestion in Greater Hobart”, https://www.transport.tas.gov.au/__data/assets/pdf_file/0016/110644/Greater_Hobart_Congestion_full_report.pdf.
2. J. Bolland and D. Ashmore, “Research into traffic peak spreading”, Transfand New Zealand Research Report No. 241. 53 pp.
3. J. Bolland,, and D. Ashmore, “Traffic peak spreading in congested urban environments", 25th Australasian Transport Research Forum, Canberra, Australia Capital Territory, 2-4 October 2002.
4. N. Holyoak and M. C. Yee, “Peak Spreading Behavior and Model Development”, 29th Australasian Transport Research Forum, Gold Coast Queensland, 27-29 September 2006, pp. 1-15.
5. OPUS International Consultants, “Wellington Regional Council State Highways 1 and 2: Peak Spreading”, Contral Laboratories Report 97-529502, 1997.
6. S. Çolak, L. Antonio and M.C. González. "Understanding congested travel in urban areas.", Nature communications 7 (2016): 10793.
7. K.Kumar, M.Parida, O. Pedersen and V.K.Katiyar, “Short term traffic flow prediction for a non-urban highway using Artificial Neural Network ”, Procedia, Social and Behavioral

Sciences,Elsevier 104, 2013 pp.755
8. N.Sohaee,H. G. Kumar and A. Sreeram, “Short-term traffic volume prediction using neural networks”, https://illidanlab.github.io/big_traffic/2018/papers/s haee2018shor tterm.pdf
9. H.Sun, H.X.Liu, H. Xiao, R.R.He and B. Ran, “Short-Term Traffic Forecasting Using the Local Linear Regression Model”, 82nd Annual Meeting of the Transportation Research Board, Washington, DC, TRB 2003 Annual Meeting CD-ROM, pp.1.
10. Y.Zhang and Y. Liu, “Comparison of Parametric and Nonparametric Techniques for Non-peak Traffic Forecasting”, International Journal of Mathematical and Computational Sciences, Vol:3, No:3, 2009, pp.172.
11. J. Xiao, H.Li, X. Wang and S.Yuan, “Traffic Peak Period Detection from an Image Processing View”, Wiley Hindawi, Journal of Advanced Transportation, Volume 2018, Article ID 2097932, pp.1.
12. P.A. Laharotte, R. Billot, E. Come, L. Oukhellou, A. Nantes, N.E.E. Faouzi , “Spatiotemporal Analysis of Bluetooth Data: Application to a Large Urban Network”, IEEE Transactions on Intelligent Transportation Systems, Vol:16, No:3: 2015 pp. 1439-1448.
13. S. Haykin, Neural networks and Learning machines, Prentice Hall, 3rd edition, New York, 2008, p.790.
14. Transportation Research Board, “Manual Highway Capacity, HCM2010”, National Research Council, Washington, DC, 2010.
15. Fasshauer G.E. Meshfree Approximation Methods with MATLAB, Interdisciplinary Mathematical Sciences - Vol.6, World Scientific New Jersey, London (2007), pp.38.

Transfer function modeling of constant work-rate tests in patients with COPD

Joren Buekers[1,2], Hanne Cryns[1], Patrick De Boever[2,3], Emiel F.M. Wouters[4], Martijn A. Spruit[4,5,6], Jan Theunis[2] and Jean-Marie Aerts[1]

[1] Measure, Model & Manage Bioresponses (M3-BIORES), KU Leuven, Belgium
[2] Environmental Risk and Health unit, VITO, Belgium
[3] Centre for Environmental Sciences, Hasselt University, Belgium
[4] Department of Research and Education, CIRO, The Netherlands
[5] REVAL - Rehabilitation Research Center, Hasselt University, Belgium
[6] Department of Respiratory Medicine, Maastricht University Medical Centre, The Netherlands
joren.buekers@kuleuven.be

Abstract. COPD is a highly prevalent lung disease that is characterized by airflow limitation. Exercise endurance of COPD patients can be quantified by means of a constant work-rate test (CWRT), performed on a cycle ergometer. The aim of this study was to develop transfer function models that can accurately describe the dynamic response of oxygen uptake (VO_2) and carbon dioxide output (VCO_2) towards the step-up/step-down of physical activity during/after a CWRT. Time constants (TC) extracted from these models can provide insights into the dynamic response of these variables. Twelve COPD patients performed a CWRT. VO_2 and VCO_2 were measured at every breath and data were divided into an exercise phase and a recovery phase. First order transfer function models could accurately describe the dynamic response of both variables in both phases ($R^2 > 0.97$; YIC < -10). TCs of both VO_2 and VCO_2 were smaller (i.e. faster dynamics) in the exercise phase compared to the recovery phase ($p < 0.05$). TCs of VO_2 during the exercise phase and of both variables in the recovery phase are inversely correlated with the lung function parameter 'forced expiratory volume in the first second' (FEV1, expressed as a % of the predicted value; $p < 0.05$), indicating that the dynamic response of VO_2 and VCO_2 becomes slower with increasing lung function impairment. Hence, these analyses can provide valuable information on how lung diseases impair the dynamic responses of the respiratory system towards changing physical activities.

Keywords: Transfer function, constant work-rate test, COPD, oxygen uptake, carbon dioxide output

1 Introduction

COPD is a highly prevalent lung disease that is characterized by airflow limitation. It is mostly caused by tobacco smoke [1]. Exercise endurance of COPD patients can be quantified by means of a submaximal, constant work-rate test (CWRT), performed on a cycle ergometer [2]. The exercise endurance time is considered the main outcome of

the CWRT. By comparing endurance time before and after pulmonary rehabilitation, the increase in physical condition after pulmonary rehabilitation can be quantified [3]. During CWRT, also a variety of physiological variables (e.g. oxygen uptake (VO_2) and carbon dioxide output (VCO_2)) are continuously measured. However, these physiological time series are rarely investigated in detail. Therefore, the aim of this study was to develop transfer function models that can accurately describe the dynamic response of VO_2 and VCO_2 towards the step-up/step-down of physical activity during/after a CWRT. From these models, time constants (TC) can be extracted that indicate the speed of the dynamic response of VO_2 and VCO_2, which can in turn provide valuable insights in the underlying biological processes.

2 Material and methods

Clinically stable COPD patients performed a CWRT during a standard baseline assessment prior to a pulmonary rehabilitation program. VO_2 and VCO_2 were measured at every breath, resulting in a non-equidistant time series. Different pre-processing approaches were tested and the most satisfying results were obtained when these non-equidistant time series were first resampled to a fixed sampling interval of 5 seconds, after which the general trend was extracted using an integrated random walk smoothing algorithm (*Figure 1*). Hereafter, the data were divided into an exercise phase (step-up of physical activity; *Figure 1*) and a recovery phase (step-down of physical activity).

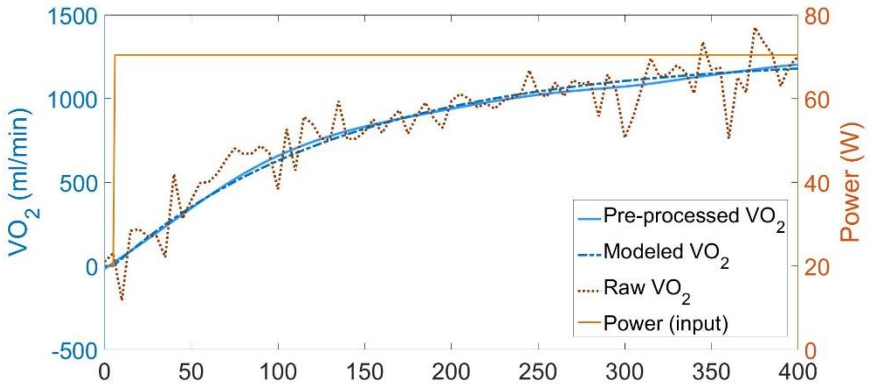


Figure 1: Example of raw, pre-processed and modeled VO_2 data during the exercise phase

After data pre-processing, transfer functions with different model orders were examined for stability and model fit (based on R^2 and Young's Identification Criterion (YIC) [4]). First order transfer functions of the following form had the best results and were used for further analyses:

$$y_k = \frac{b_0}{1 + a_1 z^{-1}} u_k + \xi_k$$

with z^{-1} = backward shift operator, y = output trend (VO_2 or VCO_2), u = input (cycling power), k = sample number, ξ = process and measurement noise, a_1 and b_0 = model parameters.

Lastly, TCs were calculated from these first order models using the following formula:

$$TC = \frac{-5}{\log_e(-a_1)} \, seconds$$

3 Results

Twelve COPD patients (8 males, 4 females) performed a CWRT (age: 63 ± 8 years; BMI: 25 ± 4 kg/m²; % predicted of the forced expiratory volume in the first second (FEV1%pred): 49 ± 14 %). First order transfer function models could accurately describe the dynamic response of both variables in the exercise and recovery phase ($R^2 > 0.97$; YIC < -10). The extracted TCs quantify the speed of the dynamic responses in the exercise phase (VO_2: 133 ± 127 s; VCO_2: 327 ± 422 s) and the recovery phase (VO_2: 800 ± 589 s; VCO_2: 1045 ± 717 s).
In the exercise phase, TCs of VO_2 were smaller than VCO_2 ($p < 0.005$). TCs of both VO_2 and VCO_2 were smaller in the exercise phase compared to the recovery phase ($p < 0.05$). Additionally, FEV1%pred was inversely correlated with TCs of VO_2 in the exercise phase ($p < 0.05$, $R^2 = 0.35$) and with TCs of both variables in the recovery phase (VO_2: $p < 0.05$, $R^2 = 0.62$; VCO2: $p < 0.01$, $R^2 = 0.77$).

4 Discussion and conclusion

First order transfer function models can accurately model the dynamic response of VO_2 and VCO_2 during the exercise and recovery phase of a CWRT. TCs extracted from these models point out differences between the dynamic response of VO_2 and VCO_2, as well as between the dynamic response during exercise and recovery phase. The inverse correlations between the extracted TCs and FEV1%pred indicate that the dynamic response of VO_2 and VCO_2 becomes slower with increasing lung function impairment. Hence, these analyses can provide valuable information on how lung diseases impair the dynamic responses of the respiratory system towards changing physical activities.

References

[1] Global Initiative for Chronic Obstructive Lung Disease (GOLD), "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease." 2016.

[2] A. van't Hul, R. Gosselink, and G. Kwakkel, "Constant-load Cycle Endurance Performance: test-retest reliability and validity in patients with COPD," *J. Cardiopulm. Rehabil.*, vol. 23, no. 2, pp. 143–150, 2003.

[3] W. Cambach, R. V. M. Chadwick-Straver, R. C. Wagenaar, A. R. J. Van Keimpema, and H. C. G. Kemper, "The effects of a community-based pulmonary rehabilitation programme on exercise tolerance and quality of life: A randomized controlled trial," *Eur. Respir. J.*, vol. 10, no. 1, pp. 104–113, 1997.

[4] P. C. Young, *Recursive Estimation and Time-Series Analysis: An Introduction for the Student and Practitioner*, 2nd ed. Springer-Verlag Berlin Heidelberg, 2011.

Adaptive R-peak Detection Using Empirical Mode Decomposition

Christina Kozia, Randa Herzallah, and David Lowe

Department of Mathematics, Aston University,
Birmingham, United Kingdom
{koziac,r.herzallah,d.lowe}@aston.ac.uk

Abstract. Accurate QRS detection plays a pivotal role in the diagnosis of heart diseases and the estimation of heart rate variability and respiration rate. The investigation of R-peak detection is a continuing concern in computer-based ECG analysis because current methods are still inaccurate and miss heart beats. This paper presents a different algorithm to the state-of-the-art Empirical Mode Decomposition based algorithms for R-peak detection. Although our algorithm is based on Empirical Mode Decomposition, it uses an adaptive threshold over a sliding window combined with a gradient-based and refractory period checks to differentiate large Q peaks and reject false R peaks. The performance of the algorithm was tested on multiple databases including the MIT-BIH Arrhythmia database, Preterm Infant Cardio-Respiratory Signals database and the Capnobase dataset, achieving a detection rate over 99%. Our modified approach outperforms other published results using Hilbert or derivative-based methods on common databases.

Keywords: R-peak detection, Empirical Mode Decomposition, Local Signal Energy

1 Introduction

The QRS complex is of focal interest in computer-based ECG signal analysis as it is the most distinguishable feature of the heart signal and embodies valuable information from which the Heart Rate Variability [1] and Respiration Rate [2] can be estimated. Signal contamination from various types of noise and the variability of the QRS morphology make the detection of the latter more complex. Moreover, the complexity of QRS identification relies on the difficulty in differentiating the R peaks from large P or T peaks [3].

Several methods for QRS detection have been proposed, from derivative based methods [3], [4] to neural networks methods [5], [6]. The majority of the algorithms consist of two stages: pre-processing and decision. The signal is pre-processed in order to enhance the QRS complex and eliminate noise and baseline wander and then a set of thresholds is applied in order to identify the real R peaks in the signal. In [3] and [4] the QRS complex is enhanced by differentiating and then integrating the signal, in order to obtain the slope and width information of the QRS complex. The decision rules are based on estimators of signal

or noise level, such as the mean and the median. In [7] and [8] it was proposed to pre-process the signal using the Hilbert transform of its first derivative. The zero-crossings in the derivative are represented as peaks in the Hilbert sequence. The regions of high probability to identify an R peak are located using a threshold based on the Root Mean Square (rms) of the Hilbert sequence. Finally, their method located the real R peaks using a second stage detector based on the heart refractory period (200 milliseconds) [9].

Over the past decade, research in computer-based ECG signal analysis has investigated the use of Empirical Mode Decomposition (EMD) [10] in R peak detection [11], [12], [13]. The EMD method acts as an effective pre-processor which amplifies the QRS complex and decomposes the signal into a set of Intrinsic Mode Functions (IMF). A method of R peak identification by summing the first three IMFs and applying a set of experimentally acquired thresholds was developed in [11]. The reconstruction of the ECG waveform by adding the first three IMFs and then applying a threshold based on 50% of the maximum amplitude was reported in [12] and [13]. The major drawback of [11] and [12] is that they have established an empirical threshold scheme. A serious weakness with the detectors being proposed in [11], [12] and [13] is that the threshold is derived from the full length ECG. Generally, difficulties arise when the signal includes very large R peaks, making the threshold high. This results in the failure to detect lower R peaks. Our method provides a solution to detect these lower R peaks by dividing the signal into segments.

Our proposed method for the detection of the QRS complex overcomes the aforementioned problems in the current state-of-the-art EMD methods by introducing an adaptive threshold which is calculated from the local energy of the reconstructed ECG signal from the EMD. The pre-processing stage of our QRS detector contains a band-pass filter in order to eliminate noise and reduce the number of the initial IMFs. The reconstruction of the signal using the EMD method facilitates the removal of low frequency interference and the absolute value of the reconstructed signal amplifies the QRS complexes. The signal is then divided into segments in order to increase the efficiency of the algorithm. Compared with the existing results on the topic, our study has three distinct features that have not been reported in the literature. Firstly, the proposed detector provides a solution for the detection of small R peaks by establishing a threshold derived from the mean of the rms over a prespecified number of most recent segments. Secondly, the threshold established relies on the local signal energy of each segment. Thirdly, the present research explores the differentiation of R peaks from large Q peaks in the absolute value of the signal, by using the first derivative of the ECG signal.

2 Empirical Mode Decomposition

EMD decomposes the signal, $x(t)$, into a series of narrow-band signals, $c_i(t)$, which are called IMFs, and fulfill special conditions. An oscillatory mode of the signal is an IMF exclusively under the conditions that: first, in the whole dataset,

the number of zero-crossings and the number of extrema are either equal or differ at most by one; and second, at any point, the mean value of the maximum and the minimum envelope is zero. The key advantage of EMD is that it is a data-driven analysis method. In each iteration the algorithm needs to decide if the i-th component, $h_i(t)$, extracted from the data, is an IMF based on the conditions mentioned above. In order to achieve this, the EMD method uses a systematic way which is called the *sifting process* and is described as follows: for a given signal $x(t)$, the extrema points are first identified, followed by approximation of the upper, $\widehat{r}(t)$, and lower, $\underline{r}(t)$, envelopes of the signal through a cubic spline interpolation. The mean is then obtained, and the i-th component, $h_i(t)$, is computed as the difference between the signal and the mean. The *sifting process* has to be repeated as many times as required to reduce the extracted signal to an IMF. For our implementation in order to terminate the EMD algorithm, the number of zero-crossings and the number of extrema are checked on the basis of whether they are equal or differ at most by one. If the final residue, $r_N(t)$, is obtained as a monotonic function, the *sifting process* is stopped, $c_i = h_i$, and the signal, $x(t)$, can be written as follows:

$$x(t) = \sum_{i=1}^{N} c_i(t) + r_N(t) \ , \tag{1}$$

where N is the total number of the extracted IMFs.

3 Proposed R-peak Detection

The proposed algorithm is based on the assumption that the QRS complex of the ECG signal can be enhanced by reconstructing the signal from the first three IMFs of the EMD. This assumption is verified on all of the tested recordings and it is shown in Section 4. However, before applying the EMD, the signal is first processed by a band-pass filter to decrease the computational cost and reduce the number of IMFs. Following the pre-processing stage, the reconstructed signal is divided into a number of segments. Then the envelope of the maxima of each segment is approximated. This is followed by the computation of the local signal energy of each segment and an averaging step for the evaluation of the threshold. Moreover, a few checks were implemented to minimize false positives and negatives including the refractory period and the calculation of the derivative of the ECG signal to discriminate large Q peaks from R peaks in the absolute of the signal. The chosen duration of the segment provides an adequate number of QRS complexes and depends on the sampling frequency. Moreover, for the averaging step the eight most recent segments were used. The number of segments to be averaged is a prespecified parameter which can be decided based on the clinical condition of the patient and whether it is expected that their ECG signal is going to be less or highly variant. However, it is recommended that just the most recent history of the ECG vital signs are kept; thus we used the eight most recent values in the current paper. To summarise, the proposed QRS complex detection algorithm is as follows:

3.1 Pre-processing Stage

1. The raw signal, $x(t)$, is first filtered with a band-pass filter, whose coefficients were designed using a Kaiser-Bessel window [14]. The band-stop frequencies were set at 8 and 20 Hertz [15] in order to amplify the QRS complex, eliminate noise and reduce the number of IMFs. The output of the filter is denoted as $x_f(t)$,
2. The EMD method is applied to $x_f(t)$ to extract the IMFs, $c_1(t) \dots c_N(t)$, where N is the total number of the extracted IMFs,
3. The signal is reconstructed by adding the first three IMFs,

$$x_r(t) = \sum_{i=1}^{3} c_i(t) \ , \tag{2}$$

   where the number of IMFs is experimentally selected and it will be discussed later,
4. Then, the absolute value of the reconstructed signal is computed, that is $a(t) = |x_r(t)|$. This makes all data points positive and implements a linear amplification of the reconstructed signal emphasising the higher frequencies.

3.2 Decision Stage

5. In order to increase the efficiency of the algorithm, we divide $a(t)$ into k segments of 3 seconds duration, that is $k = (\text{total number of samples})/(3 * fs)$. The starting point of the k-th segment should match the last R peak located in the $k-1$ segment in order increase the accuracy,
6. Compute the envelope of the maxima, $\hat{a}_k(t)$, of $a_k(t)$ for each segment k through a cubic spline interpolation of the local maxima,
7. Compute the local signal energy for each segment as,

$$RMS_k = \sqrt{\frac{1}{M} \sum_{t=1}^{M} [\hat{a}_k(t)]^2} \ , \tag{3}$$

   where k is the current segment and M is the number of samples in the segment, that is $M = 3 * fs$,
8. The threshold of the k-th segment is set to be the mean of the most recent eight RMS_k values,

$$T_k = \frac{1}{8} \sum_{j=k-7}^{k} RMS_j \ , \tag{4}$$

9. The peaks, which exceed threshold T_k in the absolute sequence $a_k(t)$, are classified as candidate peaks,
10. In order to segregate large Q peaks from R peaks, we compute the first derivative of $x_r(t)$. Peaks with a negative derivative will be investigated further at the refractory period check given next,
11. Apply a refractory period check when the R-R interval of two adjacent peaks is less than 200 milliseconds. Keep the peak with the maximum amplitude.

4 Results and Discussion

The proposed algorithm is based on the assumption that the range of the frequencies of the first three IMFs of the EMD corresponds to the QRS complex which includes high frequencies in the range 3-40 Hertz [9]. Moreover, P and T wave frequencies are about 0.7-10 Hertz [9], thus in order to enhance QRS complexes, the IMFs that correspond to P and T waves should be discarded. The following discussion shows the validity of this assumption. Fig. 1 shows the filtered ECG signal, $x_f(t)$, of recording 100 from the MIT-BIH Arrhythmia database [16] and its first five IMFs obtained after the EMD algorithm. A Fourier

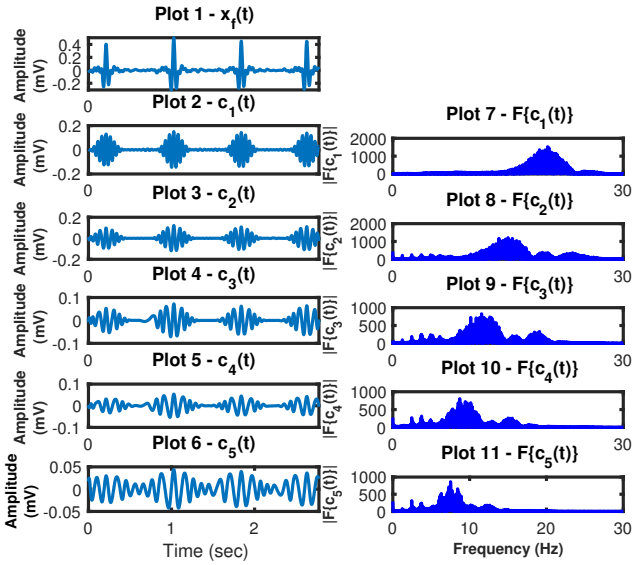


Fig. 1. The result of the EMD and the spectrum of each IMF. Plot 1 corresponds to the filtered ECG, $x_f(t)$. Plots 2 to 6 correspond to the first five IMFs. Plots 7 to 11 correspond to the Fourier transform of each IMF.

transform is applied to each IMF, in order to obtain their frequency bands. It is evident that as the order of the IMFs increases, the frequency content decreases. It can be observed that the spectra of the first three IMFs correspond to the frequency band of the QRS complex. The dominant frequencies in Plots 7-9 (Fig. 1) are about 10-20 Hertz, whereas the dominant frequencies in Plots 10 and 11 are about 2-10 Hertz, which shows that the last two IMFs correspond to P and T waves, hence they should not be used in signal reconstruction. Fig. 2 shows that the filtered signal, $x_f(t)$, can be approximated by the reconstructed signal (summation of the first three IMFs), $x_r(t)$, because the difference of the two signals (red dotted line) is small and the shape of the QRS complex is preserved.

Hence, the first three IMFs are sufficient to delineate the QRS complex. Our

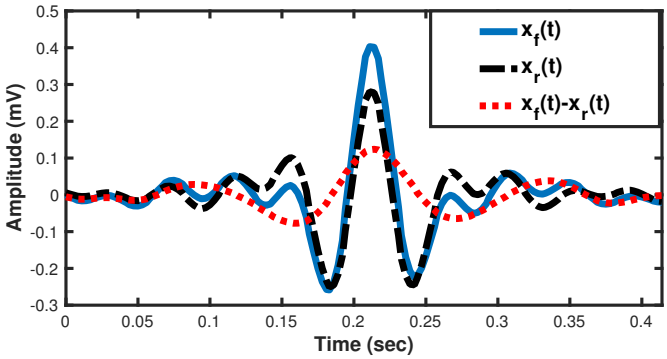


Fig. 2. Reconstruction of the filtered signal, $x_f(t)$, by the summation of the first three IMFs, $x_r(t)$, and their difference, $x_f(t) - x_r(t)$.

assumption was tested on all the recordings under study and the first three IMFs were found to be sufficient for reconstructing the signal, enhance the QRS complex and eliminate low frequency interference. Furthermore, the number of the extracted IMFs for recording MIT-BIH 100 before applying the band-pass filter was 24, and after applying the filter was 22. Following this verification step, our proposed QRS detector was tested using the entire records from the MIT-BIH Arrhythmia database [16] which belong to adults, and all the records from the Preterm Infant Cardio-Respiratory Signals database [16], [17]. Moreover, our method was tested on real data from the Capnobase dataset [18] which were collected during elective surgery and routine anaesthesia and belong to children of ages in ranges 1-14 years old. Furthermore, all databases under study provide annotated R peaks. Fig. 3 shows the sequential steps of the QRS detector. The detected R peaks are marked by an asterisk '*' on the filtered signal, $x_f(t)$ (Plot 4).

For all the databases the results obtained from the proposed method are shown in Table 1, Table 2 and Table 3. Table 4 shows a comparison of our method's performance with other existing methods. A false negative (FN) occurs when the algorithm fails to detect an actual R peak. A false positive (FP) represents a false peak detection. *Sensitivity* (Se), *Positive Predictivity* ($+P$) and *Detection Error Rate* (DER) were calculated using the following formulas respectively:

$$Se(\%) = \frac{TP}{TP + FN}\% \ , \tag{5}$$

$$+P(\%) = \frac{TP}{TP + FP}\% \ , \tag{6}$$

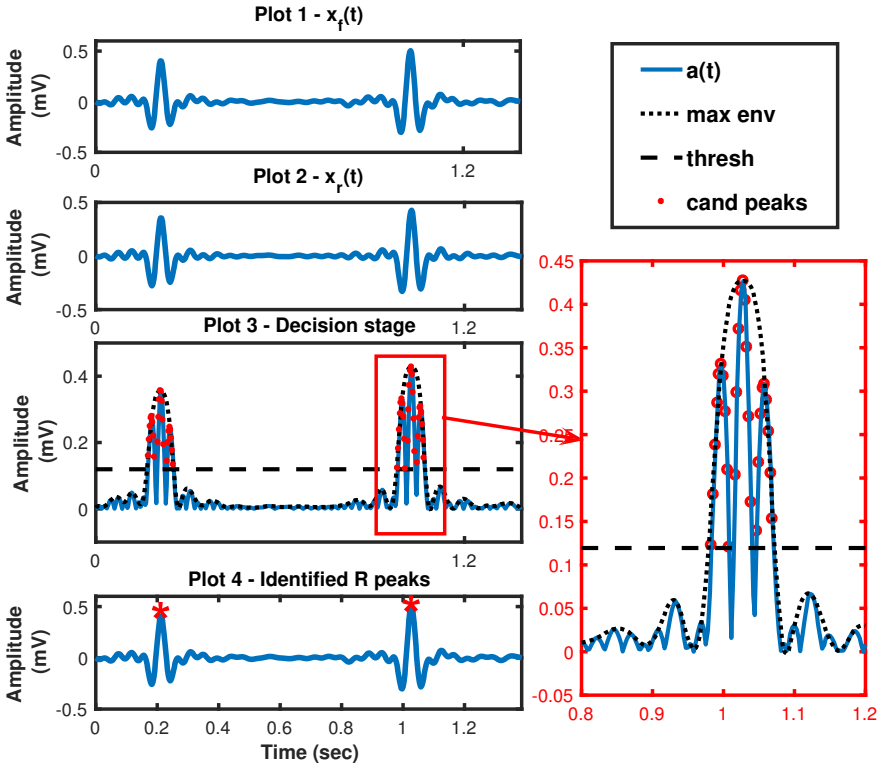


Fig. 3. Steps of the QRS detector for a part of the record 100 from the MIT-BIH database. Plot 1 corresponds to the filtered ECG signal, $x_f(t)$. Plot 2 corresponds to the reconstructed signal, $x_r(t)$. Plot 3 shows the absolute sequence, $a_k(t)$, (blue line) and its maximum envelope, $\hat{a}_k(t)$, (dotted black line) along with the threshold (dashed black horizontal line) and candidate peaks marked with a red asterisk '*'. Plot 4 shows the identified R peaks on $x_f(t)$ as red asterisk '*'.

$$DER(\%) = \frac{FP + FN}{\text{total number of R peaks}}\% \ , \tag{7}$$

where TP (true positives) is the total number of R peaks correctly identified by the detector.

Table 1. QRS detection performance using the MIT-BIH Database

MIT-BIH record	Annotated peaks	DER (%)	*Se* (%)	*+P* (%)
100	2273	0.00	100	100
101	1865	0.48	99.95	99.57
103	2084	0.00	100	100
104	2229	1.57	100	98
105	2572	2.33	99.92	97.79
106	2027	0.98	99.41	99.60
107	2137	0.47	99.81	99.72
109	2532	0.28	99.72	100
111	2124	0.66	99.95	99.39
112	2539	0.20	100	99.80
113	1795	0.11	100	99.89
115	1953	0.00	100	100
117	1535	0.00	100	100
118	2278	0.04	100	99.96
119	1987	0.25	100	99.75
121	1863	0.16	99.95	99.90
122	2476	0.00	100	100
123	1518	0.06	100	99.93
124	1619	0.30	99.81	99.77
Average	35740	0.42	99.92	99.66

As can be seen from Table 4 our method shows a better performance for the MIT-BIH records, achieving higher *Se* of 99.92% compared to 99.86% in [7] and 99.80% in [3] as well as lower DER of 0.42% compared to 14.3% in [7] and 1.33% in [3]. Furthermore, the highest *Se* percentages are reported for PICSDB and Capnobase records by our detector, compared to [3] and [7] (Table 4).

Compared to existing R peak detection methods, the following observations were found. Firstly, during our experiments we observed that some of the adult recordings from the MIT-BIH database include inverted R peaks and this increases the R-peak time-of-occurrence error. The QRS detector fails to detect the inverted R peaks. However, it identifies as a real R peak, a peak close to the inverted one which is not counted as a FP. Hence, the time difference between the actual and the detected peak is large and affects the time-of-occurrence error. The same problem occurs with existing methods [3], [4], [7], [8], [11], [12], [13]. We will further investigate this problem in future research. Secondly, another important observation, that yields high error in the detection of R peaks and was seen in some recordings from the MIT-BIH database, is that the absolute

Table 2. QRS detection performance using the PICSDB

PICSDB record	Annotated peaks	DER (%)	*Se* (%)	*+P* (%)
infant1	4671	0.08	99.95	99.95
infant2	970	1.34	100	98.70
infant3	1757	0.91	100	99.10
infant4	2300	0.00	100	100
infant5	4434	0.04	100	99.95
infant6	3974	0.30	100	99.70
infant7	4451	0.13	100	99.87
infant8	4185	0.02	100	99.98
infant9	4426	0.59	99.50	99.91
infant10	4572	0.19	100	99.80
Average	15371	0.36	99.95	99.70

Table 3. QRS detection performance using the Capnobase Dataset

Capnobase record	Annotated peaks	DER (%)	*Se* (%)	*+P* (%)
9	815	0	100	100
15	960	0	100	100
16	1012	0	100	100
18	1131	0	100	100
23	818	0	100	100
28	588	0	100	100
29	546	0	100	100
31	539	0	100	100
32	685	0	100	100
35	900	0.18	100	99.89
38	956	0	100	100
103	826	0	100	100
104	912	0	100	100
105	530	0.37	100	99.62
Average	12094	0.03	100	99.97

Table 4. Comparison of QRS detector performance with other methods

Database	Method	DER (%)	*Se* (%)	*+P* (%)
MIT-BIH	Derivative based [3]	1.33	99.80	98.85
	Hilbert transform [7]	14.23	99.86	99.71
	Our method	0.42	99.92	99.66
PICSDB	Derivative based [3]	0.34	99.86	99.81
	Hilbert transform [7]	0.14	99.92	99.84
	Our method	0.36	99.95	99.70
Capnobase	Derivative based [3]	0.03	100	99.97
	Hilbert transform [7]	0.04	100	99.95
	Our method	0.03	100	99.97

amplitude of a Q peak is larger than the R peak. This was found to identify the Q peak as a real R peak, because in the decision stage the threshold is applied to the absolute of the reconstructed signal. Fig. 4 shows part of the MIT-BIH 104 record where the absolute amplitude of some of the Q peaks is larger than the R peak. To address this issue the first derivative of the ECG signal is computed. The derivative after an R peak is negative because the signal decreases in time. The derivative after a Q peak is positive as the signal increases in time. Our method was modified and for each candidate peak also the sign of the derivative was checked. Peaks with a negative derivative were investigated further in the decision stage by applying the refractory period check of 200 milliseconds. Hence, the QRS detector proposed in this paper can efficiently distinguish Q from R peaks, whereas existing methods do not propose anything about this issue [12], [13]. Thirdly, current EMD based methods use the average of the mean of all

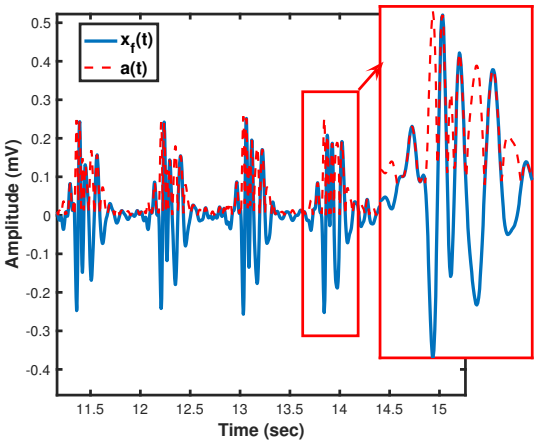


Fig. 4. Part of the filtered ECG signal, $x_f(t)$, and its absolute value, $a(t)$, from the MIT-BIH 104 record. The dashed line in the red box shows that the absolute amplitude of the Q peak exceeds the amplitude of the R peak.

segments [12] and 0.5 of the maximum amplitude [13] to calculate the threshold. When we computed these thresholds for recording MIT-BIH 104, we found the threshold of 0.5 of the maximum amplitude [13] to be high, about 0.5161, while the average of the mean [12] is found to be very low, about 0.0847, thus producing large number of FPs and FNs. Our threshold computed as the average of the rms over the full record was found to be 0.1153, thus minimising FPs and FNs. This is shown in Fig. 5 where as can be seen our peak detector, Fig. 5. Plot 2, has identified all R peaks correctly while the 0.5 of the max amplitude threshold [13], misses too many peaks as can be seen from Fig. 5. Plot 1.

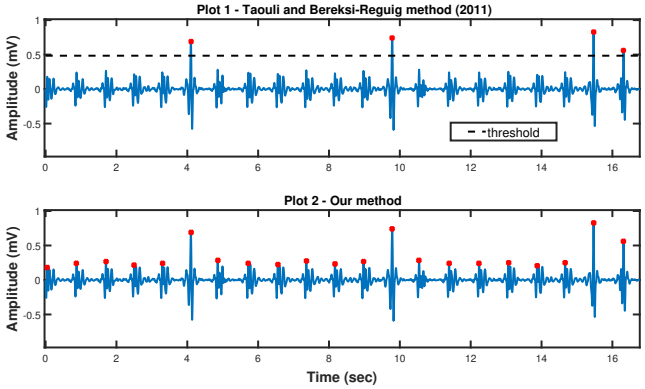


Fig. 5. Plot 1 shows the detected R peaks for a part of MIT-BIH 104 record using the method proposed in [13]. Plot 2 shows the detected R peaks for the same part of ECG using the proposed method.

5 Conclusion

To conclude, a new QRS detector was presented based on Empirical Mode Decomposition using an adaptive threshold which relies on the local signal energy. Our method provides a solution for the detection of small R peaks by establishing a threshold derived from the mean of the rms over eight segments. In addition, our detector correctly discriminates R peaks from large Q peaks in the absolute value of the signal, by using the first derivative of the ECG signal. Using the MIT-BIH Arrhythmia database, the method performed effectively with accurate QRS complex detection of 99.92%, using the Preterm Infant Cardio-Respiratory Signals database and using the Capnobase dataset the method performed effectively with an Se of 99.95% and 100%, respectively. The proposed method shows comparable results with other published methods using derivative-based [3] and Hilbert methods [7] on common databases and real data. However, the average R-peak time-of-occurrence error remains an issue to be addressed, thus the problem of the inverted R peaks is going to be part of our future work.

Acknowledgments. This work is supported by Isansys Lifecare Ltd.

References

1. Moody, G. B.: Spectral analysis of Heart Rate without resampling. Computers in Cardiology, 715-718 (1993)
2. Moody, G. B., Mark, R. G., Zoccola, A., Mantero, S.: Derivation of respiratory signals from multi-lead ECGs. Computers in Cardiology, 113-116 (1985)
3. Pan, J., Tompkins, W. J.: A real-time QRS detection algorithm. IEEE transactions on Biomedical Engineering 3, 230-236 (1985)

4. Hamilton, P. S., Tompkins, W. J.: Quantitative investigation of QRS detection rules using the MIT/BIH Arrhythmia database. IEEE transactions on Biomedical Engineering 12, 1157-1165 (1986)
5. Maglaveras, N., Stamkopoulos, N., Diamantaras, T., Pappas, T., Strintzis, M.: ECG pattern recognition and classification using non-linear transformations and neural networks: A review. International journal of Medical Informatics 52(1), 191-208 (1998)
6. Kohler, B. U., Hennig, C., Orglmeister, R.: The principles of software QRS detection. IEEE Engineering in Medicine and Biology Magazine 21(1), 42-57 (2002)
7. Benitez, D. S., Gaydecki, P. A., Zaidi, A., Fitzpatrick, A. P.: A new QRS detection algorithm based on the Hilbert transform. Computers in Cardiology 27, 379-382 (2000)
8. Benitez, D. S., Gaydecki, P. A., Zaidi, A., Fitzpatrick, A. P.: The use of Hilbert transform in ECG signal analysis. Computers in Cardiology and Medicine 31(5), 399-406 (2001)
9. Malmivuo, J., Plonsey, R.: Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic fields. Oxford University Press (1995)
10. Huang, N. E., et al.: The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 454, 903-995 (1998)
11. Yang, X. L., Tang, J. T.: Hilbert-Huang transform and Wavelet transform for ECG detection. 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM'08, 1-4 (2008)
12. Arafat. M. A., Hasan, M. K.: Automatic detection of ECG wave boundaries using Empirical Mode Decomposition. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 461-466 (2009)
13. Taouli, S., Bereksi-Reguig, F.: Detection of QRS complexes in ECG signals based on Empirical Mode Decomposition. Global Journal of Computer Science and Technology 11(20) (2011)
14. Proakis, J. G., Manolakis, D. G.: Digital Signal Processing: principles, algorithms, and applications (3rd edition). Prentice hall (1996)
15. Elgendi, M., Jonkman, M., DeBoer, F.: Frequency bands effects on QRS detection. 3rd International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSIGNALS) 5, 428-431 (2010)
16. Goldberger, A. L., et al.: Components of a New Research Resource for Complex Physiologic Signals, PhysioBank, PhysioToolkit, and Physionet, American Heart Association Journals. Circulation 101(23), 1-9 (2000)
17. Gee, A. H., Barbieri, R., Paydarfar, D., Indic, P,: Predicting Bradycardia in Preterm Infants Using Point Process Analysis of Heart Rate. IEEE Transactions on Biomedical Engineering 64(9), 2300-2308 (2017)
18. Karlen, W., Raman, S., Ansermino, J. M., Dumont, G. A.: Multiparameter respiratory rate estimation from photoplethysmogram. IEEE Transactions on Bio-medical Engineering 60(7), 1946-1953 (2013)

Understanding the behaviour of energy prices in Brazil

Abdinardo Moreira Barreto de Oliveira[1] ORCID: 0000-0002-9377-6267

Anandadeep Mandal[2]

1. Universidade Tecnológica Federal do Paraná, Departamento de Administração. Via do conhecimento, km 01, Pato Branco, Paraná, Brazil. ZIP: 85.503-390.
abdinardom@utfpr.edu.br

2. University of Birmingham, Department of Finance, Birmingham Business School. Edgbaston, Birmingham, B15 2TY, United Kingdom.
A.Mandal@bham.ac.uk

Abstract. The objective of this paper is twofold: i) to identify the relationship, in a deterministic way, between the useful water volume in the Brazilian reservoirs and the Price for Settlement of Differences (PSD) and; ii) to identify the parametric stochastic process which explains better the behaviour of the PSD series. Although there is a considerable amount of publications dealing with the commercialization of Brazilian electric energy through auctions, or dealing with models for predicting the increasing demand for electricity, little has been discussed about the econometric or stochastic behaviour of these prices, even though it is admitted, in these same publications, that they occur. It was collected 882 weekly observations of the useful water volume (%) and the PSD (R$/MWh). Regarding the first question, it was found that the useful volume of water can explain between 54% and 89% of the variations in the log PSD. Regarding the second question, it was found that a Vasicek (or Ornstein-Uhlenbeck) and a CKLS process, quite close to a Square Root Cox-Ingersoll-Ross (CIR-SR) process, better explain the stochastic behaviour of PSD. Lastly, futures studies are suggested at the end of this paper.

Keywords: Energy prices, Markov Switching model, Stochastic processes.

1 Introduction

In January of 2018, the installed capacity of generation in the Brazilian electrical system was 158,521 MW of which 129,314 MW came from the following renewable energy sources: 63.8% produced by hydroelectric plants, 9.2% produced by biomass plants, 7.86% produced by eolic plants and 0.72% produced by solar plants. On the transmission lines, Brazil has more than 140,000 km of lines at 230 kV and above, in a national interconnected system. In 2017, the average consumption of electric energy was 577,968 GWh of which 28.70% were consumed by industries, 23.17% by homes, 19.73% were due to losses and differences in the electrical system, 15.25% by commercial companies, 8.33% by the Brazilian Government and the power distributors, and 4.83% by rural properties [1].

Since 2004, after the revision of the initial model of the Brazilian Electrical Sector, the commercialization (purchase and sale) of electric power in Brazil occurs in two contracting environments: one called Regulated Contracting Environment (RCE), which purpose is to attend the majority of captive consumers (ex. homes); and another called Free Contracting Environment (FCE), which allows large consumers (ex: industries) to choose directly their electricity suppliers via bilateral contracts [2-5].

Both environments are ruled by the National Agency of Electrical Energy (ANEEL) and managed by the Chamber for Commercialization of Electrical Energy (CCEE). Under specific rules of commercialization defined by ANEEL, the power distributors must purchase electric energy only through public regulated procurement auctions in the RCE, which CCEE conducts all the auctioning process, being the power plants and the trade agents the sellers [2, 4].

These public regulated procurement auctions in Brazil are carried out by a slight version of the Anglo-Dutch model, in a descendent price format, whose first stage is a descendent clock auction (or reverse English auction), followed by a second stage sealed-bid (Dutch auction). The purposes of these auctions are: to make a long-run contract market to steer the expansion of the installed capacity; to reach low tariffs; to prevent collusion and market power behaviour; and to guarantee that the purchase of energy is supplied to consumers in an economical and understandable way [3].

Another role of CCEE is to carry out the procedures required to account for the differences between the contracts aforesaid and the current quantity of energy produced or consumed. These differences are settled in the short-run market through a spot price named Price for Settlement of Differences (PSD) [2]. This price is calculated using computational models (NEWAVE and DECOMP)[1] to find the optimal equilibrium solution between the present benefit of using water and the future benefit of its storage, measured as expected economy of fuels on thermal power plants [5-6]. In other words, the large reservoirs that feature the Brazilian power system are mostly used as energy regulators to minimize the marginal cost of operation [3].

Following a similar reasoning, [7] analyzed theoretically the economics of water as a resource for the generation of electricity in Brazil. On that approach, they stated

[1] http://www.cepel.br/produtos/todos-os-programas-computacionais (In Portuguese).

that water could be treated either as a renewable or nonrenewable good, according to the comparison between the replenish rate, which depends on the natural weather fluctuations and the heterogeneity of the environment (i.e., the speed of hydrologic cycle); and the extraction rate, which depends on the multiples uses and the demand for water. This is relevant to consider because all that affects the hydropower plants performance tend to have an important influence on the principal energy source in Brazil, mainly to the PSD definition aforementioned.

Among several critical theoretical questions that [7] addressed in their study, one important question remains not explicitly answered yet: What is the relationship, in a deterministic way, between the useful water volume in the Brazilian reservoirs and the Price for Settlement of Differences (PSD)?

The importance of answering this question is justified by the need to calculate how much the variations in the useful volume of the reservoirs can explain the variations of the PSD, as well as to calculate the relationship between the replenish rate and the extraction rate of the Brazilian reservoirs, which results can help all policymakers and practitioners interested in this issue to improve their decisions. According to [5], the expectations concerning future electricity consumption and the future flows/affluent natural energy routine in Brazil have a substantial role in the use of accumulated water in the hydroelectric reservoirs and, hence, also on the PSD.

A second approach that can be used to better understand the PSD evolution through time is the stochastic process. In this question, [5] emphasizes that there is, among other results, a strong relationship between stochasticity and the PSD time series, though they have not shown which process was used to simulate those paths.

An example of the application of Stochastic Differential Equations (SDE) on the PSD time series is seen in [8], which they simulated swing options for a hypothetical derivatives market in the FCE, the most conducive market to this, although it is influenced by the prices settled in RCE. However, they only employed an Ornstein-Uhlenbeck (or Vasicek) without jumps process, in just one of twelve existing time series (three markets in four subsystems), not checking if there were other parametric families of stochastic processes that could better explain the paths of the PSD.

In this sense, the following question is then posed: What parametric stochastic process explain better the behaviour of the PSD series? The purpose of answering this question is to verify whether an Ornstein-Uhlenbeck (or Vasicek) stochastic process continues to explain the random paths of the PSD series in a larger sample and in four distinct time series, which results is essential for all practitioners and researchers that focus on pricing derivatives for the energy market in Brazil.

The remainder of this paper is organized as follows: Section 2 describes the methodology and results for the relationship, in a deterministic way, between the useful water volume in the Brazilian reservoirs and the Price for Settlement of Differences (PSD); Section 3 demonstrates the methodology and results of the parametric stochastic process that better explain the behaviour of the PSD series; and finally, Section 4 summarizes the main findings of this study and suggest new issues for futures research.

2 The Relationship Between The Useful Water Volume In The Brazilian Reservoirs And The PSD

2.1 Estimation Models

The first model adopted in this study was to understand the behaviour of useful water volume in the Brazilian reservoirs through time. This useful water volume depends on the South America hydrologic cycle, which it is cyclically affected by meteorological conditions, especially by the El Niño and La Niña phenomenon [9, 10]. In this sense, it was employed a deterministic model with trigonometrical seasonal dependence and linear trend, as proposed by [11,12]:

$$V_t = a_1 + a_2 t + a_3 \sin\left(\frac{2\pi t}{L}\right) + a_4 \cos\left(\frac{2\pi t}{L}\right) + \varepsilon_t \quad (1)$$

$$\varphi = arctan\left(\frac{a_4}{a_3}\right)\left(\frac{180}{\pi}\right) \quad (2)$$

The parameters a_1, a_2, a_3 and a_4 in equation (1) must be estimated so that the curve fits the data appropriately. The parameter L indicates the series cycle (i.e., 365 for daily data, 52 for weekly data and so on). The parameter a_2 indicates the series trend: if positive, the replenish rate is greater than the extraction rate; otherwise, the extraction rate is greater than the replenish rate. Lastly, equation (2) calculates the series phase angle (in degrees) because normally the minimum and maximum mean do not occur on January 1st and July 1st, respectively, for weather-related series.

The second model adopted in this study was to understand the relationship between the useful water volume in the Brazilian reservoirs (V_t) and the Price for Settlement of Differences (PSD):

$$\log PSD_t = \beta_0 - \beta_1 V_t + \varepsilon_t \quad (3)$$

Whose parameters β_0 and β_1 in equation (3) must be estimated so that the curve fits the data appropriately. It was calculated the natural logarithm of PSD to ensure its nonnegativity and linearity, as well as it is expected an inverse relationship with the useful water volume because this asset is, at the same time, a renewable energy source and a scarce resource facing to a growing demand for electric energy in Brazil, subject to the same dynamics of the fossil fuels, therefore [7, 13]. This expected relation is fully consistent with the Theory of Storage and explains why price volatility is an inverse function of inventory level [14].

Both models are using a regression framework, following the same argument presented by [13]: it often proves superiority to machine learning approaches, being simpler to operationalise and more justifiable, in the sense that the effect of each variable is directly and explicitly quantified, even if more complex modelling approaches have been widely explored in the literature and applied by practitioners.

However, several extreme climate events occurred along the first two decades of the 21st century, which caused natural disasters in the centre of South America (intense floods or prolonged droughts), with some of them almost synchronized. That

occurrence might, among other things, compromise energy security, above all in countries that rely heavily on hydroelectric generation, as is the case of Brazil [9,10].

Considering that this situation can cause structural breaks on time series, it was employed in equations (1) and (3) the Markov Switching Model [15], assuming that these "turning points" have an unobservable stochastic behaviour. To perform these calculations, it was employed the MSwM R package [16], considering two or three unobservable regimes, when it is the case, with non-fixed parameters.

2.2 Data

The useful water volume (%) was collected from 01-July-2001 to 18-May-2018 on National Water Agency[2] (ANA) website, totalling 882 weekly observations for 38 reservoirs that compose the Brazilian Electrical System[3], thus distributed: 19 in Southeast/Midwest subsystem; 12 in South subsystem; 04 in Northeast subsystem and 03 in North subsystem. Together, they represent at least 92% of each subsystem.

Then, a single historical series of useful water volume was drawn up to represent each subsystem, considering the proportion of each reservoir in it, as seen in Figure 1 (black line). Moreover, the monthly ONI index was collected from NOAA[4] to find the occurrence of El Niño (light grey rectangle) / La Niña (dark grey rectangle) episodes.

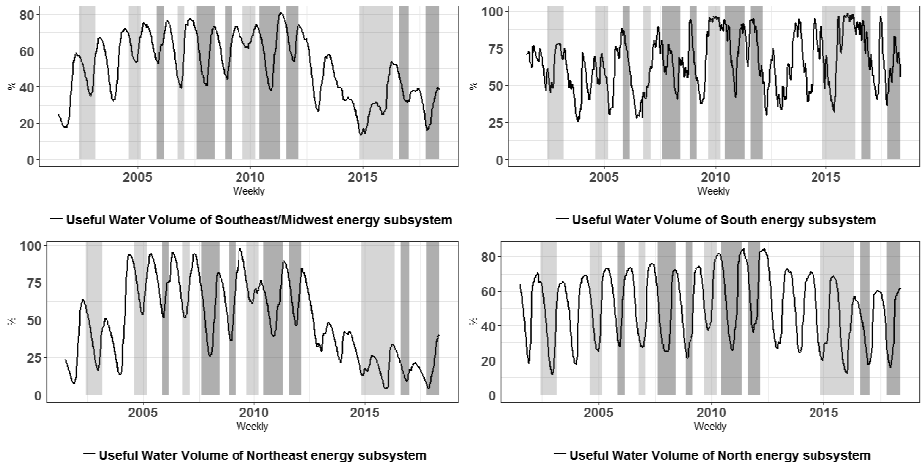


Fig. 1. Weekly useful water volume (%) in reservoirs for each Brazilian Electrical Subsystem.

The PSD (R$/MWh) was collected from 01-July-2001 to 18-May-2018 on [6] website, for the heavy market in each subsystem, as seen in Figure 2. The choice of this market is because it is the time of day where electricity consumption is the highest and spikes may occur, which is a stylized fact of electric energy price series [14].

[2] http://sar.ana.gov.br/MedicaoSin (In Portuguese).
[3] http://ons.org.br/pt/paginas/energia-agora/reservatorios (In Portuguese).
[4] http://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php.

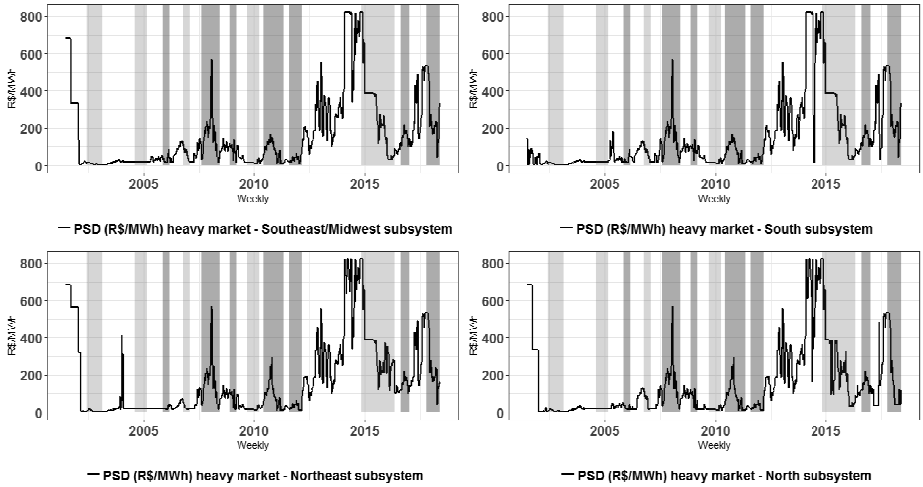


Fig. 2. Weekly PSD(R$/MWh) for each Brazilian Electrical Subsystem.

Figure 2 also shows the occurrence of El Niño (light grey rectangle) / La Niña (dark grey rectangle) episodes. Prices remained in the Brazilian currency (R$) because when converted into US dollars, the later operates as an "inflation rate", which interfered with the nominal behaviour of the original series.

At first glance, the Figures abovementioned suggest that the El Niño/La Niña phenomenon affect the Brazilian electric subsystems in different ways, sometimes acting the contrary to the expected effect. For instance, Figure 1 shows that, between 03-August-2007 and 30-May-2008 (2nd La Niña episode), all the reservoirs were decreasing, when it was expected an increase of them due to the expected rise of precipitation [10]. As an unexpected consequence, it occurred a spike in the PSD in all subsystems, as seen in Figure 2.

However, the period that attracts the most attention is between 24-February-2012, and 7-November-2014, when neither of the two phenomena was recorded. During this period, the reservoirs of the Southeast/Midwest and Northeast Subsystems have decreased considerably[5] (and the PSD reached the highest values in all subsystems, given that they are interconnected), which effects remain to this day: the Southeast/Midwest reservoirs were always below 60% and the Northeast reservoirs were always below 50% of their full capacity. The situation was not even worse because the useful water volumes of the North and South subsystems remained high.

In general, the aforementioned Figures illustrate, to some extent, that extreme climatic affect the useful water volume of Brazilian reservoirs and, thus, the price level of commercialized electricity.

[5] The phenomenon responsible for this was an anomaly occurred in the meridional Sea Surface Temperature (SST) gradient over the tropical Atlantic, which rules the meridional displacement of the Inter Tropical Convergence Zone (ITCZ). The El Niño that occurred after that worsened the situation of these reservoirs [10].

2.3 Markov Switching models results

At second glance, the Figures above said suggest that there are structural breaks in all the time series. Table 1 shows the results of the Markov Switching model applied to the useful water volume (%) of the Brazilian reservoirs (N = 882 obs., first model).

The existence of different regimes in each subsystem is confirmed by the high probability P_{ij} (P_{11}, P_{22} and P_{33}) of remaining in the same regime (all of them above 94%), as well as the high values of adjusted R^2. Each regime in the subsystems has its own features, indicating that the coefficients that predict the useful water volume change through time, which is quite specific of structural breaks.

Regarding trend, the Northeast subsystem is the only when the extraction rate is bigger than the replenish rate, its coefficient is the biggest; the South subsystem is the only which the replenish rate is always bigger than the extraction rate; the Southeast/Midwest subsystem is the only when the extraction rate is bigger than the replenish rate, the phase angle alters, indicating a change in the occurrence of the minimum and maximum values during the drought season.

Table 1. Results of Markov Switching models for useful volume (%) Brazilian reservoirs

VARIABLES		SUBSYSTEMS			
		SE/MW	**S**	**NE**	**N**
Regime 1	Intercept	80.617*	57.387*	31.243*	51.043*
	Trend(t)	-0.054*	0.023*	0.071*	0.010*
	Sin(ωt)	-5.594*	-16.015*	-10.355*	-13.643*
	Cos(ωt)	14.114*	15.593*	15.589*	17.755*
	Phase angle (φ)	-68.378º	-44.235º	-56.406º	-52.462º
	Adj. R^2	0.978	0.699	0.783	0.778
	P_{11}	0.960	0.947	0.991	0.959
Regime 2	Intercept	72.848*	60.856*	99.053*	47.178*
	Trend(t)	-0.053*	0.012*	-0.100*	-0.004*
	Sin(ωt)	1.887*	17.218*	-6.305*	0.724*
	Cos(ωt)	7.995*	-10.392*	11.813*	23.961*
	Phase angle (φ)	76.720º	-31.114º	-61.910º	88.269º
	Adj. R^2	0.728	0.664	0.944	0.948
	P_{22}	0.966	0.968	0.993	0.965
Regime 3	Intercept	41.814*	-	-	-
	Trend(t)	0.055*	-	-	-
	Sin(ωt)	-7.983*	-	-	-
	Cos(ωt)	11.146*	-	-	-
	Phase angle (φ)	-54.390º	-	-	-
	Adj. R^2	0.784	-	-	-
	P_{33}	0.987	-	-	-

Note: *coefficients are, at least, significant at 5%; SE/MW: Southeast/Midwest; S: South; NE: Northeast; N: North; The serial correlation and heteroskedasticity presented in residuals were treated via ARIMA/GARCH models separately, which results can be provided by request.

Table 2 shows the results of the Markov Switching model applied to the relationship between the useful water volume (%) in the Brazilian reservoirs and the log PSD (R$/MWh) for the heavy market (N = 882 obs., second model). Again, the existence of different regimes in each subsystem is confirmed by the high probability P_{ij} (P_{11}, P_{22} and P_{33}) of remaining in the same regime (all of them above 92%).

Regarding R^2, it was found for each subsystem (except for S_1 subsystem) that exists a regime which the useful water volume explains, at least, 54% of the variations of log PSD, according to the theoretical assumptions of the Theory of Storage [14], presenting, hence, an inverse relation. It normally happens when the extraction rate is bigger than the replenish rate, i.e. when the reservoirs levels are diminishing. On the other hand, when the reservoirs levels are increasing, exists a regime which relation is direct, from the microeconomic supply function: the greater the quantity offered the greater the price of the good, although sometimes with a low explanatory power.

Table 2. Results of Markov Switching models for log PSD (R$/MWh), heavy market

VARIABLES		SUBSYSTEMS				
		SE/MW	**S_1**	**S_2**	**NE**	**N**
Regime 1	Intercept	6.766**	6.455**	6.771**	6.658**	6.813**
	Volume (%)	-0.035**	-0.016**	-0.035**	-0.027**	-0.055**
	R^2	0.554	0.162	0,548	0.646	0.646
	P_{11}	0.989	0.984	0,981	0.966	0.958
Regime 2	Intercept	2.437**	4.079**	2.949**	5.629**	1.405**
	Volume (%)	0.008**	-0.014**	0.000	-0.029**	0.055**
	R^2	0.024	0.096	0.000	0.886	0.478
	P_{22}	0.984	0.989	0.978	0.925	0.938
Regime 3	Intercept	-	-	-	1.807**	-
	Volume (%)	-	-	-	0.014**	-
	R^2	-	-	-	0.245	-
	P_{33}	-	-	-	0.958	-

Note: **coefficients are, at least, significant at 1%; SE/MW: Southeast/Midwest; $S_{1,2}$: South; NE: Northeast; N: North; The serial correlation and heteroskedasticity presented in residuals were treated via ARIMA/GARCH models separately, which results can be provided by request.

On S_1, it appears that the useful water volume of the South subsystem does not explain its log PSD. Thus, it was employed the useful water volume of the Southeast/Midwest subsystem instead, for two reasons: it represents 70% of all Brazilian reservoirs system and there is a high Pearson correlation between their log PSD's (r = 0.89). In this sense, the S_2 model was built, which meets all the comments aforesaid to the other subsystems.

Finally, it was found for the Northeast subsystem that there are two regimes that explain the inverse relationship between useful water volume and log PSD, indicating that the coefficients change at certain reservoir level. One explanation for this is that the Northeast region of Brazil is the most affected by extreme climate changes [10], which in turn it requires a more accurate model that can capture the interference of these extreme events in the dynamics between useful water volume and energy prices.

3 The Parametric Stochastic Processes of the PSD series

3.1 Estimation model and data

The CKLS model has come to deal with a bewildering array of previously existing models used to simulate stochastic time series processes, making a common framework which different models could be clustered and their performance compared [17]. Although it has been originally created to capture the stochastic behaviour of short-term interest rate, the CKLS model is widely employed in several finance applications, solving the following stochastic differential equation [18]:

$$dX_t = (\theta_1 + \theta_2 X_t)dt + \theta_3 X_t^{\theta_4} dW_t \quad (4)$$

The parameters θ_1, θ_2, θ_3 and θ_4 must be estimated so that the curve fits the data appropriately. dW_t is a Wiener process. Table 3 presents eight drift-diffusion models that the CKLS can embed, under different parametric specifications. In all cases, $\theta_3 > 0$, because it estimates the standard deviation on the diffusion coefficient [17, 18].

Table 3. The drift-diffusion models set of CKLS stochastic processes

Stochastic Process	θ_1	θ_2	θ_4	Mean reverting	Drift-Diffusion Model
1. Merton	any	0	0	no	$dX_t = \theta_1 dt + \theta_3 dW_t$
2. Vasicek	any	any	0	yes	$dX_t = (\theta_1 + \theta_2 X_t)dt + \theta_3 dW_t$
3.CIR-SR	any	any	1/2	yes	$dX_t = (\theta_1 + \theta_2 X_t)dt + \theta_3 X_t^{1/2} dW_t$
4.Dothan	0	0	1	no	$dX_t = \theta_3 X_t dW_t$
5. GBM or B&S	0	any	1	yes	$dX_t = \theta_2 X_t dt + \theta_3 X_t dW_t$
6. Brennan-Schwartz	any	any	1	yes	$dX_t = (\theta_1 + \theta_2 X_t)dt + \theta_3 X_t dW_t$
7. CIR-VR	0	0	3/2	no	$dX_t = \theta_3 X_t^{3/2} dW_t$
8. CEV	0	any	any	yes	$dX_t = \theta_2 X_t dt + \theta_3 X_t^{\theta_4} dW_t$

Source: [17, 18].

To estimate the model's parameters, it was used the YUIMA R package via quasi-maximum likelihood estimation with standard optimizer BFGS. The choice of Δ_N met the following requirements, in order to get consistent estimators of the parameters along with some conditions on the stationarity and ergodicity of the process X [19]: $N\Delta_N = T \to \infty$; $\Delta_N \to 0$; $N \to \infty$. For log PSD weekly: $\Delta_N = 1/7$; N = 882; T = 126.

No jump processes or change-points were included into the drift-diffusion models, following the references of [8]: the absence of contractual mechanisms for the trading of futures or options in the electrical contracting environment in Brazil would make it inefficient to estimate the parameters of such processes, especially for weekly data, which distorts the characteristic of jumps with very high speed of reversion.

Regarding data, the log PSD (R$/MWh) for the heavy market was separated into an estimation sample (from 01-July-2001 to 31-December-2017; N=862 obs.) and a test sample (from 01-July-2001 to 18-May-2018; N=882 obs.).

3.2 CKLS models results

Table 4 shows the results of the estimators of the CKLS models, which they are strictly positive since the inequation $2\theta_1 > \theta^2_3$ is true for all of them. When comparing these results with the drift-diffusion models presented in Table 3, it is verified that the log PSD for the heavy market in the Southeast/Midwest and in the South subsystems follows a Vasicek (or Ornstein-Uhlenbeck) process, coinciding with the findings from [8], who states that these prices follow a mean-reverting process. Yet, the long-run equilibrium price μ (or the long-run marginal cost of production [20]) increased, from R$/MWh 41.87 [8] to R$/MWh 68.51 and R$/MWh 64.26, respectively.

Table 4. Results of CKLS models for log PSD (R$/MWh), heavy market

VARIABLES	SUBSYSTEMS			
	SE/MW	S	NE	N
θ_1	0.100**	0.223**	0.112**	0.135**
θ_2	-0.024**	-0.054**	-0.027**	-0.033**
θ_3	0.290**	0.444**	0.176**	0.191**
θ_4	0.000	0.000	0.457**	0.458**
μ	4.227	4.163	4.214	4.098

Note: **coefficients are, at least, significant at 1%; SE/MW: Southeast/Midwest; S: South; NE: Northeast; N: North; $\mu = -\theta_1/\theta_2$, is the log long-run equilibrium price.

However, the log PSD for the heavy market in the Northeast and in the North subsystems follow a CKLS process, quite close to a Square Root Cox-Ingersoll-Ross (CIR-SR) process, and following a mean-reverting process. Unlike the Vasicek process, in which the conditional volatility of changes in log PSD is constant, these models imply that the conditional volatility of changes in log PSD is proportional to log PSD [17]. For these subsystems, the long-run equilibrium price μ is R$/MWh 67.63 and R$/MWh 60.22, in that order.

Lastly, because $\theta_4 < 1$, the volatility and electricity price in the Northeast and North subsystems are inversely related, which means the presence of the leverage effect, whose models are capturing a negative volatility skew. This fact is typical of the stock market and opposite to that expected to the commodities market [14].

4 Conclusion

The objective of this paper was to bring a greater understanding of the behaviour of electric energy prices marketed in Brazil. Although there is a considerable amount of publications dealing with the commercialization of Brazilian electric energy through auctions, or dealing with models for predicting the increasing demand for electricity, little has been discussed about the econometric or stochastic behaviour of these prices, even though it is admitted, in these same publications, that they occur.

Regarding the first question posed – *What is the relationship, in a deterministic way, between the useful water volume in the Brazilian reservoirs and the Price for Settlement of Differences (PSD)?* – It was found that the useful volume of water (%)

can explain between 54% and 89% of the variations in the log PSD, depending on the chosen subsystem. However, it was confirmed the occurrence of structural breaks in both time series, whose change of regime can be attributed to the extreme climatic events that have been occurring in the country over the last 17 years.

Regarding the second question posed – *What parametric stochastic processes explain better the behaviour of the PSD series?* – It was found that a Vasicek (or Ornstein-Uhlenbeck) process explains the log PSD stochastic behaviour in the Southeast/Midwest and South heavy markets, whereas a CKLS process, quite close to a Square Root Cox-Ingersoll-Ross (CIR-SR) process, explains the log PSD stochastic behaviour in the Northeast and North heavy markets. In comparison with the results of [8], the long-run equilibrium price increased over time, possibly due to the energy prices regimes changing presented in section 2.3.

As the results above mentioned indicate considerable heterogeneity among the subsystems of electricity generation and distribution in Brazil, it is of fundamental importance that their managers have plans dedicated to dealing with the intrinsic contingencies of each of them, under penalty of possible negative externalities in one of the subsystems to spread to others.

For future studies, it is proposed two suggestions: i) to include in the equation (3) fuel prices and measure their impact on the PSD, since 18.42% of Brazilian electric energy is still produced by thermoelectric plants [1]; ii) the addition of jumps and/or change points in the stochastic models analyzed here, in order to verify their impact on the behaviour of electric energy prices marketed in Brazil, relevant for modelling options contracts.

References

1. Ministry of Mines and Energy [Ministério de Minas e Energia]. (2018). **Monthly report monitoring of the Brazilian electrical system** (Boletim mensal de monitoramento do sistema elétrico brasileiro), January. Available: http://www.mme.gov.br/. (In Portuguese).
2. SOUZA, F. C.; LEGEY, L. F. L. (2010). Dynamics of risk management tools and auctions in the second phase of the Brazilian Electricity Market reform. **Energy Policy**, v.38, n.4, p.1715-1733, April. DOI: https://doi.org/10.1016/j.enpol.2009.11.042.
3. REGO, E. E.; PARENTE, V. (2013). Brazilian experience in electricity auctions: comparing outcomes from new and old energy auctions as well as the application of the hybrid Anglo-Dutch design. **Energy Policy**, v.55, p.511-520, April. DOI: https://doi.org/10.1016/j.enpol.2012.12.042.
4. REGO, E. E. (2013). Reverse price: lessons learned from Brazilian electricity procurement auctions. **Energy Policy**, v.60, p.217-223, September. DOI: https://doi.org/10.1016/j.enpol.2013.05.007.
5. FERREIRA, P. G. C.; OLIVEIRA, F. L. C.; SOUZA, R. C (2015). The stochastic effects on the Brazilian electrical sector. **Energy Economics**, v.49, p.328-335, May. DOI: https://doi.org/10.1016/j.eneco.2015.03.004.
6. Chamber for Commercialization of Electrical Energy [Câmara de Comercialização de Energia Elétrica]. (2018). **Prices**. Available: https://www.ccee.org.br . (In Portuguese).

7. SILVA, A. S.; SOUZA, F. M. C. (2008). The economics of water resources for the generation of electricity and other uses. **Annals of Operations Research**, v.164, n.1, p.44-61, November. DOI: https://doi.org/10.1007/s10479-008-0360-z.
8. SIMÕES, M. D. P; OLIVEIRA, D. L.; PINTO, A. C. F.; KLOTZLE, M. C.; GOMES, L. L. (2011). Swing options for the Brazilian electric energy market (Opções de swing no mercado de energia elétrica brasileiro). **Revista de Economia e Administração**, v.10, n.4, p.591-610. (In Portuguese).
9. MAKHOLM, J. D. (2016). El Niño's uneven disruption of world's electricity systems. **Natural Gas & Electricity**, v.32, n.9, p.29-32. DOI: https://doi.org/10.1002/gas.21900.
10. MARENGO, J. A.; ALVES, L. M.; ALVALA, R. C. S.; CUNHA, A. P.; BRITTO, S.; MORAES, O. L. L. (2017). Climatic characteristic of the 2010-2016 drought in the semiarid Northeast Brazil region. **Annals of the Brazilian Academy of Science** (Anais da Academia Brasileira de Ciências), p.1-13. DOI: http://dx.doi.org/10.1590/0001-3765201720170206.
11. ALATON, P.; DJEHICHE, B.; STILLBERGER, D. (2002). On modelling and pricing weather derivatives. **Applied Mathematical Finance**, v.9, n.1, p.1-20. DOI: http://dx.doi.org/10.1080/13504860210132897.
12. BENTH, F. E.; BENTH, J. S. (2007). The volatility of temperature and pricing of weather derivatives. **Quantitative Finance**, v.7, n.5, p.553-561, October. DOI: https://doi.org/10.1080/14697680601155334.
13. TROTTER, I. M.; BOLKESJ∅, T. F.; FÉRES, J. G.; HOLLANDA, L. (2016). Climate change and electricity demand in Brazil: a stochastic approach. **Energy**, v.102, n.1, p.596-604. DOI: https://doi.org/10.1016/j.energy.2016.02.120.
14. GEMAN, H. (2005). **Commodities and commodity derivatives:** modeling and pricing for agricultural, metals and energy. West Sussex: John Wiley & Sons, Ltd.
15. HAMILTON, J. D. (1994). **Time series analysis**. New Jersey: Princeton University Press.
16. SANCHEZ-ESPIGARES, J. A.; LOPEZ-MORENO, A. (2014). **MSwM:** fitting Markov Switching models. R package version 1.2. Available: https://CRAN.R-project.org/package=MSwM.
17. CHAN, K. C., KAROLYI, G. A., LONGSTAFF, F. A., SANDERS, A. B. (1992). An empirical investigation of alternative models of the short-term interest rate. **The Journal of Finance**, v.47, n.3, p.1209–1227, July. DOI: https://doi.org/10.1111/j.1540-6261.1992.tb04011.x.
18. IACUS, S. M. (2011). **Option pricing and estimation of financial models with R**. West Sussex: John Wiley & Sons.
19. BROUSTE, A.; FUKASAWA, M.; HINO, H.; IACUS, S. M.; KAMATANI, K.; KOIKE, Y.; MASUDA, H.; NOMURA, R.; OGIHARA, T.; SHIMUZU, Y.; UCHIDA, M.; YOSHIDA, N. (2014). The YUIMA project: a computational framework for simulation and inference of stochastic differential equations. **Journal of Statistical Software**, v.57, n.4, p.1-51. DOI: http://dx.doi.org/10.18637/jss.v057.i04.
20. DIXIT, A. K.; PINDYCK, R. S. (1994). **Investment under uncertainty**. Princeton: Princeton University Press.

Time series Analysis for Re-Commissioning of Building Service installations

Albert-Jan Huls[1], Ben Lops[1], Wim Zeiler[2]

[1]Kropman Building Services, Nijmegen, Netherlands
[2]TU Eindhoven, Faculty of the Built Environment, Eindhoven, Netherlands
w.zeiler@tue.nl

Abstract. Heating, Ventilation and Air conditioning systems in new buildings are systematically underperforming, also known as the performance gap. Continuous Commissioning (CCx) is seen as promising to improve this situation using the data time series to monitor the conditions and performance of building Services installations. However, implementation of CCx is complex and cost-intensive when all the available data needs to be analyzed. Alternatively performance efficiencies during exploitation are periodically upgraded by performing Re-Commissioning (RCx). However, often results after RCx do not live up to their expectations and it is necessary to increase the effectiveness of data analysis. The use of the Pareto analysis and the Lean Energy Analysis (L.E.A.) makes this possible. The Pareto analysis is a stepwise approach for the identification of major causes in an identified deviation in a time series. L.E.A. specifically aims at identification and assessment of correlations between primary goals of energy consumption and the actual consumed energy by regression coefficients. By analysing time series in this way, building services inefficiencies are systematically addressed thus improving the effectiveness of the RCx process.

Keywords: Big Data · Time series · Re-Commissioning · Lean Energy Analysis · Pareto analysis

1 Introduction

Energy consumption in buildings accounts for up to 40% of total energy consumption. Over the past decades, the building industry has come aware of a recurring mismatch between design based predicted- and in-use energy consumption of buildings, often referred to as the 'energy performance gap'. Evidence on the magnitude of the gap is adding up fast, suggesting buildings tend to use 1.5 to 2.5 times more energy than predicted in their design phase (Meneres et al 2012, CarbonBuss 2018). Causes for this gap arise in all different stages of the building process, from poor assumptions and model inadequacy in the design stage to deviant occupant behaviour in the operational stage (de Wilde 2014). The amount of devices in current buildings able to capture data is massive as the connections of human activities with the internet. This new paradigm has in many cases considered as the era of the Internet of Things (IoT), or the era of the

Big Data (BD). With respect to buildings however, it seems like the major enemy is uncertainty and variability of data. The well-known issue of the performance gap is making building designs that on paper seem to be low-energy and resilient underperform when built. Why this happen remain uncertain, but the main reasons point to uncertainty on the quality of workmanship related to design-build-operate-maintain, weather conditions, and especially occupants' behaviour. The building industry is starting to realize the parallelism between the issue of low-energy robust buildings and the potential of current developments in data science. Potential solutions for better performing buildings might use technics similar to those found by data scientists that are currently making substantial progress on big data from smart infrastructure, social media or smart home applications. Initiatives started on data analysis for the design of better buildings, the management of buildings, or the understanding of occupant behaviour. In June 2018 there was the 1st international Conference on Data for Low Energy Buildings, held in Murcia, Spain. This is opening new opportunities that shown the next steps to the creation of a new approach in building industry, that could result on large added value for the clients and occupants.
In a literature review of Borgstein et al. (2016), discusses most common methods for evaluation of energy performances. In this review, 4 main methods (umbrella terms) for energy performance evaluation are categorized and discussed. The methods, as discussed in this review:

- Engineering calculation. This method contains relatively simple calculation tools, which do not replicate dynamic processes of full simulation methodologies, which means that these calculations have are lower lever of accuracy. These calculations are suitable for quick energy scans and rough performance estimations.
- Dynamic building energy performance simulation. This method involves the use of computer models to simulate the performances of a building in determined conditions. This method can be very accurate for energy evaluation, but is highly dependent of the modeler knowledge/skills, model uncertainties and detailed model input (e.g. occupant behavior).
- Statistical methods. Different evaluation methods do exist for statistical evaluation of energy performance (gaps). In table 1, taken from Borgstein, et al. (2016), different statistical methods in building energy performance evaluation/assessment, are summarized. This review mentions that the principal limitation of statistical models would seem to be their lack of a relation to physical phenomena, meaning that it can be difficult to interpret the results and identify errors.
- Machine learning. Machine learning is in full development and consists mainly of algorithms, learning from patterns in big data sets. For example, by artificial neural networks or clustering techniques, patterns in large data sets can be identified or clustered. Although this techniques seem to be very promising, further development is needed. Currently, one of the main drawbacks of machine learning it is difficult to show real, physical interpretations for the used 'black-box' approach.

Table 1. Summary of principal statistical methods applied to buildings modified from Borgstein, et al. (2016).

Statistical methods	Applications	References
Simple and multivariate linear regression	Simple models for building performance based on a few characteristics.	ASHRAE (2014)
Change-point regression	Model the non-linear effects of external conditions, e.g. below a certain external temperature, heating systems are switched on.	Zhang, et al. (2014), ASHRAE (2014)
Gaussian process and Gaussian Mixture regression	Prediction of dynamic performance, with an understanding of uncertainty. Flexible models, but more complex.	Zhang, et al. (2014)
Stochastic Frontier Analysis	Effective when there are a large number of efficient buildings and a few that inefficient. Outliers may make the method ineffectual, as residuals will be large.	Chung (2011), Buck, et al. (2007)
TOPSIS	Can be used to develop effective benchmarks, based on regressions.	Lin, et al. (2011), Wang (2015)
Data Envelopment Analysis	Evaluates the technical efficiency and improvement potential of buildings. Can only be applied to buildings within the original dataset.	Chung (2011), Lee, et al. (2011)
Correction factors	Relate building performance to physical parameters, useful for benchmarking.	Bloomfield, et al. (2010)

Although there are studies performed on causes of energy gaps and ways to avoid these causes (Wilde (2014), Menezes, et al. (2012), Morant (2012), Carbon Trust (2011)), a structured analysis, which leads to the explanation of a specific energy gap is still missing. In optimization and system recommissioning studies (Oregon Department of Energy 2014) high energy consumption situations were detected followed by a quite labor-intensive check of the building services systems. A wide variety of errors/defects was found from inefficient control strategies to leaking valves. However, there is still a lack of a structured approach for analysis of the causes of an identified energy gap. In this study the Pareto analysis (Bartlett, 2015) and LEAN Energy Analysis (L.E.A.) (Kissock et al. 2011; Donnelly et al. 2013) were used to performed such an analysis on the time series of the building services installations. Both methods will be explained briefly in the methodology section followed by the results of one of the case studies. This article is a summary of the work by Huls (2016) and is further based on Schoenmaker (2016) for Pareto analysis and Vink (2018) for L.E.A..

2 Methodology

The Pareto analysis, as discussed by Bartlett (2015), is a systematic, stepwise approach for the identification of major causes of problems. This approach is based on the principle that a large part of problems (e.g. 80%) is caused by only a small part (e.g. 20%) of all responsible causes (major causes). The challenge is to identify this '20%' that is mainly responsible for an energy performance gap. The hypothesis that this Pareto analysis might be useful as a systematic step approach to identify the energy gap is based on a study of Schoenmakers (2015).
L.E.A. is discussed and recommended (e.g. Kissock, et al. 2004, 2011, Donnelly et al. 2013) as a useful method to recognize energy inefficiency and reduce energy waste.

L.E.A. is based on the general LEAN principle that all energy which is not used for the primary goal of the energy is waste. It aims to identify the primary goals of the major energy flows in a building. The correlation between these primary goals and the energy consumption provides information about the efficiency of the energy consumption. The L.E.A. adds value to other regression/trend analysis methods by isolating regression coefficients and recognizing that they have physical meaning (Sever et al. 2011, Donnelly et al. 2013). In this way, causes were derived from the physical meaning of the regression coefficients.

Experiments

The practical applicability of the Pareto analysis and L.E.A. in the built environment was only very limited discussed. Therefore, the usefulness of these methods, as systematic analyzes of energy performance gaps in the existing built environment, was tested on 2 representative office buildings as case studies (Huls 2016, Vink 2016). The one presented here is about a traditional (1993) office building with normal building systems and control. Both methods were used in this study, to analyze the strengths, weaknesses and applicability to reinforce each other to a useful, combined method. The Pareto analysis was applied before the LEAN analysis, because it was expected that the Pareto analysis helps to focus on the key aspects. The combined, stepwise analysis in this study is shown below:

Pareto analysis, step 1: Identification of problems

The case study building was modeled according to the design of the building and the 'reference' energy performance simulated to compare with the measured energy consumption and to identify the energy gap. Two different climate files were used for the building simulations. For an indication of the 'average' energy consumption of the building, the reference climate file of NEN 5060 was used for the sensitivity analysis of the energy performances and the selection of critical parameters. This because it shows a more representative, 'average' sensitivity over a year. The second climate file contains measured climate data from the nearest official weather station of KNMI (Dutch climate institute) for a fair comparison between simulated and measured energy consumption, Fig. 1. shows the comparison for 2014.

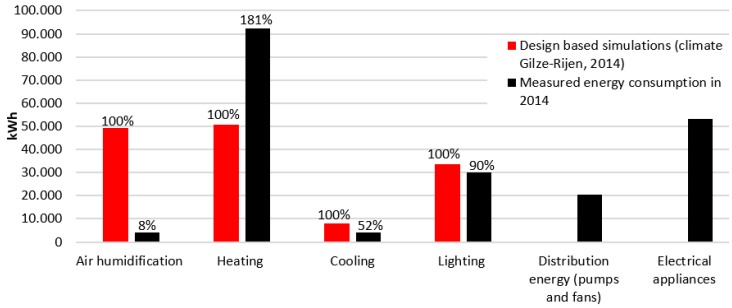


Figure 1: Annual energy consumption case study building: simulated – measured.

Pareto analysis, step 2: Identify root causes of each problem

This step deals with the identification of the major parameters in the energy performance topic. There are a lot of settings, control strategies and properties in buildings and building systems, which all together determine the behavior of the energy consumption. All these factors, see Fig. 2 are reduced to a small amount of (merged) 'main parameters'. In the further steps (3-6), the impact of these parameters is analyzed, as well as the hypothesis if 80% of an energy performance gap can be explained by these parameters. For a widely applicable method, it would be more useful to identify a few parameters for the assessment of the energy performances. Therefore, from this large quantity of parameters (indicated in Fig. 2), a relatively small selection of 15 (merged) 'main parameters' has been made, based on the input of the building model.

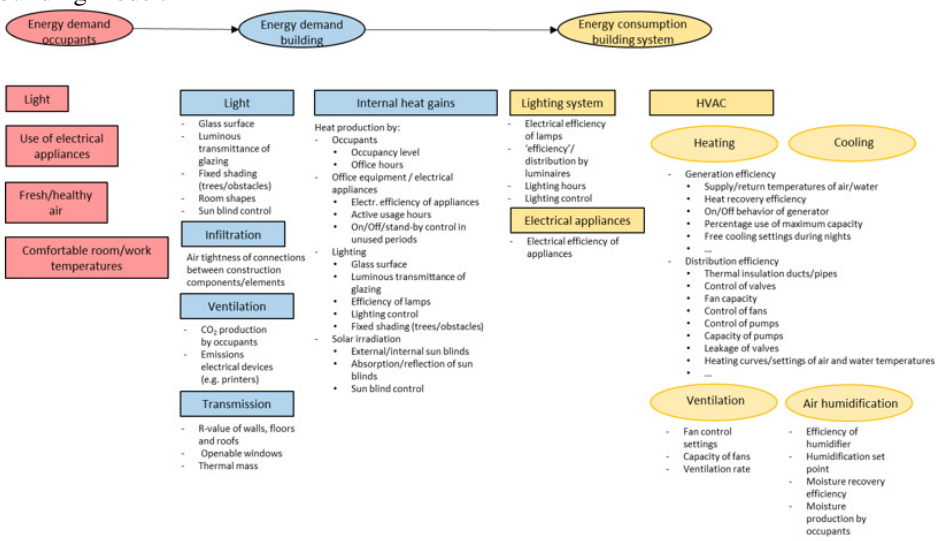


Figure 2. Indication of involved parameters in building energy consumption.

Pareto analysis, step 3/4/5: Rank, score and group problems & causes

The impact of the parameters on the annual energy consumption was assessed by a sensitivity analysis. In this way it was determined which parameters are most important in the assessment of energy performance gaps. The sensitivity analysis was performed, see Fig. 3, by varying the involved parameters individually with + and – 10% of the 'baseline' design settings. In addition to this uniform sensitivity analysis, an additional larger variety was tested for the occupancy rate (+ and – 50%), to check a more realistic variation of this parameter. In Fig. 4 the results of the sensitivity analysis are ranked by the impact on annual energy consumption, for heating, cooling and air humidification. The maximum value on the X-axis in each figure (100%) is the simulated annual energy consumption (in the reference NEN 5060 climate, because the sensitivity analysis is performed for this 'average' climate). In this way, it can easily be derived which parameters are important and should be taken into account in the assessment of the identified energy performance gaps.

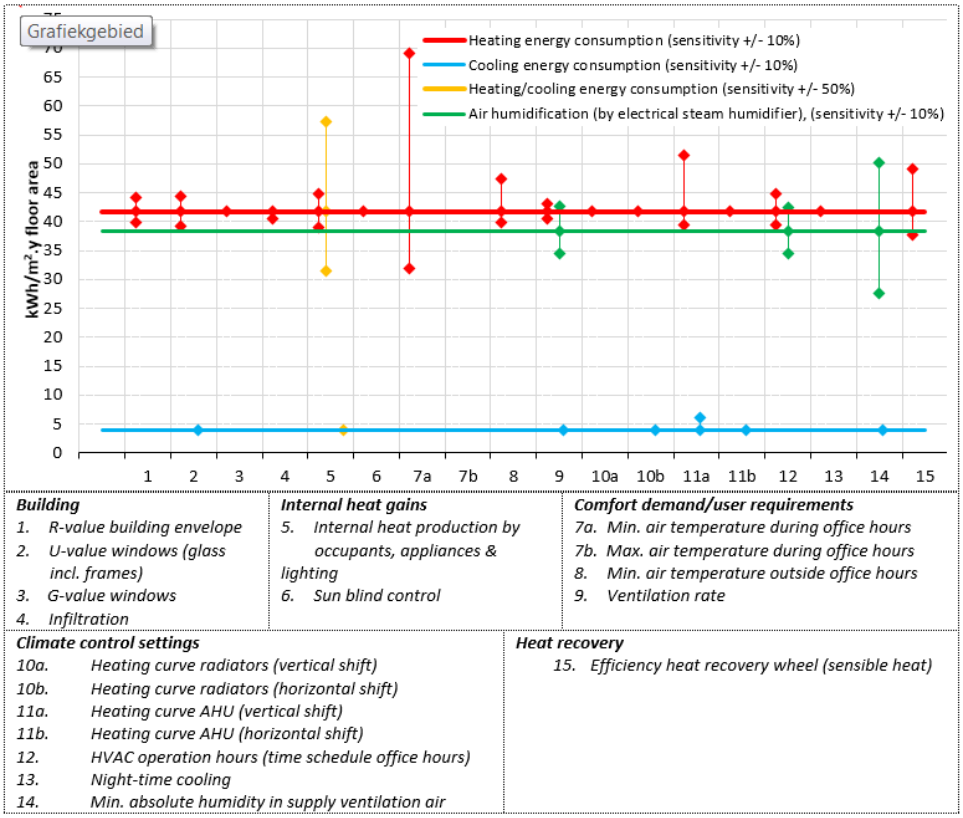


Figure 3: Results of sensitivity test of the annual energy consumption in the case study building.

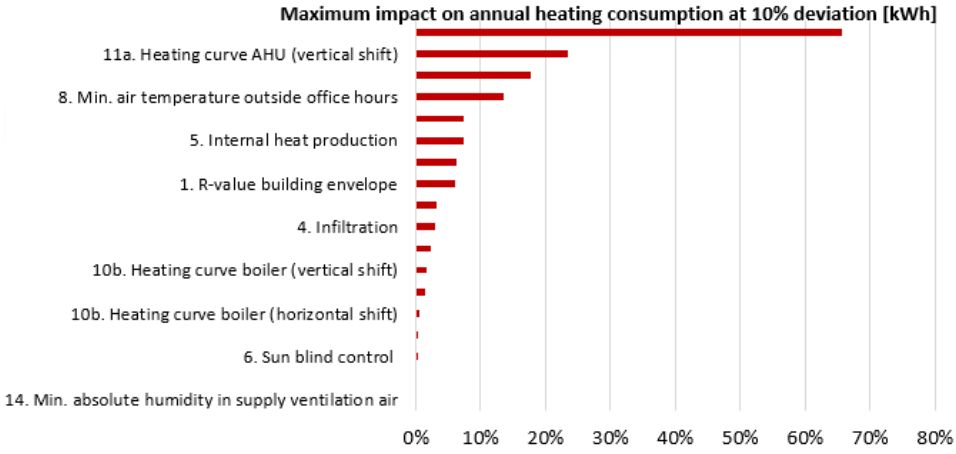
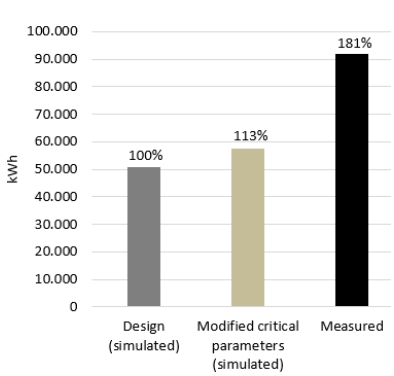


Figure 4: Results of the sensitivity analysis are ranked by the impact on annual energy consumption, for heating, cooling and air humidification.

Pareto analysis, step 6: **Assessment of energy gap by selected critical parameters**
In this step, it is investigated if the energy performance gaps in the case studies can be assessed/explained by the selected parameters from the steps before. For heating energy, the selected most critical parameters are assessed and the impact on the simulated energy consumption is investigated. The impact of the parameters is shown in Fig. 5: over a year they have only a relatively small impact on the energy consumption and only a small part of the energy gap can be explained by the selected parameters. With L.E.A. the remaining energy gap is further analyzed.

L.E.A., step 1: Collect weather & utility data
Collect the required measurement data for a straight-line comparison with the simulated energy consumption. According to the LEAN principle, the way to create a benchmark model is to consider the used energy consumption as a function of the primary goal of this energy. The primary function of room heating and cooling is to bridge the gap between the outdoor temperature and the desired indoor temperature. According to Kissock et al. (2004) building energy consumption can be adequately described by models relating energy consumption to outside air temperature. Therefore, the correlation between the energy and the outdoor temperature was analyzed, see Fig. 6. However, the correlation between the analyzed energy consumption and the outdoor temperature is weak due to several involved time-dependent parameters (as solar irradiance, internal heat production).



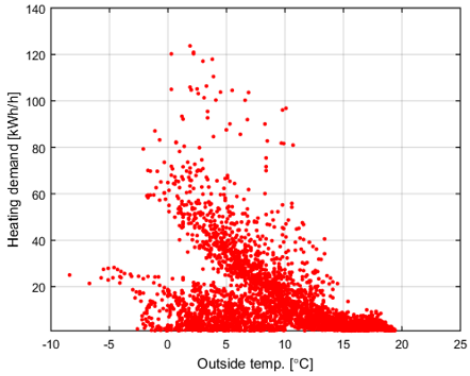


Figure 5: The effect of selected most critical parameters on annual heating consumption and Figure 6: Simulated annual heating consumption as function of outdoor air temperature.

L.E.A., step 2: Create baseline/benchmark models
This step consists of identification of characteristic correlations in energy performances of the case study building and the creation of benchmark models, which can be used to assess measured energy efficiency. Consideration of characteristic, independent variables are an effective way to assess energy efficiency. It was found that energy profiles with hourly data do not well represent the dynamic behaviour of the building systems. The heating demand was summed over a day within each time block leading to 365 data points a year, instead of 8760.

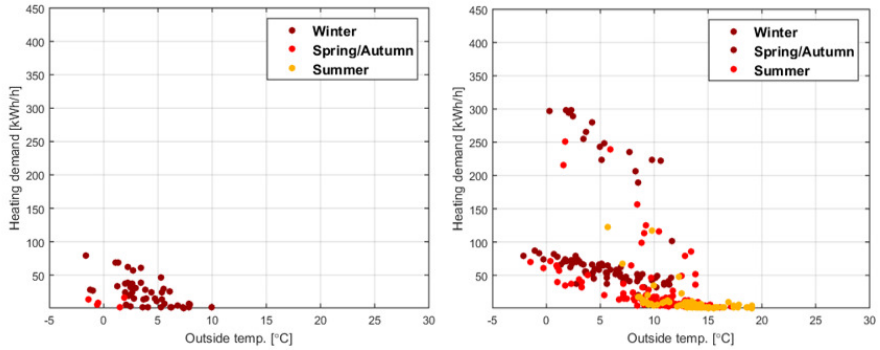


Figure 7: Simulated annual heating demand in time block A (0:00-7:00 h (Tuesday – Friday) and time block B 5:00-8:00 h (Mondays) and 7:00-8:00 h (Tuesday-Friday)

LEAN Energy Analysis, step 3: Identify energy gaps with regression coefficients of benchmark regression models

This step consists of the identification of the energy gap, according to L.E.A.. This way consists of assessment with coefficients of multi-parameterized regression models. Although multi-linear regression analysis can be applied for a better representation of the physical meaning of the data profiles, it is found that the assessment with characteristic coefficients of regression lines (or other 'best fit' curves) does not improve the ability to assess the energy consumption. The regression coefficients (e.g. slope of the regression line), do contain only general information about efficiency of the system. Since in regression lines physical information about the behavior of the building (system) is lost, compared to the energy data profiles, physical parameters cannot be derived from the regression coefficients. Involved parameters do not have that specific impact on regression lines. The combination of all parameters determine the regression coefficients, resulting in the fact that no reliable conclusions can be drawn from a difference between coefficients in the 'design' regression lines and the 'measured' regression lines.

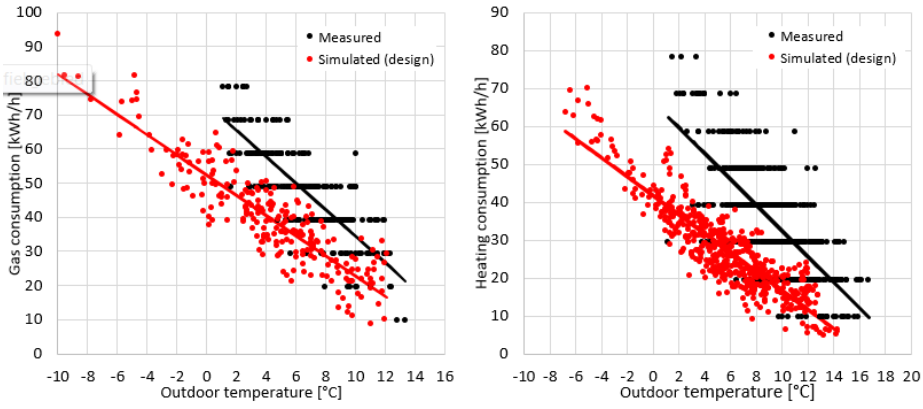


Figure 8: Results Linear regression (least squares method)

The cumulative energy consumption of the case study was assesses in each identified, characteristic time blocks. This way provides more insight in the relevance of the energy gap in each time block. Because a part of the energy gap has already been explained by the results of the Pareto analysis (the selected most critical parameters), the results of the modified building model (with the implemented real values of the critical parameters) are also shown. The most relevant seasons are shown in Fig. 9 (winter). With this comparison, it can be seen in which time blocks the critical parameters have the largest impact and which part of the energy gap is not explained by the results of the Pareto analysis. Based on these findings, the further analysis can be continued more specific and effective. As can be seen from this figures, the major energy gaps are outside the office hours, in time block A & F (0:00-7:00 h on Tuesday-Friday and during weekends (Saturday 0:00 h- Monday 5:00 h)). In the next step the energy gaps are further analyzed, to identify the cause of this relatively large difference.

Table 2: Comparison of measured and simulated heating consumption in winter period. (Huls 2016)

		Design (simulated) [kWh]	Measured [kWh]	Modified from Pareto analysis (simulated) [kWh]
A	**0:00- 7:00 h (Tuesdays - Fridays)**	1.100	14.300	7.100
B	**5:00-8:00 h (Mondays) and 7:00-8:00 h (Tuesday-Friday):**	8.100	6.900	7.500
C	**8:00-11:00 h**	9.800	11.800	11.700
D	**11:00-18:00 h**	15.000	18.000	15.700
E	**18:00-24:00 h**	200	350	50
F	**Weekend (Saturday 0:00 h. – Monday 5:00 h)**	5.900	13.200	4.100

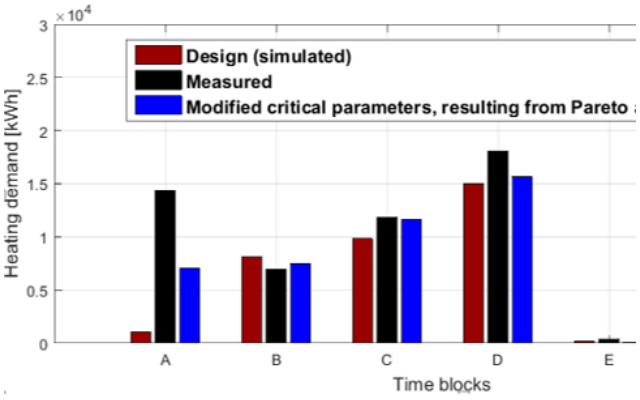


Figure 9: Winter season (January, February, November & December): comparison of the simulated and the measured heating consumption in the defined time blocks

LEAN Energy Analysis, step 4: Assessment of remaining energy gaps (which can not be explained by the results of the Pareto analysis)

In this step, the results of the LEAN Energy Analysis are used to assess the energy performance gaps, additionally to the earlier results of the Pareto Analysis. It was

concluded that the remaining energy gaps in heating consumption mainly occur in time block A and F, which are both outside office hours. Therefore, the simulated and measured heating profiles of these time blocks were considered and shown in Fig. 10. In Fig. 10, the tight correlation between outside temperature and the measured heating demand seems to indicate a heating demand from the building. Because heating is supplied by only one boiler, with only a few control 'rules', a more detailed analysis is started with this control of the boiler. In Fig. 11, the supplied heat of the boiler was analyzed during a week. The red line indicates the supplied heat by the boiler, the black line indicates the energy consumed by auxiliary energy (pumps and fans). From this black line, the 'daytime' time schedule can be derived.

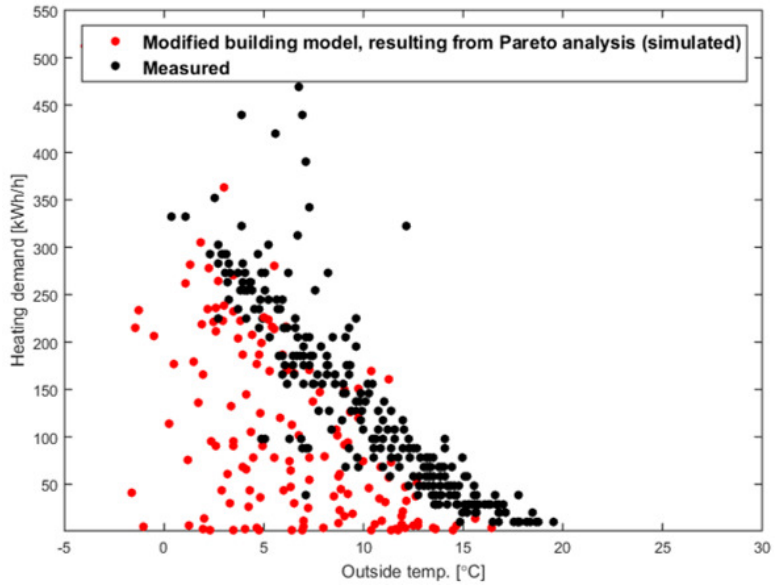


Figure 10: Simulated versus measured heating profile

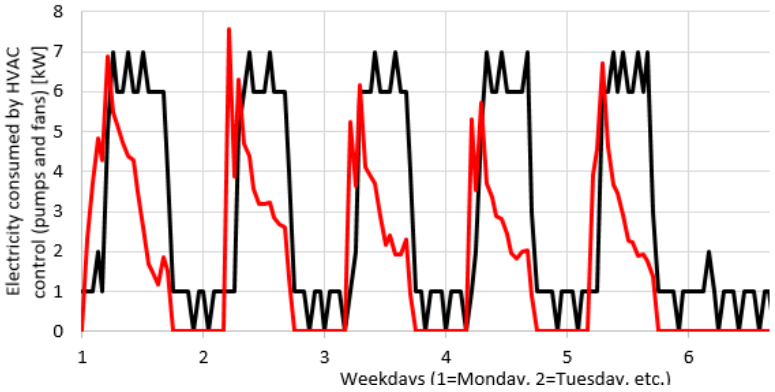


Figure 11: Supplied heat of the boiler, during a week.

Discussion: Strengths and weaknesses Pareto analysis and L.E.A.

In an evaluation of the used Pareto analysis and LEAN Energy Analysis, the strengths and weaknesses of both methods are evaluated based on the results of the three studies and 'summarized' to an assessment of the functionality and feasibility for both shown in table 3 (Huls 2016). Based on this, a combined energy assessment method is recommended using the strengths of both methods for efficient assessment of energy performance gaps.

Table 3: The strengths and weaknesses of the Pareto analysis and the LEAN Energy Analysis 'summarized' on the functionality and feasibility of both methods.

	Functionality (identification of energy gaps and their causes)		Feasibility (systematics, applicability and required amount of data)	
Pareto	+/-	Because of the interdependency of involved parameters, it is hard to define a few major parameters which can be generally considered as the 'critical parameters' and really explain the major part of energy gaps. On the other hand, it provides useful information about 'risky' parameters, improving the functionality of quick energy scans in existing buildings.	It consists of a clear, stepwise approach. By repeating this analysis for multiple case studies, correlations between building (system) characteristics and critical parameters can be investigated. With more generalizing conclusions about critical parameters, the required amount of data for an energy assessment, will decrease.	+
L.E.A	+	Causes of energy performance gaps are successfully identified by trend analysis of energy consumption within characteristic time blocks	A systematic approach helps in trend analysis, but it leads less rapidly to specific causes, compared to the Pareto analysis. Furthermore, detailed energy data (e.g. on component, room and/or floor-level) is needed to identify the real causes from trend analysis.	+/-

Conclusions

The objective of the research is the systematic identification of the major discrepancies between measured energy performances and the expected performances, which are based on a reference situation (e.g. design). To achieve this objective the Pareto analysis and the L.E.A were used to performed an analysis of the time series of two Dutch office buildings. Using the strengths of both methods results in a successful, combined method for analysis of energy performance gaps in the existing buildings. By identification and assessment of critical parameters, the most important causes in the use of building (system) were identified.

References

ASHRAE, Guideline 14. Measurement of Energy, Demand, and Water Savings, vol. 2014.

Bartlett, B.M., 2015 [online]. Powerful Strategic Planning Rules That Could Revolutionize Your Business And Life. Available at: www.benmbartlett.com

Bloomfield C., Bannister P., 2010, Energy and Water Performance Benchmarking in the Retail Sector. NABERS Shopping Centres, pp. 1-13

Borgstein E.H., Lamberts R., Hensen J.L.M., 2016, Evaluating energy performance in non-domestic buildings: a review. Energy and Buildings 128: 734-755.

Buck J., Young D., 2007, The potential for energy efficiency gains in the Canadian commercial building sector: a stochastic frontier study. Energy 32 (9):1769-1780.
Carbon Trust, 2011, Closing the Gap: Lessons learned on realising the potential of low carbon building design. Carbon Trust, London
CarbonBuzz, 2018, Evidence on the performance gap, Retrieved February 13, 2018, from http://www.carbonbuzz.org/evidencetab.jsp
Chung W., 2011, Review of building energy-use performance benchmarking methodologies. Appl. Energy 88(5): 1470-1479
Donnell M., Kummer J., Drees K., 2013, LEAN Energy Analysis; Using Regression Analysis to assess Building Energy Performance. Institute for building efficiency.
Huls A.J., 2016, Systematic energy performance assessment in operating office buildings, Identification and assessment of energy performance gaps in Dutch office buildings, using the Pareto analysis and LEAN Energy Analysis, MSc thesis, TU Eindhoven, Eindhoven, Netherlands
Kissock K., Abels B., Sever F., Ayele D., 2011, Understanding Industrial Energy Use through Lean Energy Analysis, University of Dayton, Ohio.
Kissock K., Seryak J., 2004, LEAN Energy Analysis: Identifying, discovering and tracking energy savings potential. Society of Manufacturing Engineers: Advanced Energy and Fuel Cell Technologies Conference, Livonia, MI.
Lee W.S., Kung C.K., 2011, Using climate classification to evaluate building energy performance. Energy 36 (3): 1797-1801
Lin L.C., Lee W.S., 2011, Evaluating and ranking the energy performance of office building using technique for order preference by similarity to ideal solution. Appl. Therm. Eng. 31 (16): 3521-3525
Menezes C., Cripps A., Bouchlaghem D., Buswell R., 2012, Predicted vs actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap. Appl. Energy 97: 355-364.
Morant M., 2012, The performance gap- non domestic buildings. AECOM/Constructing Excellence Wales, Cardiff
Oregon Department of Energy, 2014, Retro-commissioning in Silver Falls School District. http://www.oregon.gov/energy/CONS/BUS/comm/docs/Silverton.PDF
Schoenmakers I., Zeiler W. Boxem G., 2016, KPI's energy consumption analysis of Isolation Rooms in Hospitals, Proceedings Clima 2016, Aalborg, Denmark
Sever F., Kissock K., Brown D., Mulqueen S., 2011, Estimating Industrial Energy Savings using Inverse Simulation. ASHRAE
Vink W., Zeiler W., 2018, Re-Commissioning the HVAC of an Office Building: Systematic Performance Evaluation according to L.E.A., CIBSE Technical Symposium, London, UK.
Wang E., 2015, Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach, Applied Energy 146: 92-103.
Wilde P de., 2014, The gap between predicted and measured energy performance of buildings: a framework for investigation. Automation in Construction 41: 40-49.
Zhang Y., O'Neil Z., Dong B., Augenbroe G., 2014, Comparisons of inverse modeling approaches for predicting building energy performance, Built Environment 86: 177-190

Adaptive Methods for Energy Forecasting of Production and Demand of Solar Assisted Heating Systems

Viktor Unterberger[1,2], Thomas Nigitz[1,2], Mauro Luzzu[1], Daniel Muschick[1], Markus Gölles[1,*]

1) BIOENERGY 2020+ GmbH, Graz (Austria)
2) Institute for Automation and Control, Graz University of Technology, Graz (Austria)
*) markus.goelles@bioenergy2020.eu

Abstract. **Solar assisted heating systems use the energy of the sun to supply consumers with renewable heat and can be found all over the world where heating of buildings is necessary. For these systems forecasting methods are especially important because both heat production and demand are directly related to the weather conditions. In order to optimally plan production, storage and consumption, forecasts for both the future heat production of the thermal solar collectors as well as the future heat demand of the connected consumers are essential. For this reason this contribution presents novel, adaptive forecast methods for the solar heat production and the heat demand of consumers using weather forecasts. The developed methods are easy to implement and therefore practically applicable. The final verification of the methods shows good agreement between the predicted values and measurement data from a representative solar assisted heating system.**

Keywords: energy forecast, production forecast, demand forecast, solar heat production, heat demand

1 Introduction

Solar assisted heating systems use the energy provided by the sun to supply consumers with renewable heat and can be found all over the world where heating of buildings is necessary. Especially large-scale solar assisted heating systems, which benefit from the effect of scale regarding the solar heat cost, are an important element of future energy systems [3].

These solar assisted heating systems are typically set up with a solar collector field to generate heat and a buffer storage to store the heat and decouple the occurrence of heat production from heat consumption, at least to a certain degree. Furthermore there

are typically one or more auxiliary heating systems (e.g. a gas burner) in order to supply the connected consumers with heat in case there is a lack of solar energy and there is no heat stored in the buffer storage. A schematic representation of such a system is shown in Figure 1.

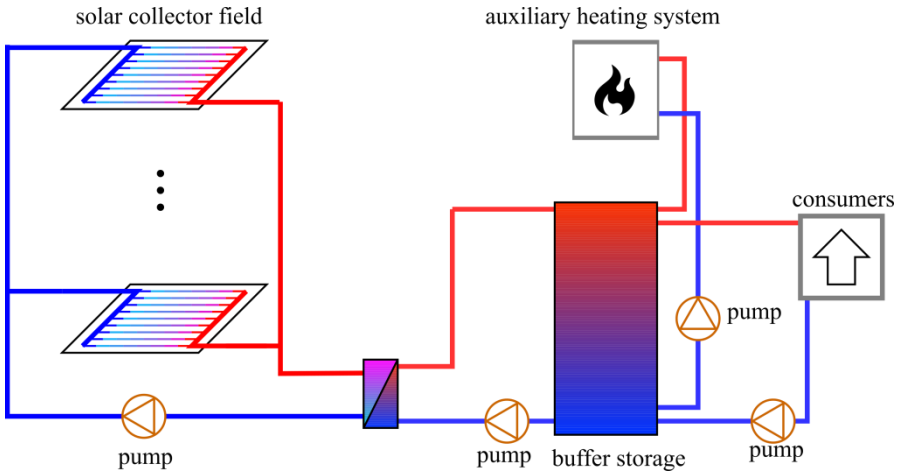


Fig. 1: Schematic representation of a typical solar assisted heating system

However, controlling such a solar assisted heating system can be particularly challenging because the daily weather fluctuations have significant effects on the system's performance, see e.g. [1, 6]. Modern high-level controllers, typically referred to as buffer management or energy management systems, thus rely on forecasts of the future heat production and demand to derive efficient control strategies. These strategies can lead to an overall efficiency improvement in the range of 2-5%, see [10,11,12].

Unfortunately, most of the available forecasting methods for solar heat production as well as for heat demand are tailored for a specific application, not adaptive and often mathematically complicated, see e.g. [2, 5, 7, 8, 9, 13]. In particular because of their complex mathematical structure (e.g. artificial neural networks), it is often not possible to easily implement them on controllers typically used in solar assisted heating systems.

For this reason this paper presents energy forecasting methods for both the solar heat production as well as the heat demand of consumers which are adaptive, easy to implement, and thus suitable for the use in controllers of solar assisted heating systems. Both methods shown in this contribution use weather forecasts from a weather service provider (*meteoblue*, https://www.meteoblue.com) as inputs and their outputs are verified with measurement data from a representative solar assisted heating system.

In the following, the general principle the forecast methods rely on will be described in Section 2. Section 3 then presents its specific implementation for forecasting the solar heat production, which is suitable for a wide range of collectors, and Section 4 describes its implementation for forecasting the heat demand of the consumers. In

both cases verifications of the presented methods based on comparisons with measurement data from a representative solar assisted heating system will be presented. Finally, Section 5 draws conclusions from the presented methods and their practical verification.

2 General Principle

The forecasting method proposed relies on the natural periodicity of the phenomena underlying both the solar heat production and the heat demand of the consumers. In the case of the solar heat production, it is obvious to assume the sun to follow approximately the same path in the sky as the day before and thus to generate approximately the same amount of heat at the same time of day as long as the weather does not change significantly. Likewise the heat demand, assuming similar ambient conditions and user behaviour, will also only change slightly from one day to the next. As a first approximation, it is thus reasonable to predict the solar heat production as well as the heat demand at a particular time of day by using the corresponding values at the same time of the previous day and possibly multiple days in the past. However, looking too far into the past must not be advised since seasonal changes or changes in the users' behaviour would not be taken into account sufficiently.
This rough prediction can be enhanced by additionally considering information on external factors also influencing the variables to be forecast, such as variations in the global irradiation due to clouds or variations in the ambient temperature. Forecasts for the global irradiation and the ambient temperature are available from weather service providers, with a typical sampling time of one hour. As the forecasting method proposed, heavily relies on these forecasts from the weather service provider, the sampling times are aligned, i.e. the time of day is considered as hour of the day.
The easiest way to express a dependency from external factors is through a multiple linear regression model:

$$\tilde{y}_{h,d} = \beta_{0,h} + \sum_{j=1}^{n} \beta_{j,h} \hat{x}_{j,h,d} \tag{1}$$

Here, $\tilde{y}_{h,d}$ denotes the prediction of the variable to be forecast for hour h of day d, $\beta_{0,h}$ is a constant offset, $\hat{x}_{j,h,d}$ is the prediction of the external factor j for hour h of day d, and the model parameters $\beta_{j,h}$ weighting the influence of the n external factors
The dependency from the external factors was found to be different for the individual hours of the day in some cases, thus, different model parameters are used for each individual hour of the day.
In order to determine the model parameters, the model error $e_{h,d}$ is defined as

$$e_{h,d} = \tilde{y}_{h,d} - y_{h,d} \tag{2}$$

with the actual value of the variable to be forecast $y_{h,d}$ for hour h of day d. By using historical values for the variable to be forecast as well as historical values for the external factors to be considered, the optimal model parameters can be calculated for a certain cost function. A reasonable approach for the cost function is the sum of the

squared errors of the model for the corresponding hours of all days in the past which should be used for the model parametrization. So it is possible to determine the optimal model parameters $\beta_{j,h}^{*}$ by solving an ordinary least-squares problem, i.e.

$$\beta_{j,h}^{*} = \text{argmin}_{\beta_{j,h}} \sum_{d=d_0-1}^{d_0-N_d} e_{h,d}^2 \ , \ j = 0, \dots, n \tag{3}$$

Here, d_0 denotes the day when the model parameters are determined, and N_d denotes the number previous days to be used for parametrization.
Apart from only considering the time of the day, it could be necessary to also distinguish between different days of the week. If different days of the week should be distinguished, separate model parameters for each time of the day need to be identified for each day of the week, or, for example, subsets such as workdays and weekend days. A further consideration of the day of the year is advised only, if the forecasting method should provide forecasts with a horizon of several weeks or months, i.e. long-term forecasts. As the forecasting method proposed should support the control of solar assisted heating systems, a forecast horizon of several days up to a week, i.e. short-term forecasts, are required. For short-term forecasts the periodicity in the variable to be forecast, can be deduced more reliably from the previous days, than from the previous year. Hence, the day of the year is considered only implicitly only, by the automatic adaptation of the model parameters when learning from only a few days in the past.

So far, the forecast for one hour only considers measurement data from the last days, but disregards measurements that are more recent because one hour is incorrectly assumed independent from the next. This would mean that the prediction for 3 o'clock p.m. of the following day could be calculated at 4 o'clock p.m. of the current day and would not change after that.

In order to reduce this shortcoming, a correction step is introduced that takes the current prediction error into account. This correction step could act in two ways, depending on the variable to be forecast. If the variable to be forecast equals the solar heat production, then a prediction error typically occurs from inaccuracies in the forecast of the weather service provider or from temporary local shading by for example buildings close to the solar collectors. Such a prediction error is likely to persist for a certain time, and should be considered by correcting the prediction with the current prediction error. If the variable to be forecast equals the heat demand, then a prediction error typically occurs from inaccuracies in the forecasts of the weather service provider or from non-periodic consumer behavior. Here only an error caused by the forecasts from the weather service provider is likely to persist for a certain time. An error caused by non-periodic consumer behavior can align with meeting a demand earlier or later than on the previous days. Then a prediction error does not persist, even more, it is likely to change its sign, which should be considered with the negative current prediction error. However, analysis of measurement data showed that a positive prediction error is more likely to persist for a certain time. So, the correction step adds the weighted prediction error, what can be written as

$$\hat{y}_{h,d} = \tilde{y}_{h,d} + (y_0 - \tilde{y}_0)\,\Phi(\Delta h), \tag{4}$$

where $\hat{y}_{h,d}$ is the corrected forecast for hour h of day d, y_0 is the most current measurement, $\tilde{y}_0$ is the corresponding forecast and Φ is a monotonically decreasing function of the time difference Δh between the most current measurement and the hour given by indices h and d.

This simple and general forecasting method needs to consider different external factors and must take into account different periodicity assumptions depending if it is used to forecast the solar heat production or the heat demand. The following chapters will go into the detailed implementation for solar heat production forecast (Section 3) and heat demand forecast (Section 4).

3 Solar heat production forecast

The heat generation of a solar thermal collector can be described by a static collector model according to the European Standard EN12975:2006:

$$\dot{Q} = A_{\text{coll}}\, c_0\, I_g - A_{\text{coll}}\, c_1\, \Delta T - A_{\text{coll}}\, c_2\, \Delta T^2 \tag{5}$$

Here $\dot{Q}$ denotes the generated heat flow, A_{coll} represents the net collector area, I_g symbolizes the global solar irradiation hitting the collector surface, ΔT is the temperature difference between collector temperature and ambient temperature, and the remaining coefficients represent the optical efficiency c_0, and the heat loss coefficients for heat conductance, c_1, as well as for heat loss through thermal radiation, c_2. The EN12975:2006 standard covers performance, durability and reliability testing of almost all collector types available on the market. Even tracking concentrating collectors were recently fully included in the scope [4]. However, the analysis of measurement data from large-scale solar assisted heating systems shows that applying this model, with its parameters $(A_{coll}, c_0, c_1, c_2)$ taken from the datasheet of the collectors, does not directly lead to satisfying results for forecasting the solar heat production even if the external factors I_g and ΔT are known. This is because the model parameters are only valid for stationary conditions in the laboratory and not for the conditions occurring during daily operation. Furthermore, these model parameters would have to change over time to account for polluted collector surfaces decreasing the optical efficiency (given by c_0), the decay of materials leading to higher heat losses of the collectors (given by c_1 and c_2) and local shading reducing the effective collector area (given by A_{coll}).

These shortcomings can be eliminated when applying the forecast method described in the previous section. Note that the static collector model is identical in structure to the linear regression model with external factors $\hat{x}_1 = I_g$, $\hat{x}_2 = \Delta T$ and $\hat{x}_3 = \Delta T^2$. The physical parameters correspond to the model parameters, i.e. $\beta_0 = 0$, $\beta_1 = A_{\text{coll}} c_0$, $\beta_2 = -A_{\text{coll}} c_1$ and $\beta_3 = -A_{\text{coll}} c_2$.

The forecast model is thus equivalent to the static collector model and is therefore valid for a wide range of collectors. Furthermore, with the forecast method the collector parameters are continuously adapted using measurement data of the produced solar heat, the temperature difference between collector temperature and ambient temperature as well as the global solar irradiation hitting the collector surface. The dependency on the time of day is considered by using different parameter sets for each hour of the day to forecast the solar heat production $\tilde{y}_{h,d}$:

$$\tilde{y}_{h,d} = \beta_{1,h}\hat{x}_{1,h,d} - \beta_{2,h}\hat{x}_{2,h,d} - \beta_{3,h}\hat{x}_{3,h,d} \quad (6)$$

By this procedure the influence of pollution of the collector fields and of decay of the materials is automatically considered.
What is more, the effects of local shading, which would have to be incorporated into the static collector model by complicated 3D modelling and shadow calculations, is automatically considered and no manual parameterization is necessary. In addition, later enhancements such as additional solar panels or changing environmental conditions such as additional buildings throwing shadows are automatically incorporated into the forecasting process.

The forecasting method for the solar heat production is evaluated for days of the year where heat production is significant but challenging to predict. Solar heat production during winter is typically very small or negligible, which makes forecasting not very interesting and its benefit small. During the summer months solar heat production is typically at its highest, but weather conditions are quite steady, which makes forecasts interesting but less challenging. In spring and autumn solar heat production can be high too, while the weather conditions can be very unsteady. This makes forecasts interesting and more challenging.

Therefore, the forecasting method for solar heat production is evaluated using both a day in summer and a day in spring for comparison. It is applied to a collector field with a collector net area of 138 m² from a solar assisted heating system built in Austria in the year 2009.
The evaluation of the day in summer is visualized in Fig. 2, showing the measured heat produced by the solar collector (black), the consecutive hourly forecasts (in colour, starting with green, going via yellow to red and blue) and the output of the static collector model with the parameters taken from the datasheet (grey).

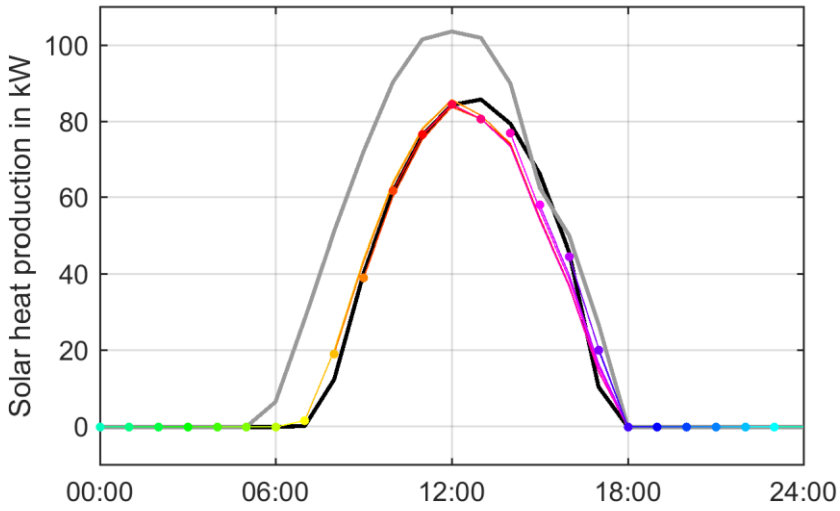


Fig. 2: Forecast of solar heat production during a day in summer

From 0 o'clock a.m. to 12 o'clock the forecast is very accurate and predicts the solar heat production very precisely. After that there is a slight variation which gets corrected over the next hours via the current prediction error, which slightly improves the next forecast. In addition, the forecast predicts the start and stop time of the solar heat production very accurately. While the static collector model (grey) will immediately predict heat output as soon as global irradiation levels increase, the actual system needs time to heat up until the temperature is above the necessary limit and only then will provide hot water. This is important information for the controller of a solar assisted heating system.

The evaluation of the day in spring is visualized in Fig. 3.

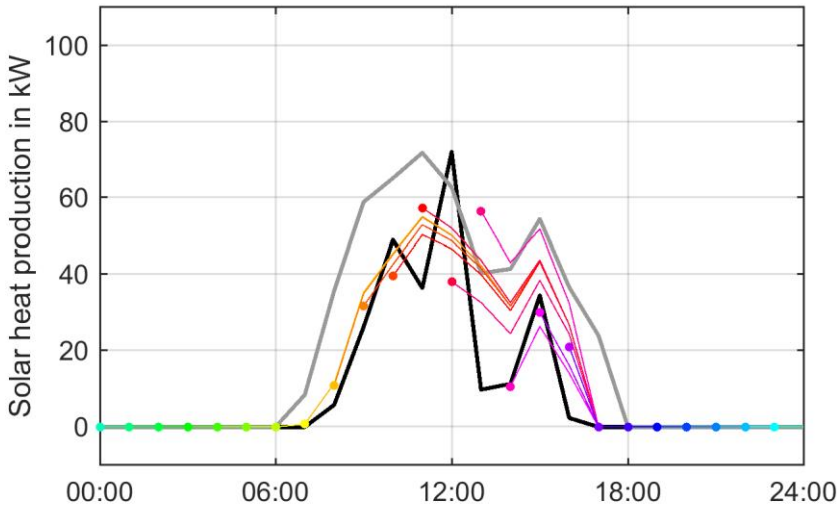


Fig. 3: Forecast of solar heat production during a day in spring

For this spring day the heat produced is about half that of the summer day. From midnight to 11 o'clock a.m. the forecast is very accurate. At 11 o'clock a.m. the forecast estimates the heat production too high, which is why the next prediction at 12 o'clock, which was already a bit too low, is corrected downwards. The opposite is true for 1 o'clock p.m., when actual production drops in an unexpected way, but the forecast is corrected upwards because of the underestimation at 12 o'clock. Days where the solar heat production drops and rises repeatedly, which can happen through rapid cloud movements, are therefore very challenging for the forecast, especially if these changes are not apparent from the global irradiation forecasts. However after 2 o'clock p.m. the forecast is sufficiently accurate again and the start and stop times of the solar heat production are predicted sufficiently well too. In contrast, the static collector model (grey) predicts the heat output as soon as global irradiation is available, which leads to wrong predictions of the start and stop times of solar heat production.

4 Heat demand forecast

The heat demand of consumers depends on many external factors such as the users' behaviour patterns and weather factors such as ambient temperature, solar irradiation, rain and wind. However, most of these factors are either hard to predict, follow a periodic pattern themselves or only have a relatively small influence on the heat demand. For example, the wind direction and speed might play a role because of higher ambient losses but they cannot get predicted sufficiently well for an individual building. Solar irradiation might play a major role especially with modern glass palaces, but its effect is mostly determined by the angle of incidence and thus the time of day. After investigating the influence of the individual factors on multiple test objects

(family homes, office buildings etc.), it was found that considering the ambient temperature T_{amb} is enough for obtaining sufficient heat demand forecasts $\tilde{y}_{h,d}$. This reduces the linear regression model (1) to:

$$\tilde{y}_{h,d} = \beta_{0,h} + \beta_{1,h}\hat{x}_{1,h,d} \tag{7}$$

with $\hat{x}_{1,h,d}$ representing the forecasted ambient temperature $\hat{T}_{amb,h,d}$ for hour h of day d. As in the common case the time of the day needs to be considered. Additionally, a distinction between workdays and weekend days improves the forecasting quality significantly. Hence, separate hourly model parameters are considered for workdays and weekend days. Seasonal variations and changing users' behaviour are considered by continuously updating these model parameters using measured heat demands and ambient temperatures of the last two weeks.

The forecasting method for the heat demand should be evaluated for days of the year where its quality is challenged. The heat demand during winter is mainly dominated by room heating, which typically makes forecasts easy. The heat demand during summer is mainly dominated by social consumer behaviour, which typically makes forecasts hard. In spring and autumn a combination of both is present. Therefore the forecasting method for heat demand is evaluated using a day in spring and a day in summer like it was used for the evaluation of the solar heat production forecast in Section 3. The consumer used for the evaluation is an office building with a connected load of 300 kW supplied by the solar assisted heating system also used for the evaluation of the solar heat production forecast in Section 3. The evaluation of the day in spring is visualized in Fig. 4, showing the measured heat demand of the consumer (in black) and the consecutive hourly forecasts (in colour).

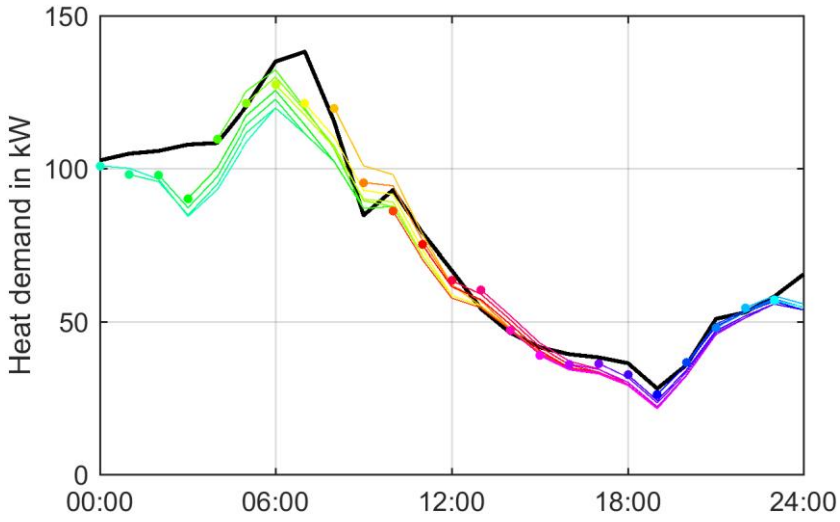


Fig. 4: Forecast of heat demand during a day in spring

From midnight to 6 o'clock a.m. the forecast underestimates the demand by a certain amount, but for the rest of the day the forecast is quite accurate. The consecutive forecasts are corrected via the current prediction error, which improves the next forecast. The evaluation of the day in summer is visualized in Fig. 5.

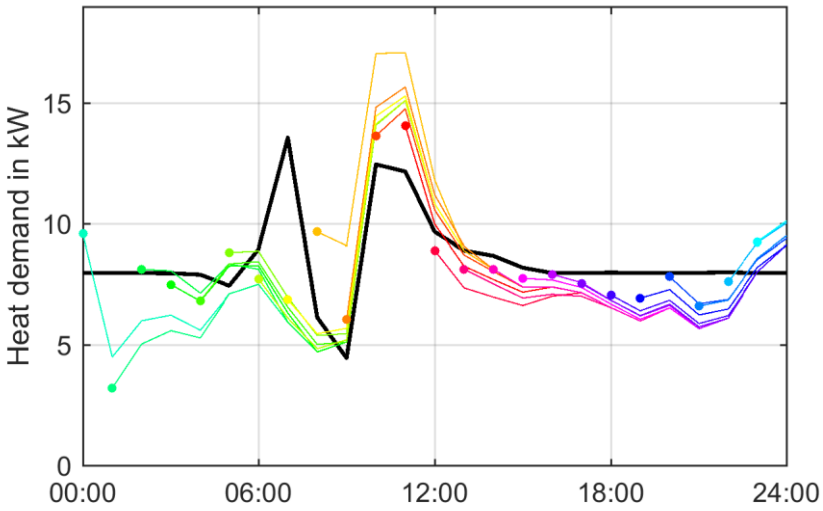


Fig. 5: Forecast of heat demand of a day in summer

Here the demand is around a tenth of the demand in spring. The forecast differs stronger from the measurements, as in summer individual actions by the users dominate the heat demand and neither periodicity nor the predicted external factors play a major role. However, while the *relative* prediction errors are quite large, the *absolute* prediction errors are quite small.

5 Conclusion

The forecasting method proposed in this article is sufficiently simple to be implemented even on very simple computers (programmable logic controllers, PLCs) but still automatically learns from past data and thus contains no parameters, which need to be tuned manually. It is based on a linear regression model, using forecasts of external factors, that accounts for changes relative to a simple periodic signal extrapolation. A simple correction step takes into account recent prediction errors to further improve the forecasts for the next hours. The forecasting method was shown to work reasonably well for forecasting solar heat production and heat demand. With few modifications, it can also be used to forecast other quantities such as electrical power demand and photovoltaic electrical energy production.

^^The only requirement of the forecasting method is the availability of measurement data for both the quantity being forecasted and the external factors influencing that

quantity, as well as forecasts of the external factors themselves. These forecasts can often be obtained from weather service providers, e.g. using web interfaces. It is important to note that the forecast quality of both production and demand naturally strongly depends on the forecast quality of the external factors such as ambient temperature and global irradiation. Therefore, it is worth considering commercial weather service providers, which typically offer better forecast quality.

By continuously updating the model parameters from past data, the method automatically adapts to changing behavior patterns or environmental conditions and needs no kind of manual parametrization. It is, therefore, a plug-and-play solution that requires no experts for implementation and parametrization. It can be used in software ranging from simple buffer management controllers to complicated, optimization-based energy management systems, to improve the control of solar assisted heating systems.

6 References

1. Camacho, E.F, Berenguel, M., Control of Solar Energy Systems1, IFAC Proceedings Volumes, Volume 45, Issue 15, 2012, Pages 848-855.

2. Fang, T., Lahdelma, R., Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system, Applied Energy 179, 2016, Pages 544-552.

3. Fisch, M.N, Guigas, M., Dalenbäck, J.O, A REVIEW OF LARGE-SCALE SOLAR HEATING SYSTEMS IN EUROPE, Solar Energy, Volume 63, Issue 6, 1998, Pages 355-366.

4. Kovacs, P., A guide to the standard EN 12975, SP – Technical Research Institute of Sweden, QAiST - IEE/08/593/SI2.529236, Deliverable D2.3, 2012.

5. Kramer, W., Bitterling, M., Artificial Neural Networks (ANN) for the Prediction of Local Outside Temperatures and Solar Yields, in Proceedings SWC2017, doi:10.18086/swc.2017.22.0.

6. Lemos, J.M, Neves-Silva, R., Igreja, J.M, Adaptive Control of Solar Energy Collector Systems, Advances in Industrial Control, 2014, Springer International Publishing.

7. Mestekemper, T., Kauermann, G., Smith, M. S, A comparison of periodic autoregressive and dynamic factor models in intraday energy demand forecasting, International Journal of Forecasting 29 (1), 2013, Pages 1-12.

8. Neto, A.H, Fiorelli, F.A.S, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy and Buildings, Volume 40, Issue 12, 2008, Pages 2169-2176.

9. Nielsen, H. A, Madsen, H., Modelling the heat consumption in district heating systems using a grey-box approach, Energy and Buildings 38 (1), 2006, Pages 63-71.

10. Quintana, H., Kummert, M., Potential of model predictive control (MPC) strategies for the operation of solar communities. Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association, 2013, Pages 2481-2488.

11. Quintana, H., Kummert, M., Optimized control strategies for solar district heating systems. Journal of Building Performance Simulation, 8, (2014), Pages 1-18.

12. Unterberger, V., Lichtenegger, K., Innerhofer, P., Gerardts, B., Gölles, M., Evaluation of the Potential for Efficiency Increase by the Application of Model-Based Control Strategies in Large-Scale Solar Thermal Plants. International Journal of Contemporary ENERGY. 2018; 4(1), Pages 549-559.

13. Zhao, H., Magoulès, F., A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews, Volume 16, Issue 6, 2012, Pages 3586-3592.

Prediction of Current by Artificial Neural Networks in a Substation in order to Schedule Thermography

Per Westerlund* and Ilias Dimoulkas

KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science
Teknikringen 31
SE-100 44 Stockholm
Sweden
perw@kth.se

Abstract. Thermography or infra-red imaging is a method that measures the temperature of a surface by receiving the infra-red radiation that the surface emits. Thermography is used in, for example, condition measuring of electrical equipment. It shows which parts are heated more than normally due to a higher resistance. Those parts will need maintenance. In order to get accurate values, thermography needs a high current in the equipment. Thus it is necessary to predict when the current will be high throughout the year. Here a neural network with two layers is used for the prediction. The data set consists of the hourly currents at a point in a Swedish substation from a period of ten years.
The purpose is to plan when to go to a substation to do thermography. As the prediction is done several months ahead, the outdoor temperature cannot be used. Hence only the time expressed as week, day and hour with different resolutions in the discretization, is used as an explanatory variable. With increasing resolutions in the discretization, the prediction error decreases. Adding inputs based on interaction does not improve the prediction. The results are however not satisfactory as the prediction error is large in comparison with the predicted values of the current and the prediction is biased. One reason is that the prediction should be several months ahead, so the actual temperature cannot be used.

Keywords: artificial neural networks, ANN, infrared sensors, IR, prediction algorithms, substations, thermography

1 Introduction

The electric grid consists of overhead lines and cables connected at substations, where there are transformers to increase and decrease the voltage. There is also

* The research is mainly funded by Energimyndigheten (Swedish Energy Agency) and Svenska kraftnät (Swedish National Grid) through SweGRIDS, Swedish Centre for Smart Grids and Energy Storage, swegrids.se.

other equipment in order to control the flow of electric power, such as circuit breakers, disconnectors and measurement transformers.

Since it is difficult to get an outage in order to measure the condition of the equipment, there are several methods that measure the condition while the equipment is still in use. One such method is thermography, which captures the infra-red (IR) radiation from a surface and calculates the temperature of the surface. Thus it can detect increasing contact resistance.

The current in a line, in a cable or in an apparatus varies throughout the day. There is also variation between the working days and the weekend and during the whole year. It is interesting to predict the current, for example, if thermography should be done to equipment connected to that line, as thermography is more accurate at high currents. [2], [3] A recommendation is that thermography should be done with an interval of between one and three years. [6, U 510:99] Normally thermography is planned well in advance (up to a year ahead), as there are no so many technicians available. Also dynamic rating is based on predicting the current of the line. Furthermore, prediction of load is important for the planning of the operation of the electric grid.

1.1 Problem definition

The main problem covered here is to plan thermography in an electrical network. Thermography consists of make infra-red images of an installation. The infra-red radiation is a measure of the temperature of a surface.

The output is the time when a technician should be sent to a certain substation. The current should be at least one third of the current that the device is rated for during the measurement in order to get valuable results.

The input consists of previous values of the current at that point at the substation. The time horizon for the prediction is about one year, as the scheduling of thermography uses that time horizon. Thus the temperature cannot be used as an explanatory variable.

Ringwood et al used both a linear model and a neural network to predict the current with three different perspectives: 24 hours, 3 years and 15 years ahead. [4] The models also include a rise throughout the years.

2 Data

The data used here is the same as in [7]. The current in a line between two substations in the Swedish transmission grid has been recorded for a period of more than 10 years. The current is given in amperes and the values are the average of each hour. The histogram of the current is shown in Fig. 1. As thermography depends only on the magnitude of the current and not on the sign, which indicates the direction in which it flows, Fig. 2 shows the histogram of the absolute value of the current. The current should be at least one third of the current that the device is rated for, in this case 3150 amperes.

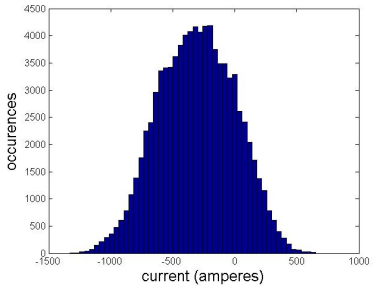


Fig. 1. Histogram of the current in the studied substation.

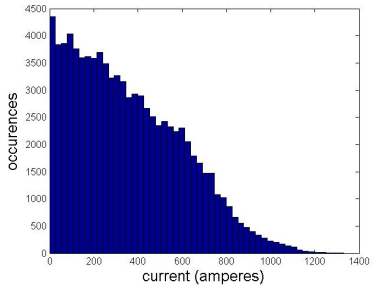


Fig. 2. Histogram of the absolute value of the current in the studied substation.

3 Preprocessing of data

As there are three cycles, the time will be represented by three variables: week, day and hour. The period length of these is a trade-off between the precision in the prediction and the amount of data for each period.

The day will be divided into periods of 24 hours, 8, 6, 4, 3, 2 and 1 hour, which are the values of the parameter n_h. If $n_h = 4$, the period 00–04 will have the same predicted current, the period 04–08 will have another one and so forth for the six periods of the day.

The weekly cycle is handled by considering all days to be the same, by dividing it into weekdays and weekends, and by having a separate prediction for each weekday. The parameter n_d has then the values 7, 2 and 1 respectively.

Similarly the year is divided into periods of 52 weeks, 13, 4 and 1 week, which is the value of the third variable is denominated n_h, The period of two weeks is not considered as it is close to one and four weeks, so its values can be 52, 13, 4 and 1.

The sign of the values is kept, although the purpose of the prediction does not need it. The reason is that the prediction can be better using all the available data.

4 Neural networks

A feed-forward neural network is used with back-propagation using the scaled conjugate gradient method to update the weights and the biases. A regularization parameter of 0.005 is used to avoid overfitting. The inputs to the network are the time of the day, the weekday and the week number according to the preceding section. Then there are two hidden layers with 20 and 10 neurons respectively and finally the output, which is the predicted current.

5 Simulations

The data set is divided into:

- 70 % for the training set, which is used to determine the parameters,
- 15 % for the validation set, which corrects for overfitting, and
- 15 % for the test set, which is used to evaluate the trained network.

The test set is the last part of the data set, about 80 weeks. The training set and the validation set are interleaved with each other and they are separated by a random procedure.

The simulations were done by Matlab using the Neural Network Toolbox, both without and with interaction between the explanatory variables. Here are some examples without interaction. Fig. 3 shows how the network is able to forecast the training part of the data, dividing the time into 13-week periods, working days and weekends, and 6-hour periods. Thus the parameters are $(n_h, n_d, n_w) = (4, 2, 13)$. Fig. 4 displays a smoothed version of the previous figure. The weekly and daily variations are smoothed out by a filter covering one week and one hour with a half step at the beginning and at the end, a uniform filter with length $7 \cdot 24 + 1 = 169$ for eliminating the seasonality. [1, sec 5.3] Figs. 5–6 are the same as Figs. 3–4 but they display the forecast of the test set. Figs. 7–10 have the finest possible discretization of the three explanatory variables, $(n_h, n_d, n_w) = (1, 1, 1)$, and they show the same as Figs. 3–6.

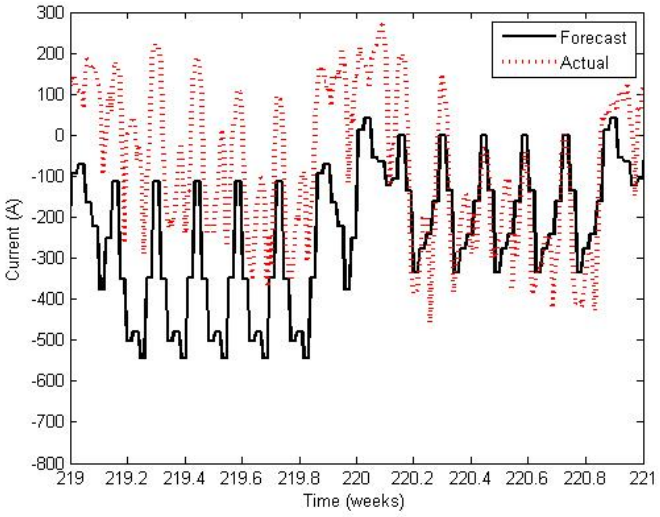


Fig. 3. Forecasting of the training set with $(n_h, n_d, n_w) = (4, 2, 13)$.

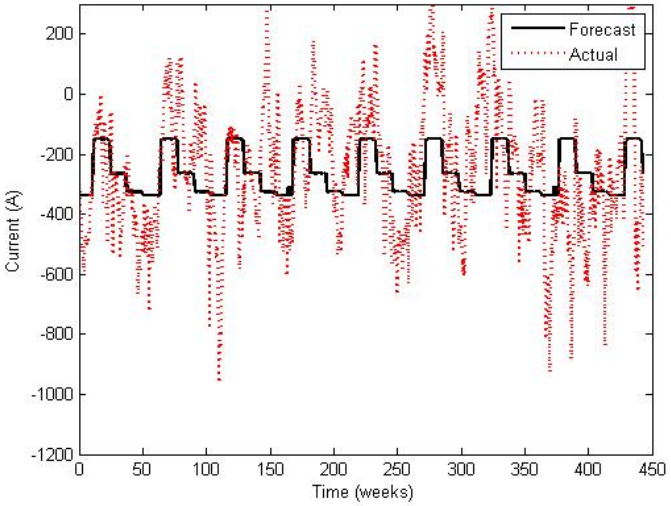


Fig. 4. Smoothed forecasting of the training set with $(n_h, n_d, n_w) = (4, 2, 13)$.

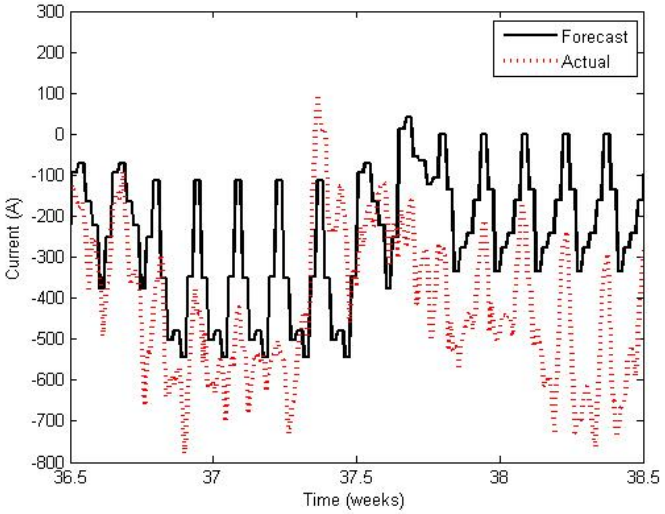


Fig. 5. Forecasting of the test set with $(n_h, n_d, n_w) = (4, 2, 13)$.

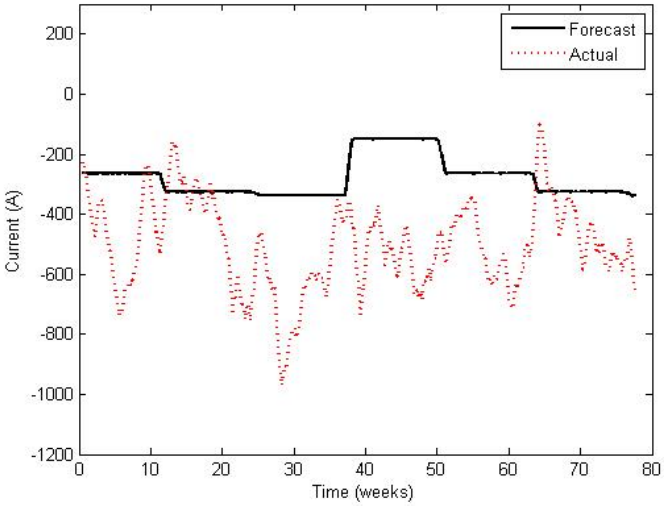


Fig. 6. Smoothed forecasting of the test set with $(n_h, n_d, n_w) = (4, 2, 13)$.

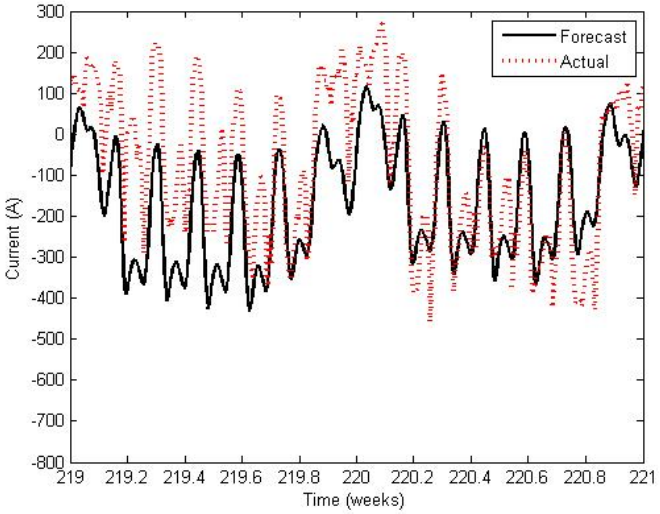


Fig. 7. Forecasting of the training set with $(n_h, n_d, n_w) = (1, 1, 1)$.

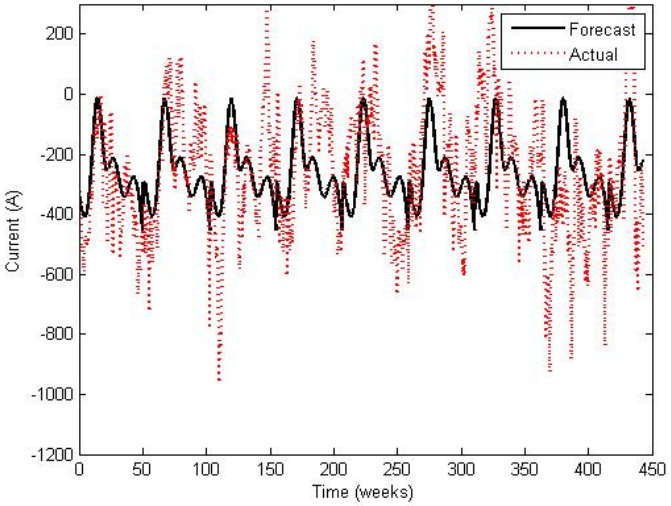


Fig. 8. Smoothed forecasting of the training set with $(n_h, n_d, n_w) = (1, 1, 1)$.

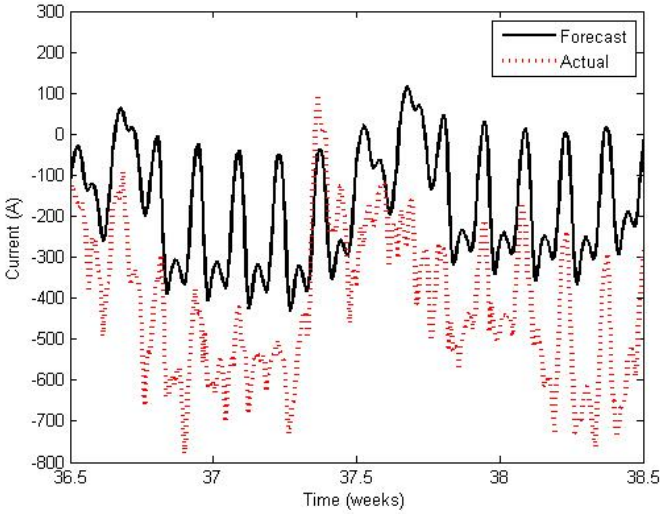


Fig. 9. Forecasting of the test set with $(n_h, n_d, n_w) = (1, 1, 1)$.

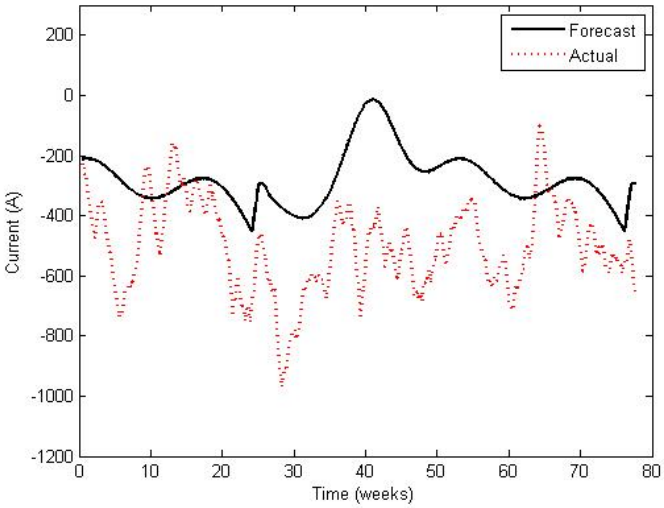


Fig. 10. Smoothed forecasting of the test set with $(n_h, n_d, n_w) = (1, 1, 1)$.

6 Results

The predictions for the test set are compared with the actual results. Both the bias, measured as the average of the difference between predicted and actual value, and three measures of the dispersion – the mean absolute error (MAE), the mean absolute error (MSE) and the root of MSE (RMSE) – are used. The results are shown in Table 1 for the two different sets of parameters without interaction. Tables 2–3 contains the results with interaction for the two parameter settings. The interaction is given by four bits, which show if there is interaction. The order is: between hour and day, between hour and week, between day and week, and, finally, between all three variables: hour, day and week. The interaction between two variables is calculated as the sum of the first variable times the number of possible values of the second variable, and the second variable. The interaction between all the three variables is calculated in a similar way. The values are presented with more decimals in order to show the differences.

n_h, n_d, n_w	4, 2, 13	1, 1, 1
MAE	269	257
MSE	104000	93900
RMSE	322	307
Bias	229	228

Table 1. Results for two different parameter settings.

Interaction	0,0,0,0	1,0,0,0	0,1,0,0	0,0,1,0	1,1,0,0	1,0,1,0	0,1,1,0	1,1,1,1
MAE	269.52	269.88	269.62	269.99	269.54	269.39	269.39	269.39
MSE	104000	104000	104000	104000	104000	104000	104000	104000
RMSE	322.42	322.73	322.47	322.72	322.32	322.18	322.18	322.18
Bias	229.69	229.82	229.52	229.57	229.87	228.96	228.96	228.96

Table 2. Results for with different interactions for $(n_h, n_d, n_w) = (4, 2, 13)$.

Interaction	0,0,0,0	1,0,0,0	0,1,0,0	0,0,1,0	1,1,0,0	1,0,1,0	0,1,1,0	1,1,1,1
MAE	255.83	256.22	256.98	255.70	257.03	257.00	257.51	259.22
MSE	93900	91900	93100	91500	92400	92500	94800	95400
RMSE	307	303.10	305.15	302.55	304.01	304.15	307.82	308.92
Bias	228.24	229.01	228.29	228.88	228.95	228.37	227.37	228.49

Table 3. Results for with different interactions for $(n_h, n_d, n_w) = (1, 1, 1)$.

7 Discussion

Starting with the coarse discretization, $(n_h, n_d, n_w) = (4, 2, 13)$, the daily pattern appears in Fig. 3 as a negative peak for each day, whereas the weekly pattern is indicated by the lower absolute level of the weekend, for example between 23.0 and 23.2. As only six weeks are plotted, just parts of two different periods of the year appear, with a shift in level between them at 24.6. The pattern with four 13-week periods is clear in Fig. 4, which covers the whole training set, as the daily and the weekly variations are smoothed out. The actual values vary considerable from the predicted values, but the prediction is not biased, which is expected. The test set is displayed in Figs. 5–6 and there is a bias between the predicted values and the actual values, which are in general lower.

Studying the results with the maximal discretization, Figs. 7 and 9 show a more adapted prediction (difference between weekdays and not so stepwise daily variation) as the input variables can assume more values. The smoothed variants, Figs. 8 and 10, show that the prediction of the test set is biased.

Table 2 shows that adding interaction variables does not result in smaller errors. By construction some kind of interaction between the explanatory variables is already available in a neural network.

From Table 1 one can deduce that the unexplained variation is about as large is the bias in the prediction. Also the mean error has the same magnitude as the predictions. Interaction gives only a slight improvement of the prediction errors. So the ability to predict the current under these conditions by using a neural network is low. The analysis of variances did not result in much better predictions [7].

Another problem with predicting when high currents occur is that they are rare in this case. Fig. 2 shows that the current is high enough (more than one third of the rated current of 3150 amperes) only during a fraction of the time. Hence, for this specific case any available method cannot perform well. However it is interesting to compare the methods and how well they perform.

8 Conclusion

The purpose is to plan when to go to a substation to do thermography. As the prediction is done several months ahead, the outdoor temperature cannot be used. Hence only the time expressed as week, day and hour with different resolutions in the discretization, is used as an explanatory variable. With increasing resolutions in the discretization, the prediction error decreases. In this case the weeks are grouped into periods of 13 weeks, the only difference is between working day and weekends, and the hours are group four by four. This coarse discretization is compared to the maximal one (all the 52 weeks, 7 days and 24 hours considered separately), whose errors are reduced to the half or to one third.

Adding inputs based on interaction of the explanatory variables does not improves the prediction, as interaction is already part of the calculations. The results are however not satisfactory as the prediction error is large in comparison with the predicted values of the current and the prediction is biased. One reason is that the prediction should be several months ahead, so the actual temperature cannot be used, a factor that influence the current.

Future work could consist of adding an explanatory variable such as the temperature in order to predict better. Although the meteorological forecasts cannot predict the temperature more than two weeks ahead, it is interesting to see how much the temperature could explain. A comparison with linear models as in [4] is also a logical step. Soares and Medeiros have also included a trend in the prediction, which appears to be useful in other cases as well. [5]

References

1. Göran Andersson, Ulf Jorner, and Anders Ågren. *Regressions- och tidsserieanalys*. Studentlitteratur, Lund, 2nd edition, 1994.
2. Tommie Lindquist. Estimating the accuracy of thermography of disconnector contacts in the field. Technical Report 2008:010, KTH, Electromagnetic Engineering, 2008. QC 20100823.
3. Tommie Lindquist. *On reliability and maintenance modelling of ageing equipment in electric power systems*. PhD thesis, KTH, Electromagnetic Engineering, 2008. Trita-EE, 2008:016.
4. J. V. Ringwood, D. Bofelli, and F. T. Murray. Forecasting electricity demand on short, medium and long time scales using neural networks. *Journal of Intelligent and Robotic Systems*, 31(1):129–147, May 2001.
5. L. J. Soares and M. C. Medeiros. Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting*, 24(4):630–644, Oct 2008.
6. Sveriges elleverantörer. *EBR. Underhåll stationer 12–420 kV (EBR. Maintenance of substations 12–420 kV)*, 1999. U 501:99–531:99.
7. Per Westerlund, Patrik Hilber, and Tommie Lindquist. Prediction of current in a substation in order to schedule thermography. In *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–7, Oct 2016.

Permutation entropy as the measure of globalization process.

Janusz Miśkiewicz

Institute of Theoretical Physics, University of Wrocław,
pl. M. Borna 9, 50-204 Wrocław, Poland
and
Department of Physics and Biophysics,
Wrocaw University of Environmental and Life Sciences
ul. Norwida 25, 50-375 Wrocaw, Poland
{janusz.miskiewicz}@ift.uni.wroc.pl

Abstract. The application of permutation entropy as the globalization measure is discussed within this paper. In order to verify stated hypothesis the following time series were investigated: CO_2 emission, CPI, employment rate and GDP. The countries were grouped such that one could easily point the strongly cooperating countries with high level of globalization and compare with the countries with low level of cooperation.

1 Introduction

Globalization is a process that reduces the distance between societies on different levels. This process initiated inventions facilitating transport, such as railways, cars, aeroplanes and communication - telegraph, telephone, radio, television, and the Internet. This process has transformed the evolution of societies from a local phenomenon to a global phenomenon by by strengthening long-range interaction. At present, communication means allow for easy and intense contacts between social groups. This process significantly influenced not only culture but, particularly, economy and politics. One of the most spectacular globalization process in the 20th century was the creation of the European Community, which at the level of political and legal regulation facilitates and improves the functioning of its members, and influence various aspects of life of EC citizens. In fact researches on globalization has a long history [1]. It was made a huge effort in finding proper description of the process. Many indexes have been introduces e.g. Global Connectedness Index, Global Entrepreneurship Index, Global Food Security Index, Global Gender Gap Report, Global Hunger Index, Global Innovation Index, Global Liveability Ranking, Global Peace Index, Global Slavery Index, Global Terrorism Index, Global Web Index, Good Country Index, Government competitiveness, Happy Planet Index, Human Development Index, Human Poverty Index, Index of Economic Freedom, Internationalization Index, KOF Index of Globalisation, Linguistic Diversity Index, Maastricht Globalization Index, Networked Readiness Index, OECD Better Life Index, Energy Globalisation Index

and many others. The common features of mentioned indexes is that they are based on a designated feature, which is considered as the characteristic of the globalization. In the present study a different approach is proposed. The globalization measure is constructed, such that it would be possible to investigate various aspects of globalization by the same measure.

2 Globalisation measure

In the 21st century the flow of information is the most dominating factor determining globalization. It is very probable that this tendency will continue in future. Thus, the measures related to the analysis of information seem to be the most promising. Another factor increasing globalization are the legal and political decisions. These are primarily regulations in the field of economy, international cooperation at the bilateral or multilateral level (e.g. the European Union). These factors unify the conditions in which companies operate. Taking into account the described factors, it seems that the proper measure of globalization should be entropy, in particular information entropy, which characterizes the degree of noise / determinism in given data sets. The main difficulty in direct use of Shannon information entropy (Eq.1) is the limitation of the available data.

$$S = -\sum_i p_i ln(p_i) \tag{1}$$

Obtaining S requires knowledge of probabilities or occurrence of some quantities or even the probability distribution function on analysed parameters. In the case of macroeconomy it is hardly possible. For example, Gross Domestic Product (GDP) is usually presented as annually time series, in consequence typical, developed economy is described by a series of 50 – 100 data points. On the other hand in a long time span one has to face the stationarity problem of such a time series. One of the possible solution of the mentioned difficulties is application of entropy related parameter i.e. Theil index [2, 3].

In the present paper the approach based on permutation entropy is proposed [4, 5]. The main difference between Shannon entropy and permutation entropy is that the latter one is focused on patterns rather than on probability distribution function, which makes it possible to apply in analysis of short time series.

The permutation entropy:
At each time s of the analysed time series $X = x_t, t = 1, \ldots, N$, a vector of D-th subsequent values is taken:

$$(x_s, x_{s+1}, \ldots, x_{s+D-1}). \tag{2}$$

This vector is sorted in an ascending order, and a permutation pattern π is created.

$$\pi = (r_0 r_1 \ldots r_{D-1}) \tag{3}$$

which fulfils

$$x_{s+r_0} \leq x_{s+r_1} \leq \ldots x_{s+r_{D-2}} \leq x_{s+r_{D-1}} \tag{4}$$

Finally the permutation entropy:

$$PE = -\sum_{i=1}^{D!} \pi_i \ln \pi_i \tag{5}$$

In the case of globalization the permutation entropy should be followed be cross-correlation analysis. Within the present study the Manhattan distance was used. Denoting the time series as A and B and theirs elements as a_i and b_i respectively the Manhattan distance between A and B in the interval $i \in (t, t+T)$ is defined:

$$MD(A,B)_{t,T} = \sum_{i=t}^{t+T} |a_i - b_i|. \tag{6}$$

The proposed method of globalization measurement consist of two steps:

- firstly the time series are converted into entropy time series and
- secondly the cross-correlation among them analysed.

Since the globalization is rather a process then a state, therefore the final outcome should describe the evolution of a system. In order to fulfil these requirements the moving time window technique is used. In fact two different time windows are required. The first one is needed to transform analysed time series into entropy time series and the second window to measure the cross-correlation among them. At each step the time window are moved by one data item (the actual shift length depends on the probing frequency). Finally a series of distant matrices are obtained. The last step is finding characteristic of the distance matrix. In the present study the mean distance of each distance matrix is calculated and presented as final result.

3 Data

The proposed methods was verified on chosen time series of GDP, CPI, employment rate and CO_2 emission. The macroeconomy time series were obtained from the OECD data base, while the CO_2 emission time series were downloaded from the European Commission web page [6]. The most comprehensive database was the GDP which consists of 93 different countries time series from 1961 till 2016. CPI time series were monthly data of 25 countries from January of 1970 till April of 2015. Employment rate covered quarterly data of 26 countries from beginning of 2001 till the end of 2017. The CO_2 emission time series covered dates 1970-2015 for 126 countries. The times series were converted into the relative increment time series i.e.

$$R(t) = \frac{p(t) - p(t-1)}{p(t-1)} \tag{7}$$

4 Results

The obtained results agree with the stated hypothesis. The time series were chosen in such a way that the globalisation processes should be observed: GDP, CPI, employment (particularly among the European countries during European Community formation) and CO_2 emission which are contrasted with the analysis of GDP time series of African, Asian or South America countries where the globalization does not exists. Quite impressive is comparison of the averaged distance between European and non-European countries for employment time series. The averaged Manhattan distance of permutation entropy for European countries is on the level 0.1 while for non-European takes value of 0.35, so three time greater. Besides that within the evolution of one can distinguish the periods 2009 - 2012, when new members entered EC which resulted further unification of employment market and decrease of permutation entropy distance. Similar observation can be made for GDP time series where the highest distances between countries of Australia and Oceania region taking value in the interval 0.85 - 0.95, while for European countries the average distance is about 0.25, so much lower, which can be related to the much higher level of integration among European countries. In this context very interesting is the result obtained for G7 countries (0.55 - 0.6). G7 is the group of the most developed countries, but due to theirs importance they are highly independent so the differences should not come as a surprise. Analogous result was obtained CO_2 emission. The highest distances among entropy time series are observed for North America countries, which is again quite obvious since USA did not ratified the Kioto protocol and Canada (which withdrew from the Kyoto Protocol in 2012). On the other side there is Europe where permutation distance among countries are more then six times lower.

5 Conclusion

Summarising, the study shows that globalization measure might be based on the permutation entropy. This measure was applied to the various sets of macroeconomy time series showing the evolution of the globalization process. The very important feature of the proposed measure its its universality. The same measure can be applied to various aspects of human activity giving the opportunity to measure its quantitatively.

References

1. Barry K. Gills and William Thompson. *Globalization and Global History*. Routledge, November 2012.
2. Janusz Miśkiewicz. Globalization—Entropy unification through the Theil index. *Physica A: Statistical Mechanics and its Applications*, 387(26):6595–6604, 2008.
3. Janusz Miśkiewicz and Marcel Ausloos. Has the world economy reached its globalization limit? *Physica A: Statistical Mechanics and its Applications*, 389(4):797–806, 2010.

4. Christoph Bandt and Bernd Pompe. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.*, 88(17):174102, April 2002.
5. Massimiliano Zanin, Luciano Zunino, Osvaldo A. Rosso, and David Papo. Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review. *Entropy*, 14(8):1553–1577, August 2012.
6. EDGAR - GHG (CO2, CH4, N2o, F-gases) emission time series 1990-2012 per region/country - European Commission.

Estimating macroeconomic uncertainty from surveys – a mixed frequency approach[☆]

Jeffrey Sheen[a], Ben Zhe Wang[a,*]

[a]*Macquarie University*

Abstract

We propose a new method of estimating economic uncertainty, using dispersions of forecasts of a wide range of financial, activity and inflation variables from both household and professional surveys at various frequencies. With a mixed-frequency state-space model, we construct *ex-ante* macroeconomic uncertainty estimates of the one-year ahead expected state of the economy. Impulse responses show uncertainty shocks lead to a contraction in economic activity, and monetary policy expansion reduces uncertainty, implying that endogenous uncertainty is an additional channel for countercyclical monetary policy.

Keywords: Economic uncertainty, survey data, mixed frequency, state-space model
JEL Classification: D80, E66, E50,C81

1. Introduction

This paper aims to provide a new monthly *ex-ante*, real-time measure of aggregate macroeconomic uncertainty that would be useful in forming economic decisions. We do so by providing an estimate based on forecast disagreements from 3 surveys and for 31 variables. We jointly consider dispersions of these variables from surveys of both consumers (the Michigan Survey of Consumers) and professionals (the Survey of Professional Forecasts and the Livingston Survey) arriving at different frequencies. We construct a mixed-frequency estimate of economic uncertainty based on the dispersion of one-year ahead forecasts of economic indicators. Macroeconomic policies, including monetary policy, are shown to have a delayed impact on these uncertainty estimates on the economy. Our forecast-based estimates reflect the delayed impact of past macroeconomic policies, as well as incorporating the evolving uncertainty of the effects of current and future expected economic policies. Therefore endogenous uncertainty may be an important channel to the real economy through which countercyclical

[☆]We thank Efrem Castelnuovo, Adrian Pagan, Mariano Kulish, and the seminar participants at the University of Melbourne and International Association of Applied Econometrics 2017 meeting for helpful comments.

[*]Corresponding author (Tel: +61 2 9850 8500)
Email addresses: jeffrey.sheen@mq.edu.au (Jeffrey Sheen), ben.wang@mq.edu.au (Ben Zhe Wang)

June 13, 2018

monetary policy might operate.

Ideally, the measure of uncertainty is related to the probability of the occurrence of relevant events. At an aggregate level, uncertainty estimates should best be represented by the dispersion of individuals' density forecasts of economic indicators. However, given the difficulty of obtaining such direct estimates consistently, economists are forced to approximate the true level of uncertainty.

We follow an important approach to estimating *ex-ante* uncertainty that uses survey dispersions of forecasts of economic indicators. The dispersion of cross-sectional survey forecasts reflects the level of disagreements among economic agents. An increase in the dispersion indicates agents disagree more with each other when forecasting an economic indicator, which suggests an increase in uncertainty. The advantage of using disagreement in surveys is that the measure is constructed from expectations of the future state of the economy. It is these expectations that should manifest in individuals' economic decisions, and thus the aggregate economy. For example, Bachmann et al. (2013) uses dispersions of firm surveys from both Germany and US as an estimate of economic uncertainty and find increases in this uncertainty reduce employment and production. They recognize a potential issue with forecast dispersions in that they could simply reflect heterogeneous, but certain forecasts. However, utilizing the micro structure of the data, they show that the forecast dispersion is highly correlated with the standard deviation of forecast errors (which are not prone to the problem of heterogenous forecasts). Therefore they conclude that forecast disagreement is a good approximation of the true uncertainty.

Two potential problems with uncertainty measures using survey data are that: (1) they are typically based on one particular survey and, (2) very often rely on one specific economic indicator in the survey, thus making it hard to generalise to the aggregate economy. While there are many available surveys with dispersions of cross-sectional responses, typically the disagreement of one indicator of one survey is selected as the measure of economic uncertainty.

The top panel of Figure 1 illustrates the first problem. It shows the 12 month ahead CPI inflation forecast dispersion from the Michigan Survey of Consumers (MCH), Survey of Professional Forecasters (SPF) and the Livingston Survey (LV). All three measures of uncertainty exhibit similar patterns—that uncertainty was high in the late 1970s to early 1980s, gradually reducing during the 1980s, remaining at a relatively low level in the so-called Great Moderation era in the 1990s and 2000s, and beginning to increase during the 2008 global financial crisis. However, each measure of uncertainty also had substantial idiosyncratic movements, which makes it hard to favour one survey over another.

The second problem in the previous survey-based literature is the practice of choosing one variable to best reflect underlying macroeconomic conditions. For example, Zarnowitz and

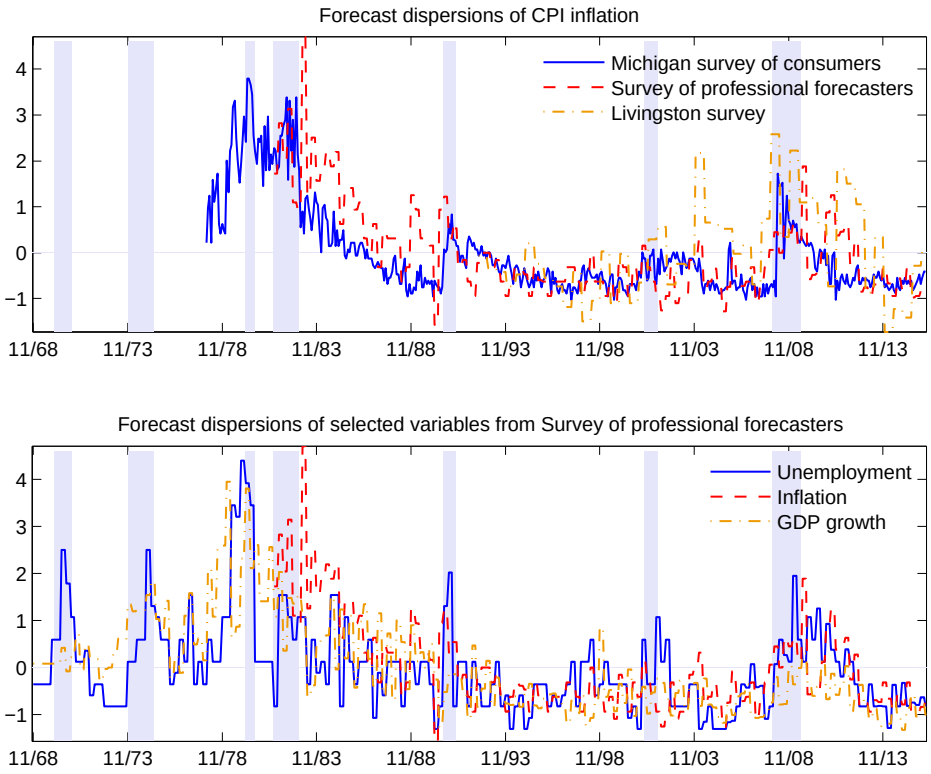


Figure 1: Comparison of uncertainty measures from (standardized) forecast dispersions

Lambros (1987) uses inflation as the underlying economic indicator, Bloom (2014) uses the forecast dispersion of GDP growth rates and Bachmann et al. (2013) utilises a dispersion measure of a qualitative question on general business cycle conditions. Choosing only one variable can be a problem. Consider the Survey of Professional Forecasters—the bottom panel of Figure 1 shows the forecast dispersion for the 12 month ahead forecast of the most important business cycle indicators: the unemployment rate, inflation and GDP growth. All three indicators follow a similar dynamic pattern, however each indicator also exhibits substantial idiosyncratic movement. This makes it hard to choose one indicator over others to represent the estimate of macroeconomic uncertainty.

We contribute to the survey-based uncertainty literature by providing a new measure of economic uncertainty that is free from these two problems. We estimate aggregate macroeconomic uncertainty from dispersions of forecasts across a wide range of economic surveys and economic indicators. Since different surveys release their forecasts at different times for variables reported at different frequencies, we employ a mixed-frequency state-space model to handle this timing issue. Our mixed-frequency approach provides a more efficient way of utilising existing economic information compared with previously published estimates from surveys.

Our paper is also related to a recent literature that approximates economic uncertainty in other ways. Beside using forecast dispersions from surveys to estimate uncertainty as we do here, there are at least three other popular approaches, which are useful but have limitations.

A first popular estimate of uncertainty is the Chicago Board Options Exchange Market Volatility—the VIX index. This gives market expectations of the 30 day ahead volatility implied by S&P500 option prices (for example see Bloom (2009), Caggiano et al. (2014) and Bekaert et al. (2013)). However, the VIX index essentially estimates the volatility of the stock market and may not be a good proxy for aggregate macroeconomic uncertainty.

The second approach estimates *ex-post* uncertainty based on the variance of forecast errors. The idea is that a rise in uncertainty implies economic variables become less predictable. For example, Jurado et al. (2015) estimate macroeconomic uncertainty based on the weighted average of the variance of forecast errors (the unpredictable component) from a large factor model of a wide range of variables. Using the first release of real activity indicators and their market expectations, Scotti (2016) constructs an uncertainty index based on a weighted average of the square of expectation errors of the first release real activity indicators. The uncertainty index is based on forecast errors, thus estimating the ex-post uncertainty of current real activity conditions. However, the weights used in calculating the indices come from the contributions of the indicators in a factor model, and would be sensitive to the data included in the model. In addition, the square of expectation errors against the first release of the data may not necessarily represent genuine uncertainty—a large expectation error could be due to measurement error of this first release. Another contribution to this approach is by Jo and Sekkel (2017) who apply a stochastic volatility factor model to construct an uncertainty index by estimating the time-varying standard deviation of the common component of the forecast errors.

As a third approach, Baker et al. (2016) construct an economic policy uncertainty index based on the text scans on key words related to economic policies and 'uncertainty' from the top 10 American newspapers. However, uncertainty based on the newspaper coverage of media reports may not fully represent the uncertainty that economic agents perceive, even though in a world without information rigidity (for example, see Mankiw and Reis (2002) and Mankiw et al. (2004)), the two might well be significantly correlated.

Lastly, Rossi and Sekhposyan (2015, 2017) propose a measure of *ex-post* uncertainty based on the position of the realised forecast error in the historical forecast error distribution. Since the forecast error can lie on the either side of mean of the distribution, they are able to distinguish 'good' and 'bad' uncertainties.

It should be recognized that using survey disagreement may not be an ideal proxy for evaluating uncertainty. There is a strand of literature evaluating this and the results are mixed.

For example, Zarnowitz and Lambros (1987) and Giordani and Söderlind (2003) provide favourable evidence, while Rich and Tracy (2010), Abel et al. (2016) and Boero et al. (2008) argue that the link between forecast disagreement and uncertainty is weak. Lahiri and Sheng (2010) decompose aggregate forecast uncertainty into disagreement and the perceived variability of future shocks, and argue that disagreement is a good proxy for aggregate uncertainty if the perceived variability of future shocks is small. Although an evaluation of whether disagreement is a good proxy for uncertainty is beyond the scope of this paper, we argue the evaluations of economic uncertainty just mentioned are all *ex-post* measures. To fulfil our objective to provide a monthly, *ex-ante*, real-time measure of macroeconomic uncertainty, we rely on the assumption that survey disagreement is a reasonable proxy for this uncertainty.

There is an increasing consensus on the countercyclical effects of the various uncertainty estimates on business cycle fluctuations. For example, using the VIX index, Bloom (2009) shows positive uncertainty shocks reduce employment and production (though subsequently become expansionary). Similarly, Jurado et al. (2015) find increases in economic uncertainty are contractionary. Caggiano et al. (2014) show that the impact of uncertainty shocks is particularly large for unemployment if one allows for a non-linearity that distinguishes recessions from other episodes. Consistent with this literature, we find that increases in economic uncertainty reduce employment and industrial production. We also find that an unexpected expansionary monetary policy change can reduce the level of uncertainty. These two results have an important implication for monetary policy design—they imply an additional countercyclical transmission mechanism.

The rest of the paper is organised as follows. Section 2 describes the survey data used in the estimation. Section 3 outlines the mixed-frequency state-space framework. Section 4 presents our estimates of aggregate macroeconomic uncertainty and compares them with other popular estimates. Section 5 shows the contractionary impact of positive uncertainty shocks and that unexpected monetary policy can help to reduce the level of uncertainty. Section 6 compares the uncertainty constructed from survey forecasts with uncertainty constructed from survey nowcasts. 7 compares the forecasting performance against the consumer inflation forecast disagreement. Section 8 concludes.

2. The survey data

We consider both surveys on consumers and professional forecasters in this study. The survey on consumers comes from the Michigan Survey of Consumers (MSC). Since January 1978, around 500 U.S. households have been surveyed each month on their one-year ahead and five-year ahead inflation expectations. The data exhibits a considerable degree of disagreement among these households in any given month, even in relatively low and stable inflation periods.

For professional forecasters, we use both the Survey of Professional Forecasters (SPF) and the Livingston Survey (LV), which are both maintained by the Federal Reserve Bank of Philadelphia. The SPF survey is the oldest quarterly survey of macroeconomic forecasts in the US, and contains expectations of a rich set of economic indicators since 1968. These variables include measures of economic activity, inflation, interest rates and spreads in financial markets. The Livingston survey is the oldest continuous biannual survey of economists' expectations, originated by columnist Joseph Livingston in 1946. The survey respondents are economists from industry, government and academia. It covers a wide range of economic indicators including measures of economic activity, inflation, interest rates and a stock price index.

	MSC: Monthly	SPF: Quarterly	LV: Biannually
Financial market indicators			
Level: AAA corporate bond yield	–	✓	–
Level: bank prime loan rate	–	–	✓
Level: 3 month T-bill rate	–	✓	✓
Level: 10 year bond rate	–	✓	✓
Growth: stock price index	–	–	✓
Inflation measures			
Growth: average weekly earnings	–	–	✓
Growth: PPI inflation	–	–	✓
Growth: CPI inflation	✓	✓	✓
Growth: core CPI inflation	–	✓	–
Growth: GDP deflator	–	✓	–
Growth: PCE inflation	–	✓	–
Growth: core PCE inflation	–	✓	–
Activity measures			
Level: unemployment rate	–	✓	✓
Growth: nominal GDP	–	✓	✓
Growth: industrial production	–	✓	✓
Growth: new housing starts	–	✓	✓
Growth: real consumption	–	✓	–
Growth: real non-residential investment	–	✓	✓
Growth: residential investment	–	✓	–
Growth: federal government spending	–	✓	–
Growth: nominal retail sales	–	–	✓
Growth: auto sales	–	–	✓
Month of that data available:	1-12	2,5,8,11	6,12

Note: MSC: Michigan Survey of Consumers; SPF: Survey of Professional Forecasters; LV: Livingston Survey

Table 1: Variables used

Table 1 shows the economic indicators used in each of the surveys. The data can be grouped into three broad categories: measures of financial market indicators, inflation and economic

activity. The only quantitative question from the MSC is about households' forecasts of expected future inflation.

Both SPF and the LV covers economic indicators in all three categories. They both ask questions about the level of returns and interest rates, with the LV focusing more on sovereign debt yields and the SPF focusing more on corporate bonds and the return on shares. Although both SPF and LV ask respondents about CPI inflation, SPF also asks about the GDP deflator and PCE inflation. On the other hand, the LV survey focuses more on the supply side, asking respondents to provide their evaluations of PPI inflation and average weekly earning growth. Both surveys also cover a wide range of economic activity indicators, including the unemployment rate, nominal GDP, industrial production, new housing starts and real investment growth. SPF further asks additional aggregate demand indicators such as consumption, residential investment and federal government spending growth. On economic activity, LV focuses more on retailing, including the growth of retail sales and auto sales.

All three surveys ask participants to provide forecasts at various horizons. To maintain consistency, we only use the 12 month ahead forecast in our benchmark estimate. This is irrespective to the frequency of the actual variable.

The varied timing of the surveys is crucial for the construction of our uncertainty indices, and thus is addressed ideally in a mixed-frequency framework. We use the month of the deadline of surveys to approximate the true information set that a respondent would possess in making forecast decisions. Telephone surveys are conducted every month for the MSC survey, which serves as the base frequency for our model. SPF surveys are sent out at the end of the first month of the indicated quarter, and the respondents are asked to respond by the middle of second month of the indicated quarter.[1] Since the true information set available at the time of making forecasting decisions is bounded by the second month of the quarter, we assume the timing of the survey is in February, May, August and November. Similarly, the biannual LV survey is mailed out in May and November, after the CPI data has been released for the previous month. The FED asks the survey to be returned before the next release of the CPI in June or December. We therefore use June and December as the months that these forecasting decisions are made.

For each of the economic indicators in these surveys, we use the inter-quartile range as the measure of disagreement between forecasters. The inter-quartile range is defined as the difference between the 75th and 25th percentile of forecasts, and it is a more outlier-robust measure of dispersion compared to the standard deviation. For each period (monthly for MSC, quarterly for SPF and biannually for LV), the inter-quartile ranges of current(one-year

[1]For example, the 2010 quarter 1 surveys were sent out by the end of January 2016, the deadline for the response was the third week of February 2016.

ahead) evaluations of economic indicators are calculated from the cross-sectional evaluations of current(one-year ahead) forecasts of economic indicators. These dispersion measures can be accessed from the University of Michigan's and the Philadelphia FED's website.[2] We standardize the dispersions series to have a mean of 0 and a standard deviation of 1, Figure 2 shows these evaluation dispersions.

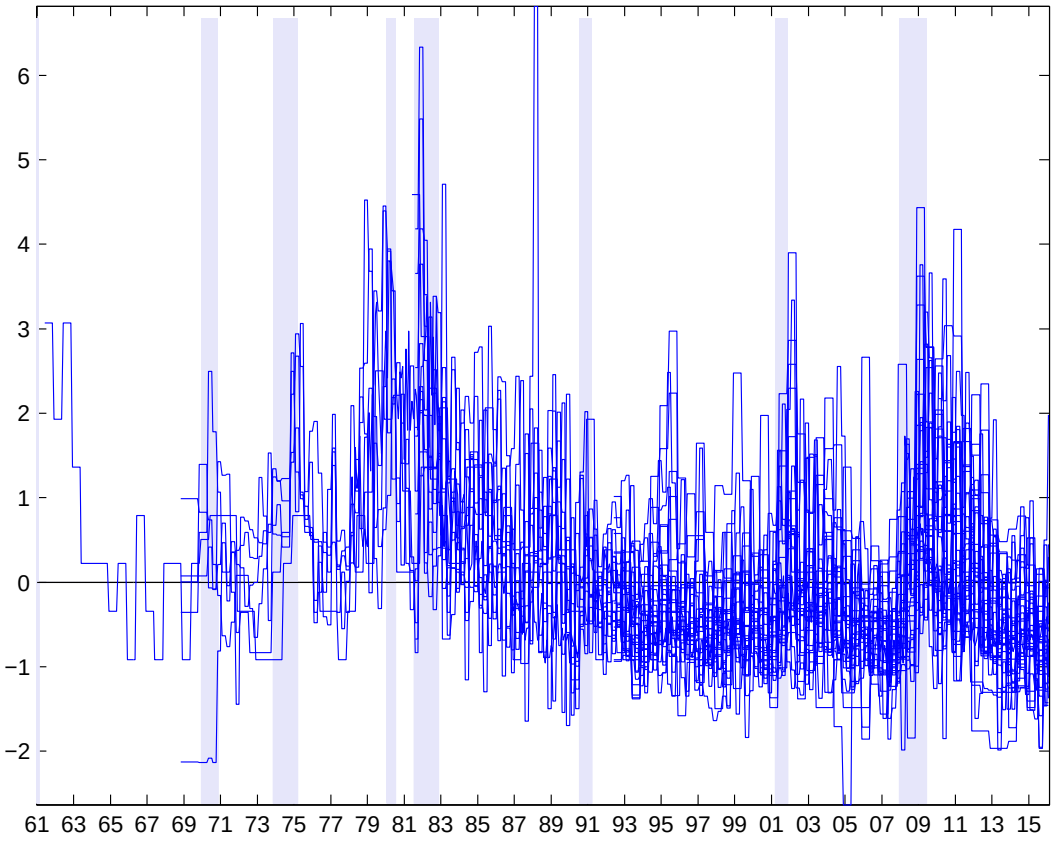


Figure 2: Survey data

In total, there are 31 dispersion series for 12-month ahead forecasts, and the sample period used is from 1961:01 to 2016:02.

Figure 2 shows time-series of forecast dispersions of one-year ahead economic indicators, with the shaded areas indicating NBER recession dates. Although there are considerable

[2], Data can be downloaded for the Michigan Survey of Consumers at https://data.sca.isr.umich.edu/data-archive/mine.php; for the Survey of Professional Forecasters, the data can be accessed at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/historical-data/dispersion-forecasts; for Livingston, the data can be accessed at https://www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey/historical-data.

idiosyncratic movements among dispersions of the different underlying economic indicators, there is also strong co-movements among these dispersions. This is particularly true when there is a recession. However Figure 2 highlights the fact that using the dispersion of only one economic indicator may be misleading in estimating aggregate macroeconomic uncertainty.

3. The econometric framework: mixed frequency state-space model

Since the frequency of MSC, SPF and LV ranges from monthly to biannual, our dataset is mixed-frequency. We accommodate this data structure by using the mixed-frequency state-space model (eg: Aruoba et al. (2009) and Sheen et al. (2015)). Denote u_t as the unobserved uncertainty index for 12 month ahead forecasts. Y_t represents the observed data series on forecast dispersions. Our monthly state-space model has the following form:

$$u_t = \rho u_{t-1} + \epsilon_t \qquad \epsilon_t \sim N(0, P) \tag{1}$$
$$Y_t = \gamma X_t + \beta u_t + \eta_t \qquad \eta_t \sim N(0, Q) \tag{2}$$

where ρ measures the persistence of the uncertainty index u_t, and ϵ_t denotes the innovation of u_t with mean zero and variance P. We allow the observed data to be conditioned by predetermined component X_t and the unobserved state, where γ and β capture the loadings of the respective components. Q is the variance covariance matrix of measurement errors, η_t. We use for X_t the previous observed value of Y_t as a predetermined component.

Denote $u_{t|t-1}$ and $\Sigma_{t|t-1}$ as the model predicted uncertainty index and their associated variance at time t given time $t-1$ information, $u_{t|t}$ and $\Sigma_{t|t}$ as the updated values given time t information, the Kalman filter recursion is given by:

$$u_{t|t-1} = \rho u_{t-1|t-1} \tag{3}$$
$$\Sigma_{t|t-1} = \rho \Sigma_{t-1|t-1} \rho' + P \tag{4}$$
$$u_{t|t} = u_{t|t-1} + K_t v_t \tag{5}$$
$$\Sigma_{t|t} = \Sigma_{t|t-1} - K_t \beta \Sigma'_{t|t-1} \tag{6}$$
$$K_t = \Sigma_{t|t-1} \beta' (Q + \beta \Sigma_{t|t-1} \beta')^{-1} \tag{7}$$
$$v_t = (Y_t - \gamma X_t - \beta u_{t|t-1}) \tag{8}$$

where K_t is referred to as the Kalman gain matrix and v_t is the prediction error. The time t log-likelihood (L_t) of the uncertainty index u_t can be evaluated via the Kalman filter. Denoting the variance of the prediction error (v_t) as $\Psi = Q + \beta \Sigma_{t|t-1} \beta'$, we have:

$$\log L_t = -\frac{1}{2} \left(N \log 2\pi + \log |\Psi| + v_t (\Psi)^{-1} v_t' \right) \tag{9}$$

where N is the number of observations of Y at time t. If not all observations are available at time t, we replace the measurement equation (eq. 2) with:

$$Y_t^* = \gamma^* X_t^* + \beta^* u_t + \eta_t^* \qquad \eta_t^* \sim N(0, Q^*) \tag{10}$$

where $Y_t^* = S \times Y_t$ and S is a selection matrix that contains the value 1 if there is valid data for the corresponding Y_t and 0 if there is missing data. Since all data series are measures of dispersions, we do not need to account for time aggregation in the model. Maximizing the likelihood is equivalent to minimizing the prediction errors, v_t. We first use a simplex method to fine tune the starting values for 20 iterations, then switch to a quasi-Newton method with BFGS updates on the Hessian matrix for the rest of the estimation.[3] We restrict the variance of the state innovations, P, to be the mean of the measurement error variances.

Since the focus of this paper is to estimate an *ex-ante* measure of uncertainty, we only consider the filtered uncertainty estimates in the rest of the paper. Appendix B compares the smoothed estimates with the filtered ones.

4. Estimates of economic uncertainty

Figure 3 shows the filtered uncertainty index. The shaded areas indicate NBER-dated recessions. The horizontal line is drawn at value 1.65—since the indices are standardised with zero mean and unit standard deviation, an uncertainty reading greater than 1.65 indicates that the uncertainty level in that period falls into the zone at least a 5% level of significance.

In general, macroeconomic uncertainty 'jumps' at the beginning of a recession and 'dives' quickly after a recession. This suggests that economic agents disagree widely on the current and future expected course of the economy during economic downturns, but re-develop consensus on the state of the economy when the economy begins to stabilize.

The estimate represents the real-time *ex-ante* uncertainty perceptions of economic agents, since it is based only on observed history. Our estimate of macroeconomic uncertainty indicates four recessions in which uncertainty was significantly elevated (at 5%): the two oil shocks in the mid- and late 1970s, the early 1980s recession and the most recent 2008 global financial crisis. The early 1990s recession shows little effect, and the uncertainty gradually declined from the mid-1980s through to the early 2000s) including during the so-called 'great moderation' era in the 1990s and 2000s, when the economy had stable economic growth with falling unemployment and inflation rates. Our estimate shows the 2008 global financial crisis to be the most uncertain time since the mid-1980s.

[3] Our estimation results are robust to the number of iterations used to fine tune the starting values.

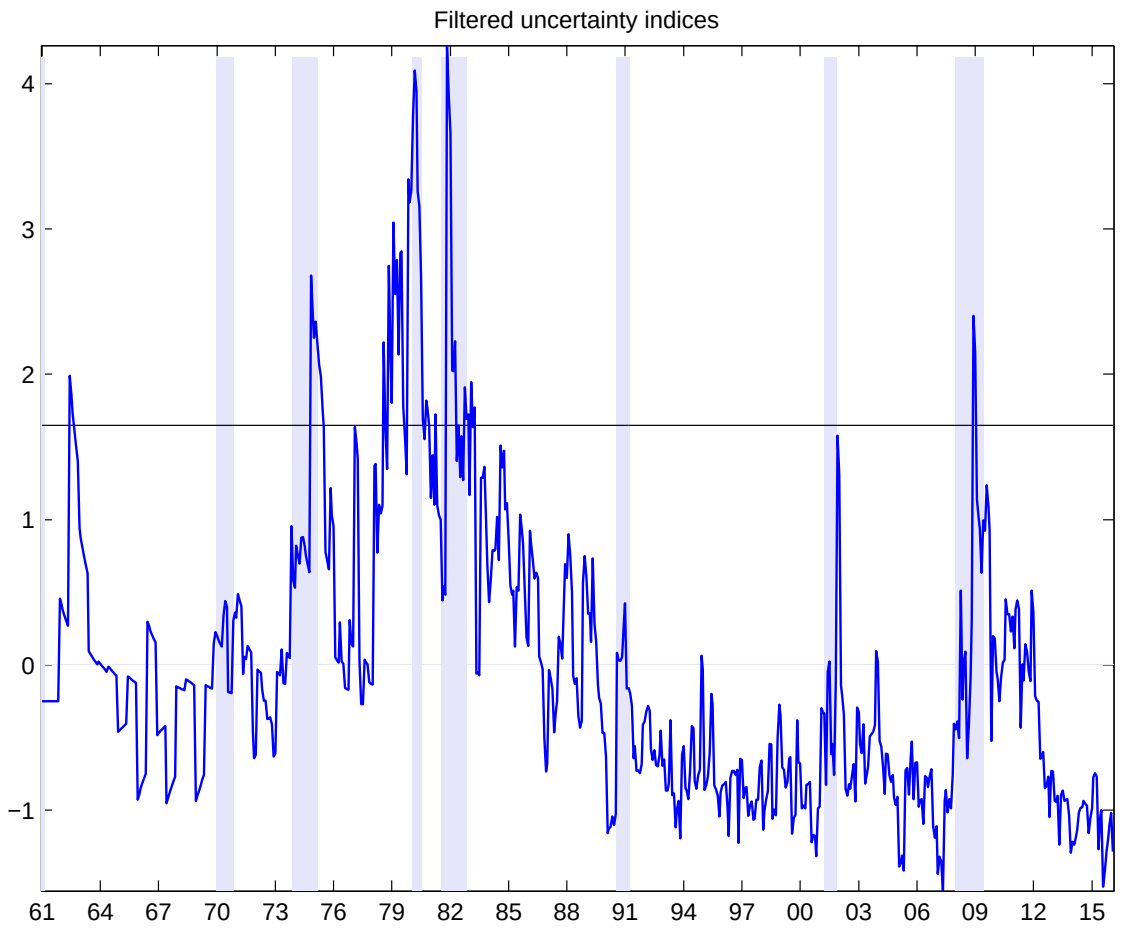


Figure 3: Estimated uncertainty index

4.1. Stylised facts of filtered uncertainty

Table 2 shows the descriptive statistics of the uncertainty index. The lowest reading of uncertainty (-1.56) is in the period between 2000:01 - 2016:02 and the maximum reading of 4.26 in the period between 1980:01-1999:12. Both the median and mean of the index shows a decreasing trend overtime, reflecting a reduction of the central tendency of the level of uncertainty. The standard deviation is at par until 1999:12, but reduces significantly in 2000:01 - 2016:02. Both skewness and kurtosis indicate a right-skewed, heavy right tail uncertainty distribution.

Overall, our estimates of the uncertainty index exhibit the following characteristics. First, the level of economic uncertainty in the U.S. gradually decreased throughout our sample period, possibly due to an improving understanding of the economy and better designed macroeconomic policies. Second, the reduction in the level of uncertainty is accompanied by a sharp reduction in the volatility of uncertainty in the period between 2000:01-2016:02,

	Full Sample	before 1979:12	1980:01 -1999:12	2000:01 -2016:02
Min	-1.56	-0.95	-1.22	-1.56
Max	4.26	3.34	4.26	2.40
Median	-0.17	0.02	-0.24	-0.74
Mean	0.00	0.33	0.11	-0.52
Standard deviation	1.00	0.90	1.13	0.68
Skewness	1.36	1.27	1.37	1.45
Kurtosis	5.13	4.15	4.94	5.58

Table 2: Descriptive statistics of filtered uncertainty

which also corresponds to the period with the lowest uncertainty reading. Third, the distribution of uncertainty is characterised by many small values and fewer larger values (due to positive skewness and kurtosis exceeding 3).

4.2. Understanding the contributions of different variables and surveys

Compared to other uncertainty estimates that have used forecast dispersions in the literature, the advantage of our approach is that we use a wide range of economic indicators and surveys in estimating macroeconomic uncertainty. According to Equation (5), the Kalman filter estimates the unobserved uncertainty based on two components: the model-predicted component $u_{t|t-1}$ and a correction component $(K_t v_t)$ given actual observations, yielding the realization of the prediction error v_t.

The Kalman gain matrix K_t contains the loading information on how much each data series contributes to the estimation of our uncertainty measure.[4] Due to the mixed frequency nature of our dataset, the Kalman filter varies the measurement equation (see Equation (10)) depending on whether there is a data observation or not for a variable. As a result, the dimension of the Kalman gain matrix also varies across time and the Kalman filter only loads on the data when there is a valid data entry.

Since in our case all data series have a positive loading, we can show the relative importance of each data series by calculating the weightings that are associated with each data series. The top panel of Figure 4 shows the weight distribution by type of variables.[5] Weights differ

[4]As explained by Meinhold and Singpurwalla (1983), the Kalman gain matrix K_t behaves like the slope parameter in a linear regression.

[5]Since observations on all variables in the sample do not begin at the same time, the loadings in the Kalman filter change whenever a new variable first becomes available. The latest first entry in our data set is in February 2007 when core CPI, PCE and core PCE inflation became available from the Survey of Professional Forecasters. Thereafter the weightings remain unchanged. The weightings shown in Figure 4 are for a typical month after February 2007.

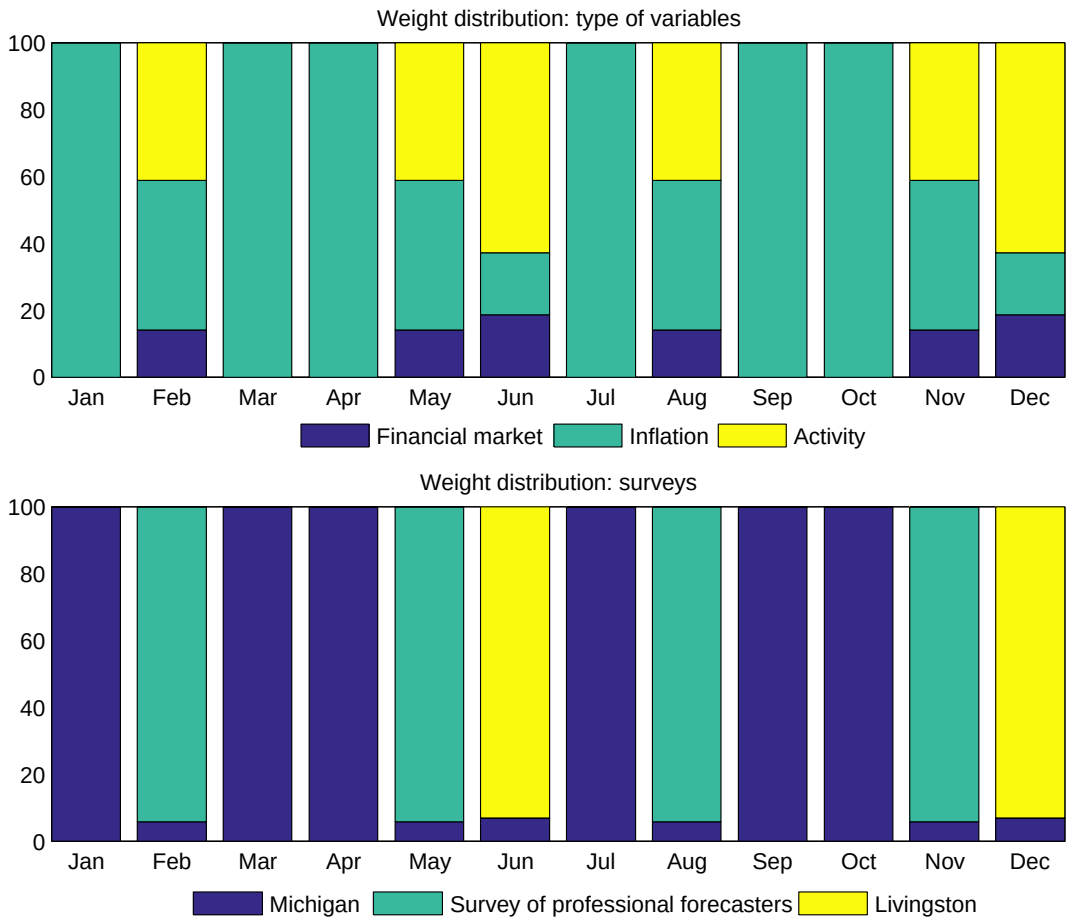


Figure 4: Weighting contributions

across months depending on the availability of variables and surveys in a particular month. When available, financial market variables play a relatively small role (less than 20 per cent) in our uncertainty index. Inflation measures account for all weights in months when only the Michigan Survey of Consumers is available (January, March, April, July, September and October). However the weights on inflation measures are significantly reduced (between 20 and 40 per cent) when financial market and real activity variables are available. Activity variables contribute substantially to the uncertainty index (between 40 and 60 per cent) whenever they are available.

The bottom panel of Figure 4 shows the weighting distributions across surveys. The Michigan Survey accounts for all weights in months when only data from this survey is available. However, this is significantly reduced (to less than 10 per cent) when the Survey of Professional Forecasters or the Livingston Survey is available. Figure 4 demonstrates the advantage of using mixed frequency data: all three surveys and all types of variables make a substan-

tial contribution to our uncertainty index, and our estimation method efficiently utilizes all relevant information for estimating our uncertainty index.

4.3. Comparing other uncertainty estimates

How does our estimate of uncertainty compare with other survey and non-survey based estimates of uncertainty?

The top panel of Figure 5 shows the (normalised) comparison of our uncertainty with survey dispersions of CPI inflation from the Michigan Survey, with CPI inflation, the unemployment rate and GDP growth from the Survey of Professional Forecasters, and with Bachmann et al. (2013) (BES) who use a dispersion measure from business survey data based on a qualitative question on firms' evaluations whether future general business activity will increase, decrease or not change. Correlations are shown in brackets. In general, our estimate is highly correlated with the first four survey dispersions. This is not surprising because our estimate is an optimally weighted average of these survey dispersions, *inter alia*. A distinguishing feature of our approach is that it makes efficient use of any relevant and available information by using multivariate data no matter at what frequency the variables arrive. For the BES estimate, the correlation is small but negative. One reason was that their monthly estimate of the index is based on one survey question about the general outlook of business conditions, so that the estimates are very noisy and non-specific. This highlights an advantage of our highly specified estimate, which utilises the cross-sectional structure of dispersion data across a range of relevant variables and surveys.

There are several non-survey based approaches to economic uncertainty. We focused on four popular estimates in this section. First, volatilities derived from financial markets have long been used to approximate the amount of risk and uncertainty. Among these volatilities, the Chicago Board Options Exchange Market Volatility index, known as the VIX index, has been widely used to approximate uncertainties as perceived in financial markets. This univariate index is constructed based on expectations of 30 day ahead option prices, and measures the implied financial market volatility. Second, Jurado et al. (2015) (JLN) construct an economic uncertainty index based on the idea that economic uncertainty should decrease if more macroeconomic variables are predictable using econometric models. Therefore they construct the index based on the prediction errors of an econometric model in a data rich environment, but restricted to data arriving monthly. Since this uncertainty is based on the predictability of variables, it can be estimated based on particular forecast horizons, we compare below our estimates with their estimates based on one-month ahead and one-year ahead forecasts. Third, using expectation error between market expectation and the first release of real activity indicators, Scotti (2016) constructs an uncertainty index based on the weighted average of the square of the expectation errors. Fourth, using key words such as 'uncertainty' and 'deficit', Baker et al. (2016) (BBD) construct an economic policy uncertainty index based on text scans of 10 leading American newspapers. The idea is that economic policy uncertainty originated in policy decisions and should be reflected in

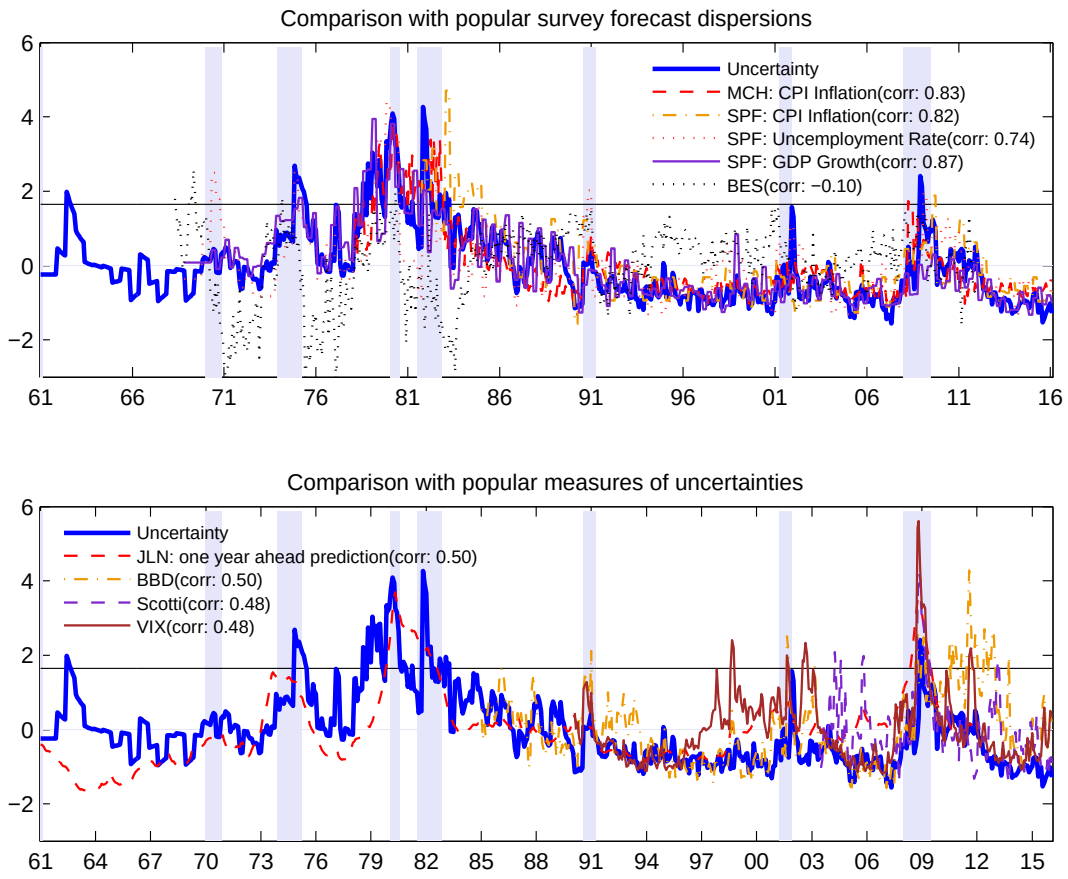


Figure 5: The top panel shows the comparison of our uncertainty index with CPI inflation from Michigan Survey (MCH) and the Survey of Professional Forecasters (SPF); the unemployment rate and GDP growth from SPF; and the business conditions uncertainty index of Bachmann et al. (2013). The bottom panel compares ours with the one-year ahead uncertainty index of Jurado et al. (2015)(JLN), the economic policy uncertainty index of Baker et al. (2016) (BBD), the monthly average of daily uncertainty index of Scotti (2016) , and the VIX index. Correlations with our index are shown in brackets.

newspaper discussions.

The bottom panel of Figure 5 shows the (normalised) comparison of our uncertainty index (blue thick line) with the VIX index (brown thin line), the JLN uncertainty index (red dashed line), the monthly average of the daily Scotti uncertainty index (purple dashed line) and the BBD economic policy uncertainty index (orange dash-dot line). Our index is positively correlated (at around 0.5) with these popular alternatives.

The main difference between the VIX index and our estimate lies in the period between 1996 to 2003, when the VIX shows high and volatile uncertainty movements. This is likely due to the fact that VIX measures implied uncertainty exclusively in financial markets, whereas

our indicator also incorporates general macroeconomic uncertainty.

The movements of our and the JLN estimate track fairly closely. Both indices show an increase in uncertainty during the 1970s oil shocks, and exhibit reducing uncertainty throughout 1980s to 1990s with a jump during the 2008 financial crisis. However, our indicator indicates higher uncertainty during the two oil shocks and lower uncertainty during the global financial crisis compared to JLN. While both estimates utilize a good range of relevant variables, ours better represents the subjective *ex-ante* uncertainty of surveyed agents, while theirs better represents objective, *ex-post* but model-dependent uncertainty.

The Scotti real activity uncertainty index is also more volatile than ours, and this is especially evident during 2003-2006 and after the financial crisis. This may reflect the fact that the Fed kept the interest rate very low, and then implemented quantitative easing, for which the market became more uncertain about its effects on real activity variables. Our estimate differs in that also incorporates financial and inflation indicators, which are likely to have had a smoothing effect.

Compared to the BBD index, ours shows generally lower and more stable uncertainty following the early 1990s recession and the 2008 global financial crisis. This reflects the fact that, while economic policies were widely discussed in the media, there would have been little consensus in acute periods about the appropriate policy response. However households and professional forecasters may have developed a much better consensus on the current and the future course of the economy, even in these acute periods.

5. Uncertainty and economic activity

It has been documented in the empirical literature that uncertainty shocks have adverse impacts on economic activity (eg: Jurado et al. (2015), Bloom (2009), Bachmann et al. (2013) and Caggiano et al. (2014)). These adverse impacts may arise because of frictions in markets. For example, Bernanke (1983) built a model with irreversible investments, so that the optimal timing decision on investment hinges on the trade-off between the potential loss of delaying the investment and the possible gain in waiting so that the outcomes become less uncertain. Increases in uncertainty will delay investment, therefore impeding rises in employment and output. Bloom (2009) built a model with labour and investment adjustment costs and showed the model is capable of driving sharp recessions when uncertainty rises and then subsequent expansions. Leduc and Liu (2016) shows increases in uncertainty effects resemble those of an adverse aggregate demand shock by lowering inflation and raising the unemployment rate. They focus on an option-value channel that arises from search frictions in combination with nominal rigidities that magnify the effects of uncertainty shocks.

In this section we employ two popular VAR models in examining how our filtered economic uncertainty estimate impacts on economic activity. The first VAR model is based on Jurado

et al. (2015), in the spirit of the Christiano et al. (2005) model. The following list contains the variables in the VAR(11) model. Industrial production is an index taken from the FRED database. Employment is measured by the total employees on non-farm payrolls. Real consumption is measured by real personal consumption expenditures. The PCE deflator is the associated chain-type price index for personal consumption expenditures. New orders are approximated by the Institute for Supply Management's new order index. The real wage is taken as the real average hourly earnings of production and non-supervisory employees and hours is the associated average weekly hours of those employees. We use the effective federal funds rate as the policy rate and S&P 500 index as the share market price. M2 is used to approximate money supply. Table A.1 shows the data sources and series IDs. Our monthly sample covers the period between 1965:06 and 2016:02.

$$\text{VAR(11):} \quad \begin{pmatrix} \log(\text{Industrial production}) \\ \log(\text{Employment}) \\ \log(\text{Real consumption}) \\ \log(\text{PCE deflator}) \\ \text{New orders} \\ \log(\text{Real wages}) \\ \log(\text{Hours}) \\ \text{Federal funds rate} \\ \log(\text{SP500}) \\ \text{Growth of M2} \\ \text{Uncertainty} \end{pmatrix}$$

The model is estimated by ordinary least squares and the structural shocks are identified using a Cholesky decomposition. Following Jurado et al. (2015), the ordering of the variables implies that shocks to all variables have an instantaneous impact on uncertainty, but uncertainty shocks do not have an instantaneous impact on all other variables. We call this model VAR(11), and it serves as our benchmark.

The assumption that uncertainty does not contemporaneously impact other variables is debatable. We therefore consider an alternative ordering that considers the most extreme case that uncertainty has an instantaneous impact on all other variables, which have no contemporaneous impacts on uncertainty. Therefore the ordering of the variables places uncertainty on the top, while the ordering of other variables remains the same. We label this model as VAR(11A).

The second VAR model (VAR(8)) is taken from Bloom (2009), and considers 8 variables including the SP500 index, uncertainty, the federal funds rate, wages, CPI, hours, employment and industrial production. The ordering of the variables implies uncertainty is only contemporaneously driven by the stock market, but fluctuations of uncertainty have contemporaneous impacts on all other six variables.

5.1. The impact of uncertainty on economic activities

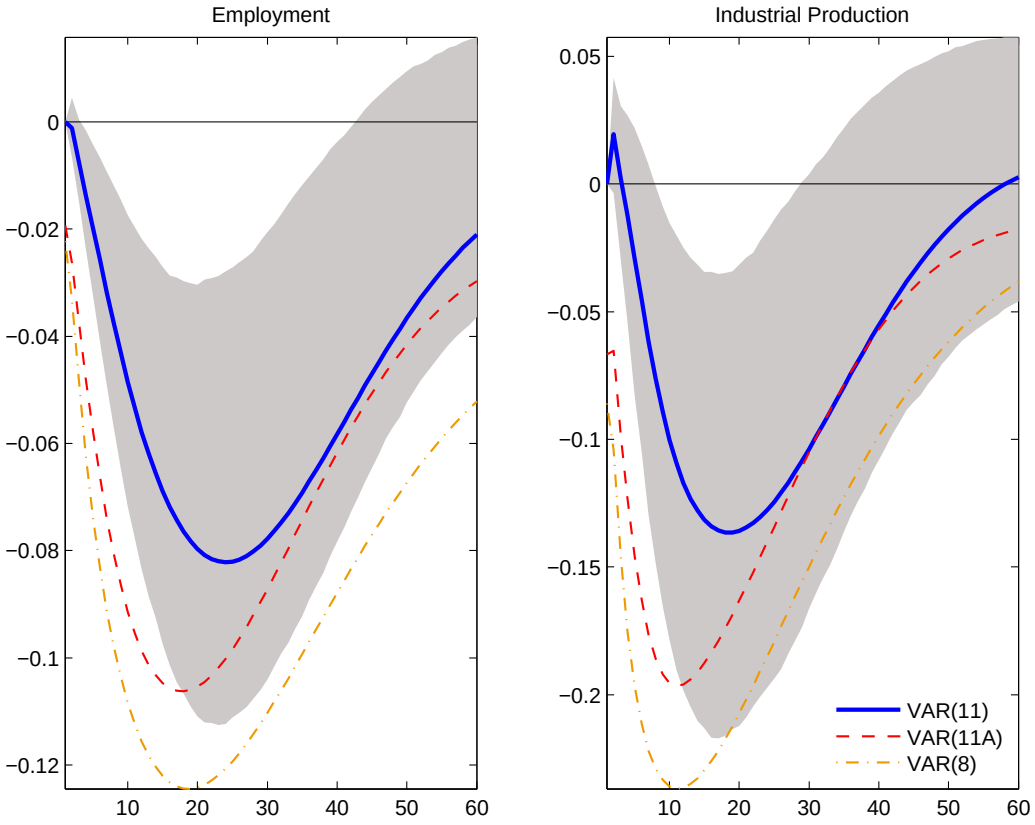


Figure 6: Impact of uncertainty on economic activity. VAR(11) and VAR(11A) are based on Jurado et al. (2015), with VAR(11) ordering uncertainty last and VAR(11A) ordering uncertainty first. VAR(8) is based on Bloom (2009). The shaded area represents the ±1 standard deviation band for our benchmark VAR(11) model.

Figure 6 shows the impulse responses of employment and industrial production to a one standard deviation shock to uncertainty for all three models, with the grey area indicating the ±1 standard deviation band for the benchmark VAR(11) model.

Focusing on the benchmark VAR(11) model. Employment falls following a positive uncertainty shock and industrial production also falls after an initial insignificant increase. The effects reaches their troughs between a year and half and two years following the shock. Comparing VAR(11) and VAR(11A), employment and industrial production respond contemporaneously by assumption in VAR(11A) but with a month lag in VAR(11). The impacts are greater in VAR(11A) compared to VAR(11), however the dynamic patterns are more or less robust to the ordering of the uncertainty estimate. Comparing VAR(11A) with VAR(8),

both employment and industrial production fall quickly initially and then subsequently recovers slowly in the medium run (>24 months).

	VAR(11)		VAR(11A)		VAR(8)	
	EMP	IP	EMP	IP	EMP	IP
$k = 1$	0.00	0.00	1.92	1.34	2.26	1.90
$k = 6$	0.43	0.09	4.13	2.21	5.90	4.24
$k = 12$	1.38	0.64	5.75	3.37	8.39	6.38
$k = 24$	2.69	1.41	6.15	3.43	8.94	6.35
$k = 36$	2.97	1.45	5.42	2.89	7.95	5.25
$k = 60$	2.73	1.20	4.60	2.35	6.85	4.19
$k = 120$	2.42	1.03	4.30	2.04	6.47	3.53

Note: k indicates the forecast horizon upon which the forecast variance decomposition is based. VAR(11) and VAR(11A) are based on Jurado et al. (2015) and VAR(8) is based on Bloom (2009).

Table 3: Variance decomposition

Table 3 shows the forecast variance decomposition of the three VAR models, with k indicating the forecasting horizon. Shocks to our *ex-ante* uncertainty seems to play a moderate role in driving employment and industrial production fluctuations. Depending on the model, unexpected uncertainty shocks account for 2.97–8.94 per cent for employment variance and 1.45–6.38 per cent for industrial production variance. [6]

In summary, positive uncertainty shocks reduce employment and industrial production. Forecast variance decomposition shows that uncertainty matters more for employment than industrial production.

5.2. Does monetary policy reduce uncertainty?

Since uncertainty is important in driving economic activity, it is important to know whether monetary policy can mitigate these uncertainties. Using VIX as the indicator, Bekaert et al. (2013) find that an expansionary monetary policy can in fact reduce uncertainty. This is an

[6] We also conduct the forecast variance decomposition using the smoothed uncertainty index and found uncertainty shocks plays a bigger role compares to *ex-ante* filtered uncertainty. The uncertainty shocks account for 5.1 and 2.4 per cent of employment and industrial production variance respectively within a year. Uncertainty shocks become more important in the medium run (11.9 and 6.8 per cent after 3 years) and remain important in the long run (10.8 and 5.8 per cent respectively at 10 years). The results are available upon request from the authors.

interesting result, since it also indicates that contractionary monetary policy raises uncertainty. Bernanke and Kuttner (2005) provide a justification that could support this result. They find that monetary policy affects stock prices mainly by affecting equity risk premia. For example, expansionary monetary policy raises stock prices by lowering these premia. Easy money reduces the riskiness of stocks by easing interest costs or improving listed firms' balance sheets, and it might decrease investors' risk aversion thus reducing precautionary saving.

While the Bekaert et al. (2013) result is based on the VIX, and thus only involves financial market uncertainty, our uncertainty estimate also includes real activity and inflation variables. We consider now whether their result extends to our wider estimate in the context of our VAR models.

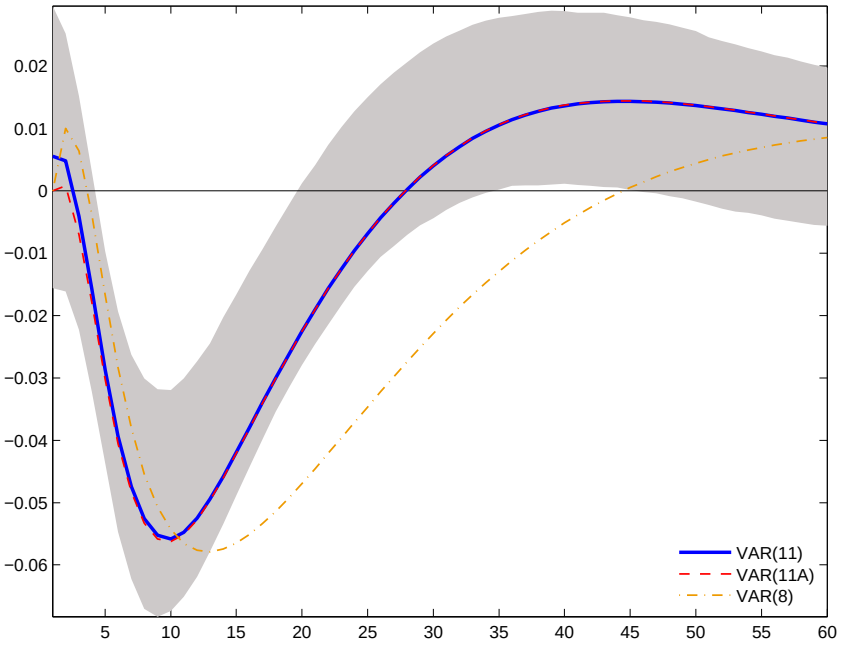


Figure 7: The impact of monetary policy expansion on uncertainty. VAR(11) and VAR(11A) are based on Jurado et al. (2015). VAR(11) orders uncertainty first and VAR(11A) orders uncertainty last. VAR(8) is based on Bloom (2009). The shaded area represents the ±1 standard deviation band for our benchmark VAR(11) model.

Figure 7 shows the impulse responses of uncertainty given a negative one standard deviation monetary policy shock (expansion). The blue line shows the impulse response of uncertainty for the VAR(11) model, the red dashed line shows the response for the VAR(11A) model, and the brown dotted line for the VAR(8) model. Consistent across all models, monetary policy shocks have a delayed impact on uncertainty, reaching a trough around 9 months after the shock under VAR(11) and VAR(11A) and a few months later under VAR(8).

	VAR(11)	VAR(11A)	VAR(8)
$k = 1$	0.85	0.00	0.00
$k = 6$	1.11	0.51	0.64
$k = 12$	6.43	6.42	1.85
$k = 24$	12.72	12.85	8.43
$k = 36$	12.71	12.83	10.82
$k = 60$	12.55	12.69	10.64
$k = 120$	12.69	12.84	11.00

Note: k indicates the forecast horizon upon which the forecast variance decomposition is based. VAR(11) and VAR(11A) are based on Jurado et al. (2015) and VAR(8) is based on Bloom (2009).

Table 4: Variance decomposition

Table 4 shows the forecast variance decomposition of the economic uncertainty estimates with respect to monetary policy shocks. Although the ordering of uncertainty shock involves two extreme assumption about the immediate impact of structural shocks, both VAR(11) and VAR(11A) are remarkably consistent about the impact of a FED rate shock on uncertainty. In the immediate short run, monetary policy is not an important driver of uncertainty, but it gradually becomes important after a year.

In summary, unexpected expansionary monetary policy shocks lower economic uncertainty in the short-run, reaching a trough around one year. Evidence from forecast variance decompositions shows monetary policy is an important driving force of economic uncertainty, especially in the medium- to long-run. Given the established significant impact of uncertainty shocks on economic activity, and the established significant decrease (increase) of perceived uncertainty with monetary policy expansions (contractions), uncertainty is clearly an additional channel through which monetary policy can be transmitted to the real economy.

Thus we get the same interesting empirical conclusion as Bekaert et al. (2013). A rationalization of our result (extending the Bernanke and Kuttner (2005) explanation) might be the following: easier (tighter) monetary policy shocks reduce (raise) equity risk premia, ease (worsen) firms' balance sheets, ease (intensify) bankruptcy risks thereby reducing (raising) economic activity uncertainty, and expand (contract) the economy.

6. Does using nowcast disagreements make a difference?

Apart from asking respondents to provide their 12-month ahead forecasts for economic variables, the SPF and LV also ask them to provide evaluations of the current state of the economy (nowcasts). The disagreements about the current state of the economy are likely to reflect the diverse interpretations of the effects of past economic shocks, past macroeconomic policies and current unexpected economic shocks. On the other hand, the disagreements of one-year ahead forecasts are more likely to reflect the diverse expectations of the future course of the economy and the evaluations of the current and the future expectations of macroeconomic policies. We therefore expect the two to be different, even if highly correlated.

Our nowcasts data come from the listed economic indicators in the SPF and LV (see Table 1). The sample period for nowcast dispersions is bounded by the availability of the SPF survey, which is 1968:11 to 2016:02. Figure 8 compares our nowcast uncertainty with our forecast uncertainty.

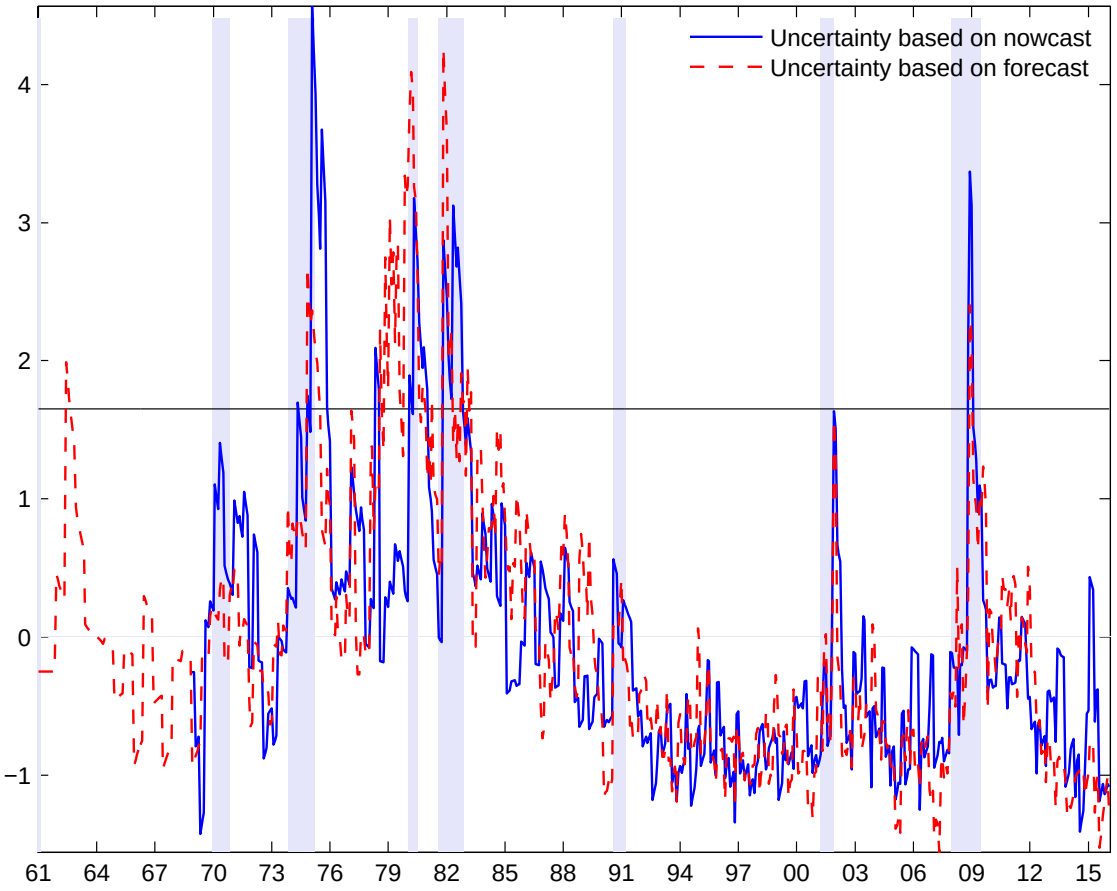


Figure 8: Estimated nowcast and forecast uncertainty indices

Both nowcast and forecast uncertainty indicate the 2008 global financial crisis was the most uncertain time since the mid-1980s. It is interesting to note that they disagree on the level of uncertainty when comparing to the two oil shocks in the 1970s. When evaluating the current state of the economy (nowcasts—blue solid line), economic agents appear equally uncertain about the impact of the unexpected economic shocks of the 2008 crisis as they were in the late 70s, but they seem to be more certain about the future course (one-year ahead forecast—red dashed line) of the economy for the 2008 financial crisis. Since uncertainty based on forecast reflects the delayed impact of current policies and the expectation of the course of future policies, our estimates of uncertainties seem to suggest that economic policies implemented in 2008 and expected after that helped to reduce uncertainty, compared to 1979. In particular, a lower reading of uncertainty based on forecasts compared to the estimate based on nowcasts in 2008 may indicate that improved forward guidance on policies in this period may have helped to reduce the level of uncertainty regarding the future course of the economy.

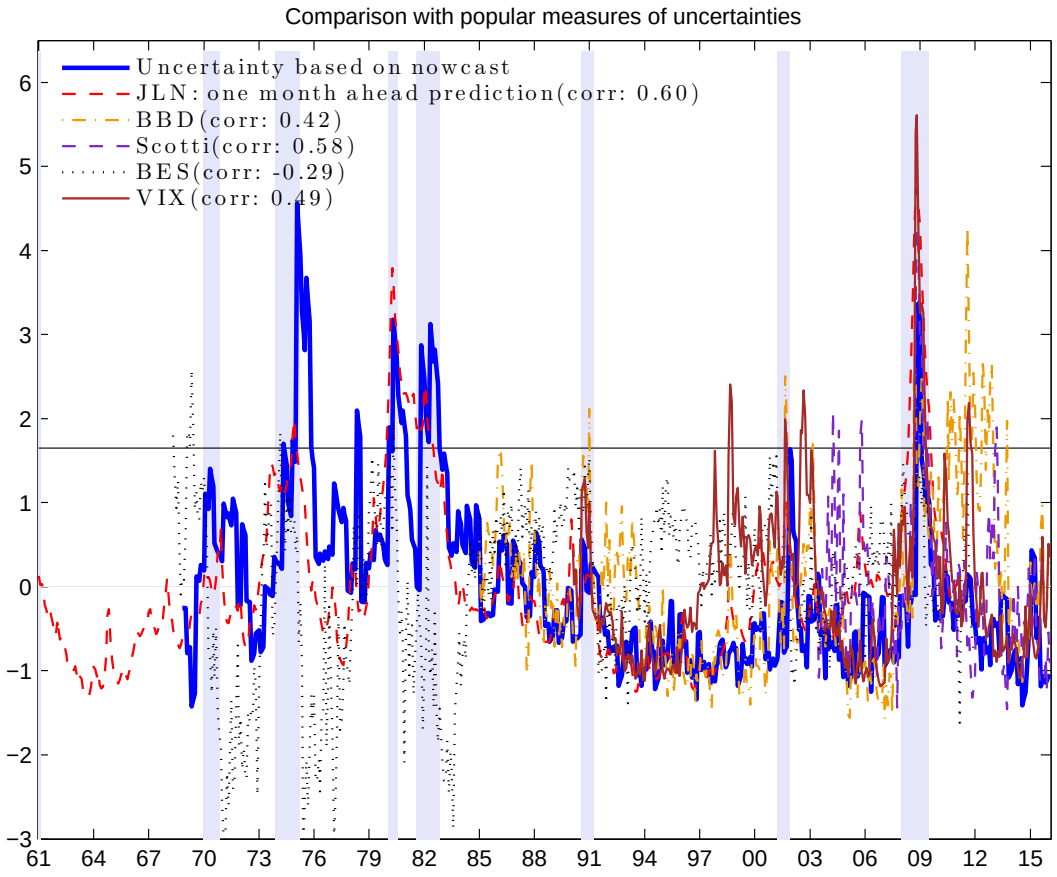


Figure 9: The figure shows the comparison of the uncertainty index based on nowcasts and the one month ahead Jurado et al. (2015)(JLN) uncertainty index, Baker et al. (2016) (BBD) economic policy uncertainty index, Scotti (2016) (Scotti) monthly average of daily uncertainty index, Bachmann et al. (2013) (BES) uncertainty index and the VIX index. Correlations with our index are shown in brackets.

Figure 9 compares our nowcast uncertainty index against other popular existing estimates. Similar to the forecast one, the index is positively correlated with all other popular estimates except the BES index where only one indicator from one survey is used. The correlation with the Scotti index has significantly increased (from 0.48 to 0.58) since the Scotti index is essentially estimating uncertainty based on nowcast errors. The correlation with the one-month ahead prediction variance of JLN also significantly increases from 0.5 to 0.6.

Impulse responses and forecast variance decompositions of the three VAR models in Section 5 and 5.2 using the nowcasts deliver the same qualitative conclusions: uncertainty shocks are contractionary for employment and industrial production, and easier monetary policy reduces the level of uncertainty.[7]

7. Does adding additional surveys help to forecast economic activity?

Theoretically, the state-space model outlined in Section 3 should provide a better estimate of uncertainty than any individual variable disagreement by eliciting the common co-movements (signal) among all 31 variables and removing idiosyncratic noise. Figure 4 shows that all categories of variables across all three surveys indeed contributed to the estimate of our uncertainty index.

How useful are theoretically better estimates for forecasting economic activity? Though a thorough evaluation of forecasting performance is beyond the scope of this paper, we provide a simple comparison of the forecasting performance of economic activity variables (industrial production and employment) using either the estimated uncertainty index or the monthly disagreement of just inflation expectation from the MCH.

For each of the three VAR models, we calculate extending window forecasts for both inflation disagreement and our estimated uncertainty index. We use 10 years of data from January 1978 to January 1988 as the training sample, then for every subsequent month we re-estimate the models and produce forecasts. The out-of-sample forecast errors are calculated for the period between February 1988 and January 2016.[8] Given the economic activity indicator of interest, the VAR model and the forecasting horizon, we compare the forecasting performance by taking the ratio of the mean squared error of the forecasts based on the estimated macroeconomic uncertainty and those based on the inflation forecast disagreement. A score smaller than 1 indicates the forecasting performance using our estimated uncertainty is better than that using the inflation forecast disagreement. The one-sided Diebold and Mariano (1995) test is used to evaluate the statistical significance of the comparison.

[7] These results are available upon request from the corresponding author.
[8] The lag order of the VAR models is selected using the BIC information criteria.

	1-Month	6-Month	12-Month	24-Month	60-Month
			VAR(11)		
Industrial Production	1.03	1.01	0.97**	0.98***	0.96***
Employment	1.15	1.20	1.09	1.01	1.01
			VAR(11A)		
Industrial Production	1.03	1.01	0.97**	0.98***	0.96***
Employment	1.06	1.18	0.97	0.82*	0.83*
			VAR(8)		
Industrial Production	1.04	1.04	0.99*	0.97***	0.90***
Employment	1.18	1.22	1.11	1.03	0.97***

Note: Forecasts are compared using the ratio of the mean squared error—
a number less than one indicates the estimated uncertainty
measure has a better forecasting performance than the
monthly inflation forecast disagreement of the MCH. A one-sided Diebold and Mariano (1995)
test evaluates the statistical significance of the comparison.
*, **, *** represent significance at 10%, 5% and 1% level of significance.

Table 5: Forecasting performance

Table 5 presents the results. Focusing on industrial production, it is interesting that our estimated uncertainty forecasts better in the mid- to long-run ($>$ 12 months ahead). This result is robust across all three VAR models used. Our estimated uncertainty also outperforms the inflation forecast disagreement in forecasting employment in the long-run (60 months ahead) in the VAR(11A) and VAR(8) models, but this is not the case for VAR(11) models.[9] In short, based on the particular models, our estimated uncertainty indicator is almost surely more useful for forecasting industrial production in the mid- to long-run, and employment in the long run (though with less conviction).

8. Conclusions

We have provided a new estimate of *ex-ante* macroeconomic uncertainty in the US using disagreements of forecasts of a wide range of economic variables from both household and professional surveys arriving at various frequencies. Our innovation is to implement a mixed-frequency state-space model to efficiently combine forecast disagreements of a wide range of economic indicators from three widely used surveys. While our estimates of uncertainty are highly correlated with most other existing estimates, there are important differences particularly in acute periods, when such estimates really matter. Our mixed-frequency estimate

[9]Using the inflation forecast disagreement helps to forecast 1 month ahead short-run industrial production and short to medium-run employment for the VAR(11) and VAR(8) model only. The results are available upon request.

efficiently uses information arriving at different frequencies, which distinguishes it from other existing estimates.

Uncertainty clearly matters and government can and should try to achieve significant improvements in the economy by finding ways to reduce disagreements among economic agents. Our uncertainty estimates reflect disagreements and have significant and important impacts on economic activity—in particular, employment and industrial production. If uncertainty in terms of survey disagreement rises by one standard deviation, employment and industrial production fall to a trough of -.08 to -.13% in 1 to 2 years. We also show that monetary policy expansions significantly reduce uncertainty, with a trough after a year of -.05 to -.06%—conversely, tighter monetary policy will worsen uncertainty within a year. Therefore endogenous uncertainty represents an extra channel (beyond the cost and availability of funds) through which counter-cyclical monetary policy operates. A detailed investigation into the details of the size and time variations of this uncertainty channel across business cycles is left for future research.

9. Reference

Abel, J., Rich, R., Song, J., Tracy, J., 2016. The measurement and behavior of uncertainty: evidence from the ECB survey of professional forecasters. Journal of Applied Econometrics 31 (3), 533–550.

Aruoba, S. B., Diebold, F. X., Scotti, C., 2009. Real-time measurement of business conditions. Journal of Business & Economic Statistics 27 (4), 417–427.

Bachmann, R., Elstner, S., Sims, E. R., 2013. Uncertainty and economic activity: Evidence from business survey data. American Economic Journal: Macroeconomics 5 (2), 217–249.

Baker, S. R., Bloom, N., Davis, S. J., 2016. Measuring economic policy uncertainty. Quarterly Journal of Economics 131 (4), 1593–1636.

Bekaert, G., Hoerova, M., Duca, M. L., 2013. Risk, uncertainty and monetary policy. Journal of Monetary Economics 60 (7), 771–788.

Bernanke, B. S., 1983. Irreversibility, uncertainty, and cyclical investment. The Quarterly Journal of Economics 98 (1), 85–106.

Bernanke, B. S., Kuttner, K. N., 2005. What explains the stock market's reaction to federal reserve policy? The Journal of finance 60 (3), 1221–1257.

Bloom, N., 2009. The Impact of Uncertainty Shocks. Econometrica 77 (3), 623–685.

Bloom, N., 2014. Fluctuations in Uncertainty. Journal of Economic Perspectives 28 (2), 153–76.

Boero, G., Smith, J., Wallis, K. F., 2008. Uncertainty and disagreement in economic prediction: the bank of england survey of external forecasters. The Economic Journal 118 (530), 1107–1127.

Caggiano, G., Castelnuovo, E., Groshenny, N., 2014. Uncertainty shocks and unemployment dynamics in us recessions. Journal of Monetary Economics 67, 78–92.

Christiano, L. J., Eichenbaum, M., Evans, C. L., 2005. Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy. Journal of Political Economy 113 (1), 1–45.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 253–263.

Giordani, P., Söderlind, P., 2003. Inflation forecast uncertainty. European Economic Review 47 (6), 1037–1059.

Jo, S., Sekkel, R., 2017. Macroeconomic uncertainty through the lens of professional forecasters. Journal of Business & Economic Statistics, 1–11.

Jurado, K., Ludvigson, S. C., Ng, S., 2015. Measuring uncertainty. The American Economic Review 105 (3), 1177–1216.

Lahiri, K., Sheng, X., 2010. Measuring forecast uncertainty by disagreement: The missing link. Journal of Applied Econometrics 25 (4), 514–538.

Leduc, S., Liu, Z., 2016. Uncertainty shocks are aggregate demand shocks. Journal of Monetary Economics 82, 20 – 35.

Mankiw, N. G., Reis, R., 2002. Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve. Quarterly Journal of Economics 117 (4).

Mankiw, N. G., Reis, R., Wolfers, J., 2004. Disagreement about Inflation Expectations. In: NBER Macroeconomics Annual 2003, Volume 18. NBER Chapters. National Bureau of Economic Research, Inc, pp. 209–270.

Meinhold, R. J., Singpurwalla, N. D., 1983. Understanding the Kalman filter. The American Statistician 37 (2), pp. 123–127.

Rich, R., Tracy, J., 2010. The relationships among expected inflation, disagreement, and uncertainty: evidence from matched point and density forecasts. The Review of Economics and Statistics 92 (1), 200–207.

Rossi, B., Sekhposyan, T., 2015. Macroeconomic uncertainty indices based on nowcast and forecast error distributions. American Economic Review 105 (5), 650–55.

Rossi, B., Sekhposyan, T., 2017. Macroeconomic uncertainty indices for the euro area and its individual member countries. Empirical Economics 53 (1), 41–62.

Scotti, C., 2016. Surprise and uncertainty indexes: Real-time aggregation of real-activity macro-surprises. Journal of Monetary Economics 82, 1–19.

Sheen, J., Trück, S., Wang, B. Z., 2015. Daily business and external condition indices for the Australian economy. Economic Record 91 (S1), 38–53.

Zarnowitz, V., Lambros, L. A., 1987. Consensus and uncertainty in economic prediction. Journal of Political Economy 95 (3), 591–621.

Appendix A. Data

Variable	Source	Series ID
Industrial production	FRED	INDPRO
Employment	FRED	PAYEMS
Real consumption	FRED	DPCERA3M086SBEA
Hours	FRED	AWHNONAG
Wages	FRED	AHETPI
Fed rate	FRED	FEDFUNDS
M2	FRED	M2SL
CPI	FRED	CPIAUCSL
PCE deflator	Datastream	USCP...CE
New orders	Datastream	USNAPMNO
SP500	Datastream	S&PCOMP

Table A.1: Data source for VAR models

Appendix B. Smoothed estimates

If our focus was to accurately estimate economic uncertainty indices using all available information, we should apply the Kalman smoother to the states u_t after the parameters and state are jointly estimated. The mean of the smoothed uncertainty index is given by:

$$U_t = u_{t|T} = u_{t|t} + J_t(u_{t+1|T} - u_{t+1|t}) \tag{B.1}$$

where T is the length of data and $J_t = \Sigma_{t|t}\rho'(\Sigma_{t+1|t})^{-1}$. The key difference between the Kalman filter and the Kalman smoother is that the filter uses information available up until time t to estimate time t uncertainty, whereas the smoother uses all available information(in the whole sample) to estimate time t uncertainty. The filtered estimate represents the real-time *ex-ante* uncertainty perceptions of economic agents, since it is based only on observed history. The smoothed estimate gives the econometrician's perspective since it makes use of the whole span of the data set. Figure B.1 compares the filtered uncertainty index and the smoothed uncertainty index.

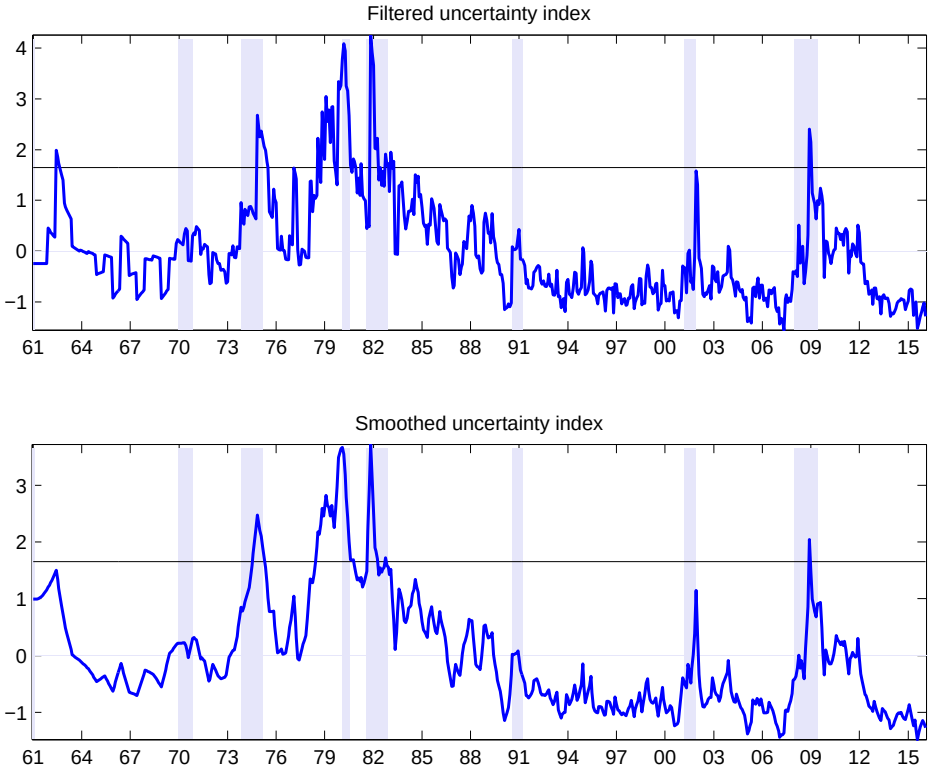


Figure B.1: Estimated uncertainty indices

In most cases the filtered and smoothed uncertainty measures closely track each other. The only major difference is that the filtered uncertainty was significant at the 5% level in the early 1960s but this was not true for smoothed uncertainty.

External Migration as a Factor of Economic Growth: Econometric Analysis for CIS Countries

Kseniya Bondarenko

Analytical Center for the Government of the Russian Federation
Moscow, Russia
xenabondarenko@gmail.com

Abstract. Recent intensification of CIS countries instability has led to an intensification of migration flows in the post-soviet space. By providing multidirectional impact on the various growth factors, migration may both accelerate the countries' economic development and slow it down. The paper analyses the benefits and losses caused by migration processes – both in labor-exporting (the countries of Central Asia, Ukraine) and labor-importing countries (Russia). Using international organizations databases and the instrumental variables regression approach we confirmed the hypothesis that in general for some developing labor-exporting countries high level of labor migration constrains economic growth. Still, in other developing countries with other economic conditions the situation may be quite different. We find the reasons when migration can contribute to economic growth. The calculated threshold value for labor migrant remittances defines the value that inversely changes the impact of migration process on economic growth. We estimated elasticity coefficients of economic growth for certain countries of Central Asia. We defined practical recommendations to increase economic growth inclusivity of these countries by providing special economic measures, such as accelerating development of small industrial businesses and creating large numbers of new sustainable jobs.

Keywords: economic growth, migration, remittances, institutional quality, human capital.

1 Introduction

Labor, along with capital and efficiency, has always been considered [17, 16, 41, 14, 42] as the main factor of economic growth in the framework forecasting models for macroeconomic dynamics. However, the accelerated processes of globalization of the world economy, the increased level of population mobility, the uneven pace of development of various countries and regions in the world push many researchers to make adjustments to the traditional understanding of labor migration influence on economic growth.

Some of these factors, hypothetically, should positively influence economic dynamics while others may negatively affect them. The former can include the influx of labor migrants for countries importing labor resources and the growth of household incomes in the countries that export labor migrants - recipients of remittances. The latter, on the contrary can include low quality (qualification) of incoming labor mi-

grants, shadow economy increase in the countries importing labor and a reduction in the number of employed in labor-exporting countries. This raises the question of who wins and who wins more from migration in the short and long-term: labor-exporting countries or importing countries?

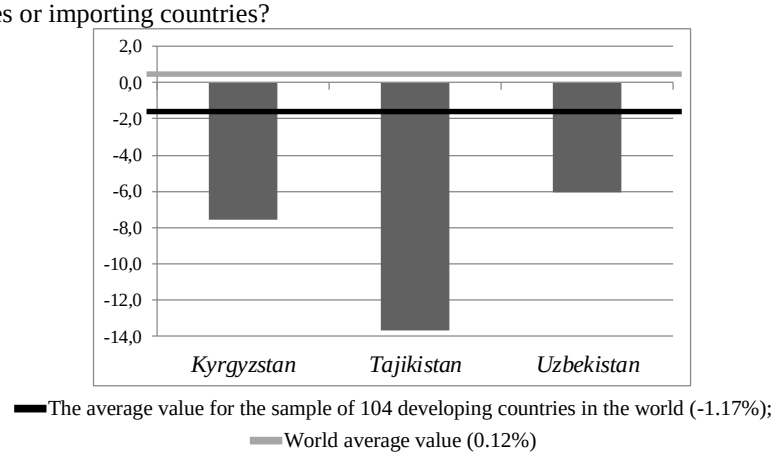


Fig. 1. Indicators of net migration of selected countries in Central Asia in 2012, % to population. Source: calculated by authors on the basis of WB development indicators; FMS of the Russian Federation and statistical committees of the countries of the Central Asian region

This problem has become especially relevant in the post-Soviet area where two poles of migration have emerged in the last decade: Russia as the main importer of labor resources and the other nearby countries where a significant part of the labor resources is mastered by Russian labor market [21]. For some countries, migration is a critical factor in economic growth. It is true in particular for Kyrgyzstan and Tajikistan where remittances of labor migrants are comparable or even exceed the amount of their budgeted revenues. The importance of migration for the countries of Central Asia is also proved by the fact that the magnitude of migratory flows in these countries significantly exceeds the average estimate for the world's developing countries (Figure 1). Due to the growing importance of the migration processes, the quantitative impact of this factor on the economic growth in CIS countries that supply more developed countries with their labor resources is the topic of growing scientific interest.

2 Literature Review

Existing literature on this issue shows that a number of authors [17, 42, 41] agree on the positive impact of migration process on poverty eradication, income growth and other development factors of the world economy. The relationship between poverty and migration is a rather complex economic process. Migration decisions strongly depend on individuals and their living conditions, as well as on existing social conditions. Factors such as age, level of education, financial situation, availability of land ownership and employment opportunities directly affect the participation of individu-

als and households in migration processes, as well as their interrelationships and the presence of social ties (social status and position in a society) [25, 38, 35].

Studies indicate that in many developing countries a large proportion of the money available to households consists of migrants' incomes. Nevertheless, the cash income received does not always contribute to the improvement of the well-being of households, as in some cases the money is used to repay loans and arrears or to spend on other unproductive expenses [35].

The research results confirm that the incomes and the level of consumption of households with some share of income coming from migrants, is generally higher than for similar households that do not have funds received from outside[40]. However, the advantages in economic development accumulated in the course of migration processes are not confined to individual households. Remittances from labor migrants play an important role in improving the welfare of individual countries and their regions [30].

Many researchers believe that international migration of human capital has a strong negative impact both on the overall level of the intellectual potential of the country of origin and on its economic growth [5]. In particular, the migration of skilled labor has a negative impact on human capital in the country, as well as on sustainable economic growth in the country.

The impact of migration processes on the economic growth of some particular country was explored by a number of authors [16, 9, 14]. In particular, Bildirici et al. [9] established that the effect of migration on growth depends on the level of development of the countries concerned and on the fact whether the country belongs to the importer or exporter migrant country. Thus, for the period from 1990 to 2001 for the least developed countries that tend to become migrant exporters, the increasing number of migrants also increases the rate of GDP growth, while for the total sample of 77 countries, including the more developed countries, the GDP growth rate is reduces.

When studying the remittances one can encounter the results of research with both direct and indirect influence of this factor on economic growth in country/economy. This relationship was examined in academic papers [13, 43, 22, 29, 3]. Some researchers have noticed that remittances can increase long-term economic growth in countries with high quality of political and economic institutions [13], in countries of Central and Eastern Europe [22] and in countries with an average income less than $ 1,200 per capita [43]. However, despite the positive impact of remittances on economic growth and poverty reduction, they are essentially sources of external shocks and economic instability [29]. Thus, a sharp increase or decrease in the volume of remittances can have a strong negative effect on economic growth, primarily as a result of a possible weakening of trade sector and a decrease in its external competitiveness.

The description of the qualitative influence factors of migration processes requires additional variables to be added to the equation of economic growth. These variable can include the proportion of migrants with higher education, the proportion of students studying scientific disciplines and developing technology from the total number of students within the country [18]. It was also found that migration of labor resources

with higher education from the country negatively affects the economic growth of this country. Thus, if the country accepts foreign students to study at their universities while the share of such students in the total number of students increases, it also contributes to the development of the country.

Thus, the results of empirical studies indicate the importance of accounting for remittances and migration processes in explaining the dynamics of economic growth. A rapid growth of migration processes, reflecting the process of increasing the uneven economic development of various countries and regions of the world, increases the significance of these factors among other determinants of growth dynamics.

However, the results of the above-mentioned studies are difficult to use to analyze the prospects for economic development of the CIS countries.

First, the statistics used in the studies do not include data from the post-crisis (2010-2014) period, when a new global financial architecture began to form.

Second, the sampling in most of the studies is not homogeneous, as it includes both developed and developing countries, reflecting the patterns of development in countries with already formed and just emerging markets, rich and poor in natural resources of states.

Third, the regression models used are conservative in nature and do not allow establishing the threshold value of migration and qualitative factors. Once this value is excessed, it changes the direction of the impact of individual factors on the dependent variable, which in certain cases is peculiar to the developing countries [37].

This article puts forward and examines the hypothesis that in some developing countries that are labor exporters, a high level of labor migration restrains economic growth, while there are other cases when a different situation occurs.

3 Modelling

Methodological basis to study this problem represents multi-factor regression models. The generalized model we use includes migration variables [9,13, 22] and control variables used in classical models economic growth [4]. In the generalized form, the theoretical model can be represented as follows:

$$growth = a1 + a2\ ln\ (GDPpcl990) + a3\ migr + aX + e, \quad (1)$$

where growth is the average annual increase of GDP (as a percentage of the previous year for the period 1990-2012); a1 is a constant; a2, a - regression coefficients before the dependent variables of the equation; ln(GDPpc1990) - natural logarithm of GDP PPP per capita for 1990; migr - share of net migration as a percentage of total population of the country (average for 1992-2012); X - vector of control variables; e - random error. Please note that vector X includes control variables, i.e. hdi (the average value of the human development index for 1990-2012), urpop (the average value of the share of the urban population of the country for 1990-2012), investments and a number of other elements in traditional models of economic growth.

The model we use is an extended traditional model of economic growth where endogeneity problem is a key issue in econometric modelling [26]. Migration affects

economic growth, but there is a vice versa relationship where economic growth encourages people to migrate abroad or attracts potential migrants to arrive. For example, Lucas [33] notes that an increase in the level of economic development of labor-importing country stimulates the growth of migration there.

As the problem of endogeneity arises, we use instrumental variable approach (IV method) to make consistent estimation of our regression model. This is consistent with the study of Dolado et al. [19] who used IV regressions in migration-growth modelling.

As hypothetical instruments we check the following variables for endogeneity, consistency and over-identifying restrictions: subsidies and benefits to citizens of the country, the level of social welfare guarantees [6], the share of migrants in the host population [10], geographical structure of bilateral migration flows [6, 34], area of labor-exporting country, as far as unemployment and GDP per capita indicators in host countries [5].

We checked instrumental variables for endogeneity and implied Sargan test for testing over-identifying restrictions and Wu-Hausman test to evaluate the consistency of estimators. We found that only three of all instrumental variables mentioned above, i.e. i) social support (net oda received, % GNI), ii) the share of migrants in host population and iii) the area of the exporting country are effective to be used in our regression. While considering the first stage of instrumental variable equation, we also checked the strength of the instruments. The found the coefficients of instrumental variables and their statistical significance at the first stage of the regression as follows:

$$migr = —15{,}82 + \underset{0{,}074*}{0{,}41\ oda} + \underset{0{,}000***}{85{,}06\ mstock} + \underset{0016\ **}{0{,}30ln(land)} + Xe + e, \quad (2)$$

Coefficient of Determination (R-squared) in the first stage of regression is 0.402; the number of observations is 156; F-test = 10.90; where migr is migration; oda – net development assistance; mstock – the ratio of migrants to the population; land –the area of the country; e – regression error; Xe are the variables in the basic model (1) (initial level of development, gross investment, human development index, inflow of foreign direct investment, etc.); p-value in parentheses.

The instruments' coefficients are statistically significant and have expected sings at the first stage of the regression equation, so that they are strong and relevant. Thus, the introduction of the three above-mentioned instrumental variables can be considered as the solution of the endogeneity problem in the model.

The second stage estimates the regression parameters within the theoretical model (1) and showed statistical significance of net migration. The results of the second stage are presented in the next section.

4 Data description

Labor migration has not yet been adequately measured in existing statistical databases. The data on labor migrants and their remittances is abrupt and fragmentary.

The key reasons are as follows: i) lack of a generally accepted definition of labor migration and the difficulties in registering migrants, ii) large-scale illegal migration, iii) the problems of allocating remittances etc.

In these conditions, the only alternative is the indicator of net migration included in the information resource of the World Bank WDI, for which there is comparable statistics, since 1990, for a wide range of countries. This approach was used in most of the studies mentioned above. In addition to the World Bank database WDI 2014 and other international organizations databases, we used information from national migration services and statistical agencies, i.e. Federal State Statistics Service of Russia, The State Committee of the Republic of Uzbekistan on Statistics, Agency on statistics under the President of the Republic of Tajikistan, National Statistical Committee of the Kyrgyz Republic, etc.

We highlight our analysis on developing countries and found the data available for 104 countries. In order to check for the stability of regression coefficients, a wider sample of 158 countries was used, including the countries from the sample of 104 developing countries and 54 developed countries.

5 Results

The results of econometric analysis (see Table 1) showed that most of the key factors of economic growth included in the model (the level of initial economic development, the quality of human capital, gross investment) are statistically significant both for the sample of developing countries (104 countries, model 2), and for the full sample (158 countries, model 1). The signs of the regression coefficients correspond to theoretical basics of economic science.

We confirm the importance of migration for developing countries by narrowing the sample of countries. When accounting for the general sample we receive statistically insignificant coefficient of "net migration" in the model. Still, when we have the sample of developing countries "net migration" becomes positive and statistically significant. The positive value of the coefficient means that in the long run when the inflow of migrants exceeds the outflow (Russia, EU), this factor accelerates the economic growth, while in countries where the outflow of migrants prevails (Tajikistan, Kyrgyzstan, Uzbekistan) slows economic growth down.

The coefficient allows us to estimate the magnitude of accelerating (decelerating) economic growth for the sample of developing countries. Its value of 0.2 represents the change of growth by two percentage point caused by a 10 percentage point change of net labor. The share of predictable variance increases when limiting the sample of countries, i.e. for the general sample it equals to 43% and for the sample of developing countries it reaches 48%.

When including the factor "the sum of international transfers of labor migrants" (model 3) into the model, the share of predictable variance increases further and reaches 52%. The factor is statistically significant, but its negative sign does look illogical. For labor-importing countries (for example, Russia), the outflow of financial assets can to some extent slow down their economic growth. Still, for labor-exporting

countries (including Central Asian countries), the inflow of foreign currency to the households should increase their disposable income, broaden consumer demand and stimulate growth.

Table 1. Results of econometric analysis: the impact of migration on economic growth

Variable description	Variable	Model # (dependent variable - growth)				
		1	2	3	4	5
The initial level of development (GDP in 1990)	GDPpc 1990	-1.244*** (0.216)	-1.356*** (0.300)	-1.493*** (0.280)	-1.555*** (0.284)	-1.661*** (0.291)
Gross investment,%	*invest*	0.0793** (0.0346)	0.0874** (0.0405)	0.0955** (0.0399)	0.0955** (0.0395)	0.0761* (0.0398)
Human Development Index	*hdi*	9.602*** (1.931)	10.09*** (2.231)	10.32*** (2.129)	10.38*** (2.129)	9.774*** (2.044)
Inflow of direct foreign investments,%	*fdi*	0.0711** (0.0284)	0.0495 (0.0341)	0.0485 (0.0317)	0.0535 (0.0333)	0.0588* (0.0308)
Sum of exports and imports,%	*trade*	-0.00399 (0.00308)	-0.00612 (0.00492)	-0.00285 (0.00421)	-0.00247 (0.00415)	-0.00256 (0.00414)
Share of urban population,%	*urpop*	0.00655 (0.00679)	0.00265 (0.00839)	0.00391 (0.00820)	0.00634 (0.00854)	0.00937 (0.00795)
Rule of law	*RoL*					0.638*** (0.229)
Share of net migration in the population, %	*migr*	0.0372 (0.0604)	0.206*** (0.0749)	0.118* (0.0719)	0.173** (0.0795)	0.156** (0.0770)
International transfers, % of GDP	*rem*			-0.0635*** (0.0237)	-0.0859*** (0.0264)	-0.0866*** (0.0266)
The interaction factor (centered)	*mp2*				-0.0110* (0.00651)	-0.0109* (0.00650)
Constant	*_cons*	3.307*** (1.035)	4.213*** (1.435)	4.765*** (1.379)	5.106*** (1.406)	6.801*** (1.595)
R squared, %	*R*	0.43	0.48	0.52	0.53	0.55
Number of countries	*N*	158	104	104	104	104

Source: results of econometric analysis of statistics of relevant indicators for selected countries (World Bank, 2014, UNDP, 2014, World Justice Project, 2014) for 1990-2012. *Note.* In the table, the symbols ***, ** and * indicate the significance ratings at 1% -, 5% - and 10% levels, respectively. Standard errors in parentheses.

One possible explanation here is the inefficient use of these financial resources in labor-exporting countries, as discussed above[1].

[1] In Tajikistan, for example, the average indicator of net migration is -175 thousand people (the value is negative, since the outflow of human capital from the country significantly exceeds the inflow), and the average population - 6451 thousand people for the period 1992-2012. Accordingly, the share of net migration in the population is -2.7%. The contribution of a negative indicator of the share of net migration leads to a slowdown in economic growth. In the example for Tajikistan, this is growth = 0.173 (-2.7) = -0.468%.

Thus, the topic of our interest is not limited just to the volume of remittances the country sends or/and receives; it rather covers the impact of remittances on economic growth. In other words, it makes sense to examine whether there is a threshold value between remittances received and economic growth. We suppose that when in some labor-exporting country the volume of received remittances prevails the threshold value, migration and remittances make positive contribution into that country's economic growth and vice versa.

Such an approach has been already studied by Giuliano & Ruiz-Arranz [39] who examined the impact of financial development on economic growth for 1975-2002. They identified the opposite effects of remittances on economic growth, depending on the level of financial system in developing countries. The threshold estimate allowed dividing the sample of countries into two categories: i) the countries where financial system development exceeded the threshold value and where money transfers stimulated growth, and ii) the countries with a poor financial system, where transfers restrained economic growth.

To test the hypothesis we included into the model (1) the centered factor of interaction *mp2* between remittances and migration indicators [31].

The hypothesis is considered in detail in the work [39]. This study examined the impact of financial development on economic growth. Based on the analysis of the country statistics for the period 1975-2002. The opposite effects of remittances on economic growth have been identified, depending on the level of development of the financial system in relation to the developing countries of the world. The received estimate of the threshold value for remittances allowed to distribute all countries of the sample used into two categories: countries whose financial system development level exceeded a certain threshold value, and, consequently, money transfers stimulated growth, and countries with a weak financial system, where transfers limited the rate of economic growth.

To test the hypothesis with regard to remittances of migrants, the model (1) included the centered factor of interaction of transfer revenues mp2 with the indicator of the migration process [31], which has the following form:

$$mp\ 2 = (migr - migr_{aver})\ (rem - rem_{aver}), \quad (3)$$

where $migra_{ver}$ is the average unweighted estimate of net migration (% population) in the sample of 104 countries (-1.17%); rem_{aver} is the average unweighted estimate of aggregate current transfers in the sample of 104 countries (4.9% GDP).

This approach to a considerable degree allows to overcome the problem of multicollinearity in the regression model. It thus increases the accuracy of parameters estimation and improves the quality of their statistical testing.

As a result of this approach, the original equation (1) is as follows:

$$growth = \alpha 1 + \alpha 2\ ln\ (GDPpc1990) + \alpha 3invest + \alpha 4hdi + \alpha 5\ fdi + \\ + \alpha 6trade + \alpha 7urpop + \alpha 8rem + \alpha 9migr + \alpha 10mp2 + \varepsilon. \quad (4)$$

Based on the regression model above, we can determine the threshold value of migrants' remittances using the following condition:

$$threshold\_value = d\,(growth)\,/d(migr\,) = 0. \quad (5)$$

The derivative when adjusting for (3) and (4) is based on the following:

$$d(growth)\,/\,d(migr) = \alpha 9 + \alpha 10 rem - \alpha 10\ \mathrm{rem}_{aver} = 0. \quad (6)$$

Hence we find the threshold value for remittances:

$$rem=(\alpha 10\ \mathrm{rem}_{aver} - \alpha 9)/\,\alpha 10. \quad (7)$$

Here the sign and magnitude of economic growth elasticity for migration depends on the amount of money received from abroad. If the value of transfers (% GDP) is less than the threshold value, the elasticity (or the gap) will be negative. The larger the gap is, the more migration restrains economic growth of labor-exporting countries.

Substituting the parameters of model 4 (see Table) (–0.011, +4.918, +0.173) into the ratio (7), we obtain rem = 20.6%.

For example, for Tajikistan, where remittances exceed 52% of GDP and thus significantly exceed the threshold value, migration growth by 1 pp (for example, from –13% to –14% of total population, or –1 pp) will lead to GDP growth by the amount

$$\Delta growth=–0{,}011(52{,}0–20{,}65)(–1)=0{,}345\ p.p.$$

Herewith, the obtained estimates of economic growth elasticity for migration will be the higher, the greater is the gap between the current and the threshold level of remittances.

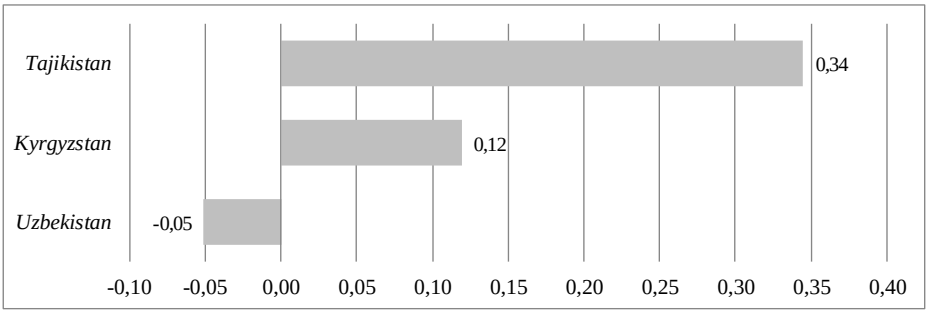


Fig. 2. Changes in GDP per capita (in pp) caused by to the changes in net migration by 1 pp.
Note. calculations are based on the results of econometric analysis

Figure 2 presents the elasticities for three labor-exporting countries in the Central Asian region. For Uzbekistan, where the amount of remittances is less than the threshold value, the growth of migration will reduce the rate of GDP growth, i.e. elasticity estimates will be negative.

For Tajikistan and Kyrgyzstan the estimates have positive signs because the volume of remittances exceeds the threshold value.

This means, that in Uzbekistan migration does not promote economic growth, it rather slows it down. But in the other two labor-exporting countries migration is a sig-

nificant factor of promoting economic growth mostly due to high volumes of remittances received by the households there.

6 Conclusion

The results show that the export of labor, unlike the export of goods and services, cannot be the source of long-term economic growth in developing countries. In some cases transfers can stimulate domestic demand and further expand the growth of production.

These cases refer to unique economic situations which unite two simultaneous conditions: i) the share of imported products in the market of the country is not large, which is not typical for the majority of developing countries, and ii) the amount of transfers exceeds a certain threshold value.

Labor-exporting countries (for example, Central Asian countries) face the challenges of developing and implementing special programs to create new sustainable and decent workplaces that would be able to absorb free labor resources. This will require large-scale investments, imported technologies and the transition to a new industrial policy. One of the most convenient solutions here can be the transfer to industrial and export-oriented small business supported by state programs. The other one is the improvement of institutions quality in a country aimed to improve the effectiveness of economic policy, ensure the guarantees of private property rights, observe the contract discipline and social obligations of the state, and improve the investment climate.

In labor-importing countries (i.e. Russia in our research) migration is positively related to economic growth. Still, these countries also face challenges such as export of remittances outside the country (outflow of capital) which negatively affects balance of payments, sharpening of certain social problems, increased competition in the labor market and decreasing the earnings of native population.

To fix these problems, both labor-exporting and labor-importing countries (especially if they have significant bilateral migrant flows) should constantly work together on the improvement of their common rules and agreements in the field of migration and remittances.

Here we highlight the importance of the results obtained to justify the current situation of migration processes and their prospects in the long-term socio-economic development of countries located in the post-Soviet space.

Further we intend to develop our research by introducing an econometric model based on time series forecasting. We intend to calculate the impact migration can have on economic growth, using different econometric models for different post-Soviet countries.

7 References

1. Orefice G. et al. (2010.) Skilled Migration and Economic Performances: evidence from OECD countries //Schweizerische Zeitschrift fur Volkswirtschaftund Statistik. Vol. 146. No. 4. P. 781.
2. Altonji J. G., Card D. (1991) The effects of immigration on the labor market outcomes of less-skilled natives //Immigration, trade and the labor market. University of Chicago Press, P. 201-234.
3. Barguellil A., ZAIEM M. H., Zmami M. (2013) Remittances, Education and Economic Growth: A Panel Data Analysis //Journal of Business Studies Quarterly. Vol. 4. No. 3.
4. Barro R. J. (1991) Economic growth in a cross section of countries //The quarterly journal of economics. Vol. 106. No. 2. P. 407-443.
5. Beine M., Docquier F., Rapoport H. (2001). Brain drain and economic growth: theory and evidence //Journal of development economics. Vol. 64, No. 1. P. 275-289.
6. Berthélemy J. C., Beuran M., Maurel M. (2009) Aid and Migration: Substitutes or Complements? //World Development. Vol. 37. No. 10. P. 1589-1599.
7. Bertocchi G., Strozzi C. (2008) International migration and the role of institutions //Public Choice. Vol. 137. No. 1-2. P. 81-102.
8. Betts A. (ed.). (2011) Global migration governance. Oxford : Oxford University Press,
9. Bildirici M. et al. (2005) Determinants of human capital theory, growth and brain drain; an econometric analysis for 77 countries //Applied Econometrics and International Development. Vol. 5. No. 2.
10. Card D. (2001) "Immigrants Ináows, Native Outáows, and the Local Labor Market Impacts of Higher Immigration", Journal of Labor Economics, vol. 19 n.1, pp.22
11. Card D. (2005) Is the new immigration really so bad? The economic Journal, 115.
12. Card D., Lewis E. G. (2007) The diffusion of Mexican immigrants during the 1990s: Explanations and impacts // University of Chicago Press, P. 193-228.
13. Catrinescu N. et al. (2009) Remittances, institutions, and economic growth //World Development. Vol. 37. No. 1. P. 81-92.
14. Chen and Fang (2013). Migration, social security, and economic growth// Economic Modelling. Vol. 32. p.386–399
15. Cortes P. (2008) The effect of low-skilled immigration on US prices: evidence from CPI data //Journal of political Economy. Vol. 116. No. 3. P. 381-422.
16. Di Maria C., Lazarova E. A. (2012) Migration, human capital formation, and growth: An empirical investigation //World Development. Vol. 40. No. 5. P. 938-955.
17. Díaz-Briquets S.et al. (1997). Refugee remittances: Conceptual issues and the Cuban and Nicaraguan experiences//International Migration Review. Vol. 31, No 2. P.411-437.
18. Di Maria C., Stryszowski P. (2009) Migration, human capital accumulation and economic development //Journal of Development Economics. Vol. 90. No. 2.P. 306-313.
19. Dolado, J., A. Goria, and A. Ichino (1993), "Immigration, Human Capital and Growth in the Host Country: Evidence from Pooled Country Data," Journal of Population Economics, Vol. 7. No 2. P. 193–215.
20. Dustmann C. (2003) Return migration, wage differentials, and the optimal migration duration //European Economic Review. Vol. 47. No 2. P. 353-369.
21. Gontmakher E., (2013). Global migration processes: the need of global regulation // Voprosy ekononiki No. 10. P. 136-146.
22. Goschin Z. (2014) Remittances as an Economic Development Factor. Empirical Evidence from the CEE Countries //Procedia Economics and Finance. Vol. 10. P. 54-60.

23. Groizard J. L., Llull J. (2007) Brain drain and Human Capital Formation in Developing Countries. Are there Really Winners?. Universitat de les Illes Balears, No. 28.
24. Groizard J. L., Llull J. (2006) Skilled migration and growth. Testing brain drain and brain gain theories. Universitat de les Illes Balears, Departament d'Economía Aplicada, No.20
25. Haberfeld Y., Menaria R.K., Sahoo B.B., Vyas R.N. (1999), "Seasonal Migration of Rural Labour in India", Population Research and Policy Review, Vol. 18, P. 473-489.
26. Durlauf S. N., Johnson P. A., Temple J. R. W. (2005) Growth econometrics //Handbook of economic growth. Vol. 1. P. 555-677.
27. Hsiao C. (2014) Analysis of panel data. Cambridge university press, Vol. 54.
28. Hunt, J. (1992). The Impact of the 1962 Repatriates from Algeria on the French Labor Market. Industrial and Labor Relations Review 45 (3), 556-572.
29. Imai K. S. et al. Remittances, growth and poverty: New evidence from Asian countries //Journal of Policy Modeling. 2014. Vol. 36. No. 3. P. 524-538.
30. Hoddinott J. (1994) A model of migration and remittances applied to Western Kenya //Oxford Economic Papers. P. 459-476.
31. Kraemer H., Blasey C. (2004). Centring in regression analyses: a strategy to prevent errors in statistical inference //International journal of methods in psychiatric research. Vol. 13, No. 3.P. 141-151.
32. Lin, J. (2012). New structural economics: A framework for rethinking development and policy. World Bank Publications.
33. Lucas R. E. B. (1997) Internal migration in developing countries //Handbook of population and family economics. Vol. 1. P. 721-798.
34. Mayda A. M. (2008) Why are people more pro-trade than pro-migration? //Economics Letters. Vol. 101. No. 3. P. 160-163.
35. Mosse D., et al. (2002), "Brokered livelihoods: debt, labour migration and development in western India", Journal of Development Studies, 38(5), 59-88.
36. Orefice G. et al. (2010) Skilled Migration and Economic Performances: evidence from OECD countries // SZVS. Vol. 146. No. 4. P. 781.
37. PolterovichV., Popov V. (2007). Democratization and Economic Growth // Obshestvennienauki i sovremennost'. No. 2. P. 13-27.
38. Rogaly B. et al. (2001) Seasonal migration, social change and migrants' rights: Lessons from West Bengal //Economic and Political Weekly. P. 4547-4559.
39. Giuliano P., Ruiz-Arranz M. (2009) Remittances, financial development, and growth //Journal of Development Economics. Vol. 90. No. 1. P. 144-152.
40. Samal C. (2006). Remittances and Sustainable Livelihoods in Semi-Arid Areas. Asia Pacific Development Journal 13(2), 73-92.
41. Sana M., Massey D. (2005). Household Composition, Family Migration, and Community Context: Migrant Remittances in Four Countries//Social Science Quarterly. Vol. 86, No. 2.
42. Woodruff C., Zenteno R. (2001). Remittances and microenterprises in Mexico //UCSD, Graduate School of International Relations and Pacific Studies WP.
43. Ziesemer T. H. W. (2012) Worker remittances, migration, accumulation and growth in poor developing countries: Survey and analysis of direct and indirect effects //Economic Modelling. Vol. 29. No. 2. P. 103-118.

Business Cycle Synchronizaiton: The effects of cycle extraction methods.

Kanya Paramaguru*

September 4, 2018

Abstract

Many papers investigate the macroeconomic determinants of business cycle synchronization, often investigating the role of trade. However, the cycle extraction methods used in these studies may play a role in determining the results. This study seeks to understand the effect that the choice of business cycle extraction methods can have on the perceived synchronization between two countries. I compare the benchmark method that is used which is the Hodrick-Prescott Filter, to an approach that is rarely used in these kinds of studies. This is a structural decomposition approach with a fourier transformation built in. When comparing these methods, the total business cycle synchronization within the EU seems overstated in the existing literature.

*Surname1: affiliation1, address1, email1. Surname2: affiliation2, address2, email2.
Acknowledgements: I would like to thank

1 Introduction

2 Hodrick Prescott Filter

The main contribution of this study is to investigate whether a bias exists in the output synchronization literature owing to the use of the certain cyclical extraction methods. Understanding business cycles and their fluctuations is important in achieving stable growth. The debate about how to extract cycles has existed for many years. In recent years, the main methods used in macroeconomics for cycle extraction has been the Hodrick Prescott filter(Baxter & King 1999) . (HP filter from here onwards).

The HPfilter works by penalizing the cyclical component of a time series and then makes the assumption that everything that remains forms the trend component of a time series.

$$\min_{\tau}(\sum_{t=1}^{T}(y_t-\tau_t)^2+\lambda\sum_{t=2}^{T-1}[(\tau_{t+1}-\tau_t)-(\tau_t-\tau_{t-1})]^2)$$

The first term of the equation is the sum of the squared deviations $d_t = y_t - \tau_t$ which penalizes the cyclical component. Let $y_t \, fort = 1, 2, ..., T$ denote the logarithms of a time series variable. The series $y_t\$$, is made up of a trend component $\tau_t\tau_t$, a cyclical component c_t, and an error component ϵ_t such that $y_t \ = \tau_t \ + c_t \ + \epsilon_t$ .Given an adequately chosen, positive value of λ , there is a trend component that will solve The second term is a multiple λ of the sum of the squares of the trend component's second differences. This second term penalizes variations in the growth rate of the trend component. The larger the value of λ, the higher is the penalty(Kim 2004).

Whilst a useful technique in econometrics time-series, it has some limitations many studies have pointed out it's limitations(King & Rebelo 1993). Rebelo and king (1993) argue that the HP Filter excludes components from the cycle that are actually an important part of the cycle and that this can have real impacts on subsequent policy analysis. The cyclical component is estimated by the specification of λ One is that the Hodrick Prescott filter

is that it does not adequately account for shocks in the time-series and these shocks are often interpreted as part of the underlying trend component.

The underlying data generating process of the HP filter is subject to being spurious.as if $\lambda = 0$ there is no cyclical component. As a result, the one sided HP filter suffers from an end point problem. That is that the choice of the end affects the whole of the smoothing process.This may particularly be a problem for time series that end in the first few periods on the recession.

I use the Hodrick Prescott filter to extract the real business cycles from logged GDP series of the EU. The first step is to apply the filter to log seasonally adjusted GDP series. I use the model as specified in equation X. For the lambda which is the cyclical penalizing parameter, I use the value of 1600 which is the standard in the literature for quarterly data

I calculate the R squared of the cyclical components for all the unique bi-lateral pairs of the EU and display in the histogram in Fig 3. The median value of correlation across the EU is 0.66. Furthermore there appears to be no countries that exhibit counter-cyclical synchronization.

3 Structural Time Series with Fourier Expansion

In this next section I introduce the use of structural time series decomposition to extract real business cycles. One of the main benefits of UCM is that the series and trend is updated with new information.

Structural time series models decompose a time series in to three components (Harvey 1990). A trend component, a seasonal component and an irregular component: The trend component is locally estimated by the following equation:

$$y_t = \mu_t + \gamma_t + \varepsilon_t$$
$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t$$
$$\beta_t = \beta_{t-1} + \zeta_t$$
$$t = 1.....T$$

Where y_t is the series that is being decomposed. μ_t is the trend component. β is the slope. γ_t is the seasonal component. η_t and ζ are assumed to be serially and mutually uncorrelated with zero mean and variance σ^2 $(\eta_t, \zeta_t \sim NID(0, \sigma^2_{\eta\zeta}))$

The seasonal component is determined as follows:

$$\gamma_t = \sum_{s-1}^{j=1} \gamma_{t-1} + \omega_t$$

The cyclical component is estimated with a series of sinusoidal functions :

$$\begin{bmatrix} \varphi_t \\ \varphi_t^* \end{bmatrix} = \begin{bmatrix} cos\lambda_c & sin\lambda_c \\ -sin\lambda_c & cos\lambda_c \end{bmatrix} \begin{bmatrix} \varphi_{t-1} \\ \varphi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}$$

In this study , I run three forms of the baseline decomposition with various restrictions imposed. I follow the model structure as outlined in (Macchiarelli,2013). Macchiarelli (2013) uses timeseries of GDP and inflation for CEE countries to see if the dynamics are related. The study runs nine different models on the data and uses the AIC to see which model has the best fit. All models start with the basic form of a trend cycle decomposition and then various restrictions on the variance of the level and slope component are added to see if they provide a better model fit. The cyclical components are estimated using an arma 2 process.

- Model 1 is the structural decomposition with the variance on the level fixed at zero but the variance on the slope remains determined by the model.
- Model 2 is the structural decomposition with the variance on the slope fixed at zero but the variance on the level remains determined by the model.
- Model 3 is the structural decomposition with the variance on the level and slope both fixed at zero.

  Next I follow the specification as outlined in (Macchiarelli,2013).The next group of models work by replacing the stochastic equation for the trend with a data with a fourier expansion process. The stochastic

trend in the earlier model is replaced with a more general specification that includes a fourier transform. This is the fourier approximation allows us to represent a cycle as a series of sinusoidal functions. I run three further models on the raw data but this time I incorporate a first order fourier expansion. By incorporating this flexible functional for to determine the trend, we allow for the possibility that there are multiple peaks and troughs in the time series. This is a more realistic determination of the real business cycle.

$$x_{i,j} = \sum_{h=0}^{2} \delta_{i,h} t^h + \sum_{k=1}^{n} \alpha_{i,k} sin(\frac{2\pi kt}{T}) + \sum_{k=1}^{n} \beta_{i,k} cos(\frac{2\pi kt}{T})$$

Where k is the order of the expansion. Where $n < \frac{T}{2}$ and n refers to the number of frequencies contained in the approximation and $t = 1, ..., T$ is a linear trend

- Model 4 is a first order fourier approximation with a sine and cosine expansion.
- Model 5 is a first order fourier approximation with a time trend and sine and cosine expansion
- Model 6 is is a first order fourier approximation with a quadtratic time trend and sine and cosine expansion. The final three models that are run on the raw data are a pure second order fourier expansion.
- Model 7 is a second order fourier approximation a sine and cosine expansion.
- Model 8 is a second order fourier approximation with a time trend and sine and cosine expansion
- Model 9 is is a second order fourier approximation with a quadtratic time trend and sine and cosine expansion.

Once models have been run, I extract the cyclical component of each model on a country basis. Based on the Akaike information criterion, I choose the model which fits the data the best. The nine models are run on quarterly

pre-seasonally adjusted logged gdp for all 28 countries. The data runs from 2000q1-2017q4.

Although the model specification allows for a seasonal component to be extracted, I use pre-seasonally adjusted data as seasonal holidays and working day adjustments are made by the respective statistical agencies of the member states.

The first order fourier expansion without a time trend provides the best fit for nearly all the countries in the sample (Model 4). However, this difference between Model 3 and Model 4 is marginal as the AIC remains around the same level for both models. This results is different to previous papers(Macchiarelli 2013) who finds that a variance of model fits best for the 10 countries in his sample. This could be because of the time sample that I have used which includes the crisis and post crisis periods. There is also the possibility that the filtering methods might be sensitive to the frequency of data and therefore the use of quarterly data may change what is the best model selected. The second order fourier expansions do not perform as well as the rest of the models and the model with the weakest is the second order fourier expansion with a quadratic time trend. This is probably intuitive because business cycles do not display quite as erratic behaviour.

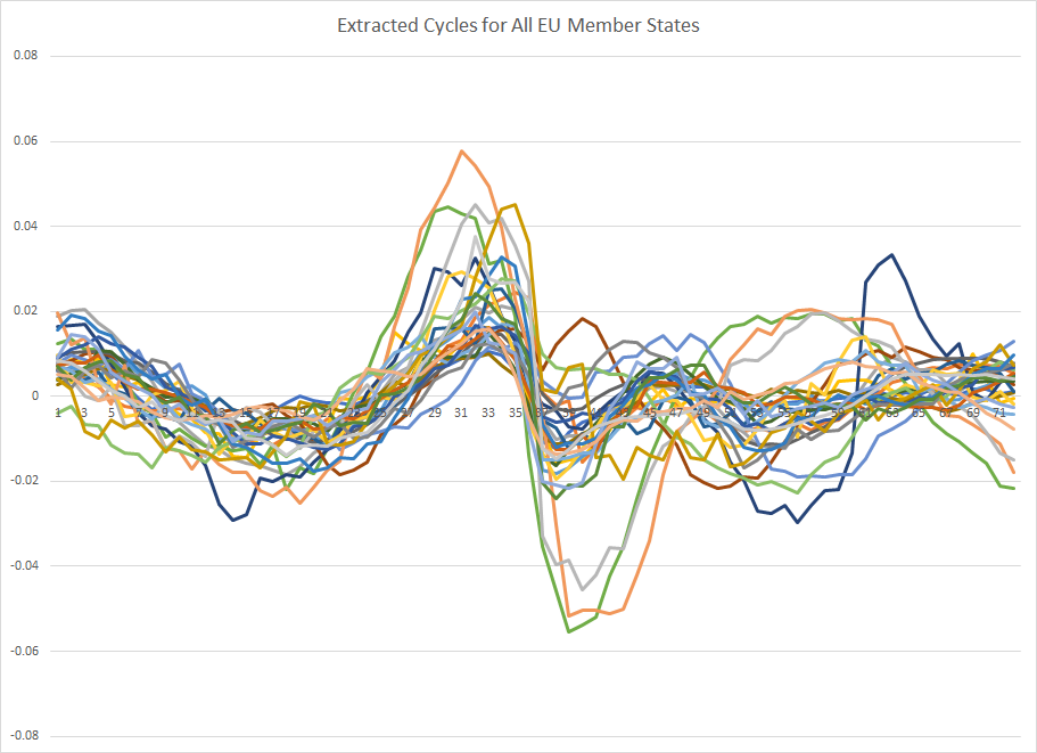


Figure 1:

Figure 1 shows the extracted cycles for all 28 member states. Although it is hard to really see individual country behaviour, there are a few things that really stand out. The synchronicity that is attained during the downturn of the cycle is made clear however their seems to be a lot more variance in the post crisis period than the pre-crisis period. Ireland stands out in this graphs as having a better than average post-crisis recovery. Once I have chosen the correct cycles for each of the countries, I then obtain a pure r correlation for the cycles on a bilateral basis for all the country pairs.

The majority of the country pairs exhibit pro-cyclical business cycle synchronization. Whilst there are a handful who exhibit counter-cyclical behaviour. The median value is a synchronization which is above half. Which on a global scale shows that EU countries are already pretty synchronized. The median value for the rsquared of business cycles in 0.59. Countries

within the EU appear to already have some synchronicity(Dées & Zorell 2012). When looking at an EU only sample, the variation in the dependant variable is much more limited compared to when observing a global data-set. Factors of geography are likely to promote synchronicity in the kind of exogenous shocks that these countries face. However, the positive side is that purely EU dynamics are captured which reduces the risk of outliers affecting the result. Also, there is a wider choice of data available for use in the estimation.

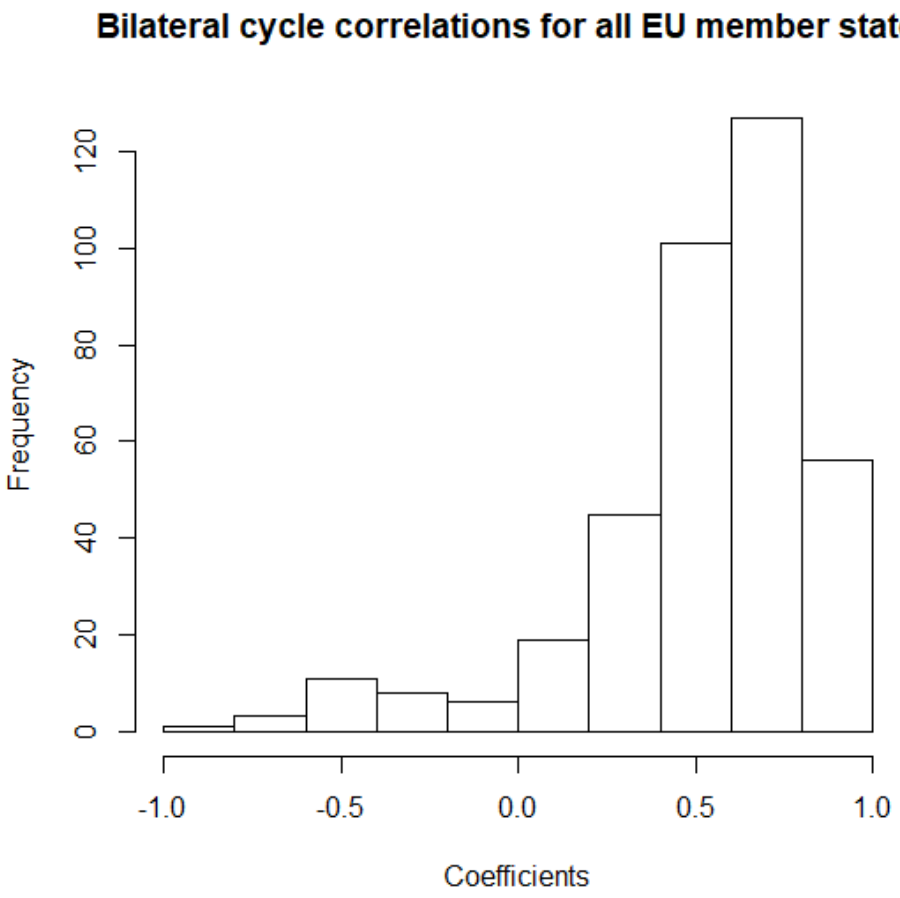

Bilateral cycle correlations for all EU member states
Frequency
0
20
40
60
80
100
120
-1.0
-0.5
0.0
0.5
1.0
Coefficients

Figure 2:

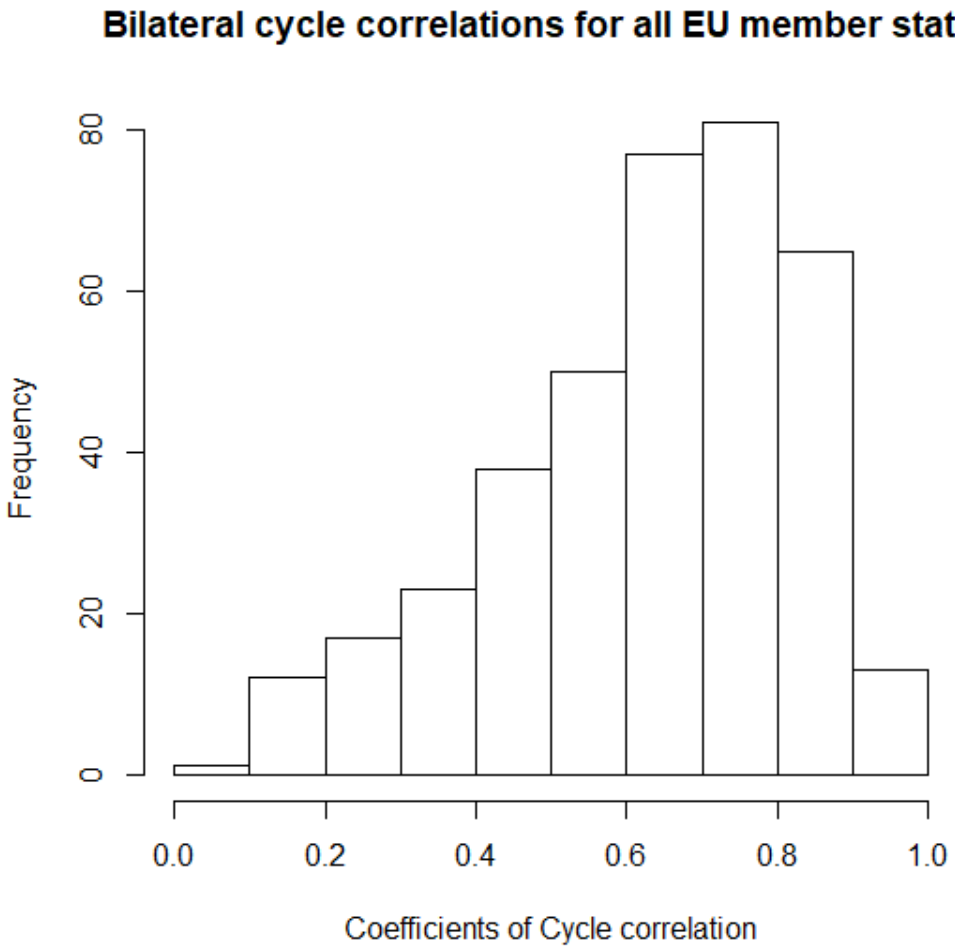


Figure 3:

The correlations between the extracted cycles for all the unique

country pairs in the EU is calculated and displayed in a histogram in figure 3.

4 Discussion on Results

The use of the Hodrick-Prescott filter seems to over estimate the degree to which countries within the EU are synchronized. This is in comparison to methods that incorporates a local trend. Furthermore, there are no pairs of countries that observe counter-cyclical correlation when using the benchmark filter, whereas the structural decomposition does exhibit so counter-cyclicality within the EU. This is significant for the literature of business cycle synchronization determinants, as it increases the variance within the most commonly used dependant variable in the study.

References

Baxter, Marianne, and Robert G King. 1999. "Measuring business cycles: approximate band-pass filters for economic time series." *Review of economics and statistics*, 81(4): 575–593.

Dées, Stéphane, and Nico Zorell. 2012. "Business cycle synchronisation: disentangling trade and financial linkages." *Open Economies Review*, 23(4): 623–643.

Harvey, Andrew C. 1990. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Kim, Hyeongwoo. 2004. "Hodrick-Prescott Filter." *Kim Hyeongwoo. March*, 12.

King, Robert G, and Sergio Rebelo. 1993. "Low Frequency Filtering and Real Business Cycles." *Journal of Economic dynamics and Control*, 17: 207–231.

Macchiarelli, Corrado. 2013. "GDP-Inflation cyclical similarities in the CEE countries and the euro area."

Localized Online Weather Predictions with Overnight Adaption*

Michael Zauner[1], Michaela Killian[1], and Martin Kozek[1]

Vienna University of Technology,
Institute of Mechanics and Mechatronics,
Getreidemarkt 9, 1060 Vienna, Austria
michael.zauner@tuwien.ac.at

Abstract. This paper describes an approach to online forecasting of ambient temperature and solar irradiation. The proposed method creates a localized prediction with an improvement over the available weather predictions ranging from 52% to 92% in ambient temperature forecast and 8% to 42% for solar irradiation forecast. This localized forecast can be used for improved predictions in smart homes or PV power plants for a more efficient operation. A new method for adapting the parameters of the autoregressive model with external input (ARX) for the solar irradiation over the night is proposed. This allows the model to be tuned to changing weather conditions without relying on external inputs.

Keywords: Weather prediction · Local optimization · Parametric model

1 Introduction

With the ever growing energy demand and the exhaustion of non-renewable resources the efficient usage of renewable energy sources (wind, solar, tidal and biomass) gets more important [4]. Due to the weather dependent nature of those renewable energy sources it is challenging to balance energy production and consumption in global electrical grids and in decentralized smart grids with smart consumers (e.g. smart homes). Therefore it is vitally important to have accurate forecasting models for those renewable energy sources [6]. The most important factor influencing solar power production via PV (photovoltaic) systems is solar irradiation [9], followed by meteorological parameters like ambient temperature and relative humidity [4]. Another application where solar irradiation and ambient temperature are important factors are the heating and cooling tasks of residential buildings [5]. There a model predictive controller for an HVAC (heating, ventilation and air conditioning) system cannot only provide better comfort for the residents, but also save energy if accurate predictions are available.
The well-known modern numerical weather forecasting services (WFS) use discrete cells for simulating weather predictions. The initial conditions for those

* This work was supported by the project "intelliEE-Home" (FFG. No. 853663) in cooperation with evon GmbH.

simulations are gathered by land based weather stations and satellite images. This results in poor localized predictions as the forecast is valid for the whole cell. The idea of this paper is to create a localized weather prediction based on the forecasts of the WFS and the past and current local sensor data.
The WFS provide information up to several days ahead [9]. The temporal resolution of those WFS predictions is typically limited to 1 hour. This resolution is usually not sufficient for the performance of most applications [3]. A common approach for short term forecasting is based on sky imaging [2] and time-series models [1, 4, 7, 9].
While predictions based on sky imaging provide good results in the range up to a few minutes, they also suffer from drawbacks: the devices are expensive, they require a lot of maintenance and predictions are only usable when the cloud cover is not too high or too low.
The usage of localized sensors and global predictions provided by WFS allow for more accurate localized predictions and a higher temporal resolution. While many authors use non-linear methods like ANN (Artificial Neuronal Networks) [7], GMDH (Group Method of Data Handling) [4] or SVM (Support Vector Machines) [9] for forecasting, a linear autoregressive model with exogenous input (ARX) is proposed in this paper. The advantages of using ARX models over ANN, GMDH and SVM are that less parameters have to be optimized and the optimization can be done in real time. Bacher et.al [1] proposed a similar modeling approach in their work. The main differences to the proposed work are the usage of a diurnal component and a clear sky approximation via smoothing kernels. The method proposed here includes furthermore an overnight prediction scheme to accommodate unmeasured changes in weather conditions.
With the usage of ARX-models the proposed method can learn statistical differences between local conditions (provided via the sensors) and the WFS predicted conditions.
The remainder of the paper is structured as follows: Section 2 explains the methods and algorithms used for the ambient temperature forecast. Section 3 highlights the necessary changes for the solar irradiation forecast. Section 4 briefly explains the simulation setup as well as the results of the simulations for ambient temperature forecast and solar irradiation forecast. In the end Sect. 5 concludes this paper.

2 Ambient Temperature Forecast

The WFS is assumed to provide an ambient temperature prediction for the next 80 hours. In the first 65 hours hourly prediction values are available. After that the WFS only provides predictions in 3 hour intervals. An example WFS prediction for the ambient temperature can be seen in Fig. 1.
The WFS prediction is linearly interpolated and a new time series $\vartheta_{\mathrm{pred}}(k)$ with the uniform sampling time of $T_s = 0.25$ h is constructed. Where $k = \{1, ..., T\}$ with T being the final time step where predictions are available.
It is assumed that a local weather station is measuring the local ambient tem-

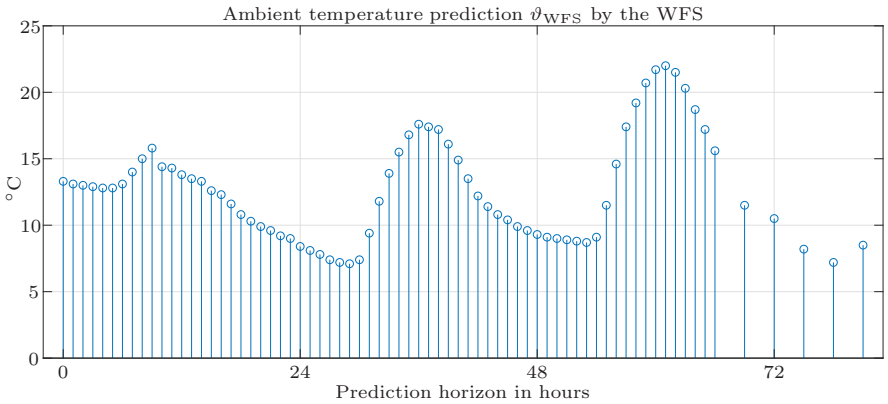


Fig. 1. Temperature prediction from the WFS. The hourly ambient temperature predictions are provided for the next 65 hours and after that in 3 h intervals for an additional 15 hours.

perature $\vartheta_{\mathrm{amb}}(k)$ every 0.25 hours. The last n measurements of the local temperature are saved in the system.
At every time step the vector

$$\boldsymbol{x}^T(k) = [\vartheta_{\mathrm{amb}}(k-n+1), \ldots, \vartheta_{\mathrm{amb}}(k), \vartheta_{\mathrm{WFS}}(k), \ldots, \vartheta_{\mathrm{WFS}}(k+m-1)] \quad (1)$$

is constructed, where $\vartheta_{\mathrm{WFS}}(k)$ is the latest WFS prediction for the current time step and $\vartheta_{\mathrm{amb}}(k)$ is the current measured ambient temperature. The variables $n \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$ represent the order of the denominator and nominator in the resulting ARX model.
Using the **W**eighted **R**ecursive **L**east **S**quares algorithm (WRLS) shown in (2a)-(2c),

$$\boldsymbol{\gamma}(k) = \frac{\boldsymbol{P}(k)\boldsymbol{x}(k)}{\boldsymbol{x}^T(k)\boldsymbol{P}(k)\boldsymbol{x}(k) + \lambda}, \quad (2a)$$

$$\hat{\boldsymbol{\theta}}(k+1) = \hat{\boldsymbol{\theta}}(k) + \boldsymbol{\gamma}(k)[\vartheta_{\mathrm{amb}}(k+1) - \boldsymbol{x}^T(k)\hat{\boldsymbol{\theta}}(k)], \quad (2b)$$

$$\boldsymbol{P}(k+1) = \frac{1}{\lambda}[\boldsymbol{I} - \boldsymbol{\gamma}(k)\boldsymbol{x}^T(k)]\boldsymbol{P}(k), \quad (2c)$$

with $\boldsymbol{P}(k) \in \mathbb{R}^{(n+m)\times(n+m)}$ being the parameter-covariance matrix, $\hat{\boldsymbol{\theta}}(k) \in \mathbb{R}^{(n+m)}$ representing the estimated parameter vector, and $\boldsymbol{\gamma}(k) \in \mathbb{R}^{(n+m)}$ the correction vector. The scalar value $\lambda \leq 1$ represents the sensitivity of the algorithm to more recent values. Choosing a λ closer to 1 will increase the weight of past samples that are significant to the current parameter estimation. With $\lambda = 1$ the WRLS algorithm will behave like a regular recursive least squares algorithm. Furthermore $\boldsymbol{I} \in \mathbb{R}^{(n+m)\times(n+m)}$ is defined as the unity matrix.
The initial value for the parameter-covariance matrix $\boldsymbol{P}$ is chosen as $\boldsymbol{P}(0) = \alpha\boldsymbol{I}$ where $\alpha \gg 1$. The initial value for $\hat{\boldsymbol{\theta}}(0)$ is chosen as a random $(n+m)\times 1$ vector.

The future predictions for the ambient temperature, denoted by $\hat{\vartheta}_{\mathrm{amb}}(k|j+1)$ where $j+1$ represents any given future time step and k the current time step are given by (3). Note that $j \geq k$ and $j < T$ must hold.

$$\hat{\vartheta}_{\mathrm{amb}}(k|j+1) = \hat{\boldsymbol{x}}^T(k|j)\hat{\boldsymbol{\theta}}(k), \tag{3}$$

where

$$\hat{\boldsymbol{x}}^T(k|j) = [\tilde{\vartheta}_{\mathrm{amb}}(k|j-n+1), \ldots, \tilde{\vartheta}_{\mathrm{amb}}(k|j), \vartheta_{\mathrm{WFS}}(j), \ldots, \vartheta_{\mathrm{WFS}}(j+m-1)], \tag{4}$$

with $\tilde{\vartheta}_{\mathrm{amb}}(k|i)$ being defined as

$$\tilde{\vartheta}_{\mathrm{amb}}(k|i) = \begin{cases} \vartheta_{\mathrm{amb}}(i) & \text{if } \ i \leq k \\ \hat{\vartheta}_{\mathrm{amb}}(k|i) & \text{else.} \end{cases} \tag{5}$$

In (4) $\vartheta_{\mathrm{WFS}}(j)$ is the most recent prediction for the time step j. Equation (5) recursively calculates predictions by (3) until only current or past measurements are needed for the formulation of $\hat{\boldsymbol{x}}^T$ in (4).
The predicted future values have to be recalculated after every new measurement since the parameter vector $\hat{\boldsymbol{\theta}}(k)$ is updated in (2b).

3 Solar Irradiation

The WFS provides the hourly solar irradiation predictions for the next 43 hours. Because of the diurnal and annual periodicity of the sun the scheme presented in Sect. 2 cannot be used without modifications. The absence of measurements during the night does not allow for parameter adaption during the night. This is problematic since the weather conditions could change drastically overnight.
In a first step the global horizontal irradiance (GHI), provided by the sensors, is transformed into the clear sky index. The clear sky index τ_{cs} is defined by

$$G = G_{\mathrm{cs}} \cdot \tau_{\mathrm{cs}}, \tag{6}$$

where G is the current global horizontal irradiation (in $\mathrm{W/m^2}$) and G_{cs} is the clear sky global horizontal irradiation (in $\mathrm{W/m^2}$). The clear sky index τ_{cs} is an indication for the transmissivity of the atmosphere. The GHI for clear sky conditions is calculated via the toolbox provided by Sandia National Laboratories [8].
As previously the vector

$$\boldsymbol{x}^T(k) = [\tau_{\mathrm{cs}}(k-n+1), \ldots, \tau_{\mathrm{cs}}(k), \tau_{\mathrm{WFS}}(k+1), \ldots, \tau_{\mathrm{WFS}}(k+m)], \tag{7}$$

is created at every time step $k = \{1, ..., T\}$, where T is the final time step where predictions are available and $\tau_{\mathrm{WFS}}(k+1)$ the next clear sky index calculated with the WFS data. It is important to note that the current prediction $\tau_{\mathrm{WFS}}(k)$ is not used, instead the next future prediction $\tau_{\mathrm{WFS}}(k+1)$ is included. This corresponds to a negative input dead time. The variables $n \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$

represent again the order of the denominator and nominator in the resulting ARX model.
During the day the WRLS algorithm shown in (2a) - (5) can be applied to compute the predictions for the clear sky index $\hat{\tau}_{\text{cs}}$.
During the night no calculations are possible due to the lack of measurements. In the morning new initial values for $\boldsymbol{P}$ and $\hat{\boldsymbol{\theta}}$ are needed since the weather conditions could have changed significantly overnight.
To calculate the new initial values the latest predictions from sunrise to sunset are queried before sunrise and concentrated into the vector $\boldsymbol{\tau}_{\text{pred}}$. These predictions are then compared against measured solar irradiation time-series of past days in a database.

Require: $\exists\ \tau_{\text{pred},i} \in \mathbb{R}^{1\times 1}, \tau_{\text{datab}} \in \mathbb{R}^{n_d \times 48}$
1: normalize τ_{pred} to 12h day $\Rightarrow \tilde{\tau}_{\text{pred}} \in \mathbb{R}^{1\times 48}$
2: **for** $j = 1$ to n_d **do**
3: $\quad e_j = \sqrt{\frac{1}{48}\sum_i^{48}(\tilde{\tau}_{\text{pred},i} - \tau_{\text{datab},j,i})^2}$
4: **end for**
5: sort e_j ascending
6: **return** $\tau_{\text{datab},j}$ of $[e_1, e_2, ..., e_K]$

Algorithm 1: Find the K most similar solar days

Algorithm 1 showcases an example on how to search a database with n_d normalized entries. For a normalized solar day the time between sunrise and sunset is defined as 12 hours. Therefore a single normalized solar day consists of 48 entries when sampled at 15 minutes intervals. Line 3 of the algorithm uses the euclidean distance to calculate the similarity, but other distances could be considered too. Algorithm 1 returns the K most similar solar days in the database.
With the usage of the clear sky index and normalizing the solar days to 12 hours, comparisons between the daily solar conditions can be drawn regardless of the time of the year.
Algorithm 2 describes the overnight prediction process. The prediction for the next day and the K most similar solar days from the database along with λ are the inputs. After normalizing the database entries and initializing $\boldsymbol{P}$ and $\hat{\boldsymbol{\theta}}$ the WRLS algorithm shown in (2a) - (5) is executed. In this WRLS algorithm $\tau_{\text{WFS}} = \tau_{\text{pred}}$ and $\tau_{\text{CS}} = \tilde{\tau}_{\text{datab},j}$ according to (7). The overnight prediction algorithm then returns the new initial values for $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{P}$ for the WRLS algorithm that is active during the next day.
The choice of K should be large enough for the parameters to settle during the overnight prediction.

Require: $\exists\ \tau_{\mathrm{pred},i} \in \mathbb{R}^{1\times 1}, \tau_{\mathrm{datab}} \in \mathbb{R}^{K\times 48}, \lambda \in \mathbb{R}^{1\times 1}$
1: normalize τ_{datab} to size(τ_{pred}) $\Rightarrow \tilde{\tau}_{\mathrm{datab}} \in \mathbb{R}^{K\times I}$
2: $\hat{\boldsymbol{\theta}}(0) = \alpha * \boldsymbol{I}$
3: $\boldsymbol{P}(0) = \mathrm{rand}$
4: **for** $j = 1$ to K **do**
5: $\quad[\hat{\boldsymbol{\theta}}(j), \boldsymbol{P}(j)] = \mathrm{WRLS}(\tau_{\mathrm{pred}}, \tilde{\tau}_{\mathrm{datab},j}, \hat{\boldsymbol{\theta}}(j-1), \boldsymbol{P}(j-1), \lambda)$
6: **end for**
7: **return** $\hat{\boldsymbol{\theta}}(K), \boldsymbol{P}(K)$

Algorithm 2: Overnight prediction

4 Results

In this section the simulation setup as well as the results of the proposed localized weather prediction algorithm described in Sec. 2 and Sec. 3 are presented. Both ambient temperature and solar irradiation forecasting simulations use data collected over a period of 36 days. The local weather station collects measurements for ambient temperature (in deg C) and solar irradiation (in W/m^2) with a common sampling time of 15 minutes. The database for the overnight prediction consists of 181 collected daily solar irradiation values from a different weather station from a different time-frame. The dataset entries were already normalized to the normalized solar day with the correct sampling time of 15 minutes.

4.1 Ambient Temperature Forecast

The values for n and m in (1) represent the order of the ARX model. The denominator-order is set by n and represents how many past measurements are used in the model. The nominator order is defined by m and corresponds to the amount of future predictions used.
Since the ARX model represents a stochastic system rather than a physical system only an optimal model order can be determined. This optimal order was determined by calculating the global significance of the model and the significance of the individual parameters and choosing the values that offer the best trade-off. The parameters used for the simulation can be found in Table 1. There Np represents the maximal prediction horizon.

Table 1. Parameters used for the ambient temperature forecast

Variable	Value
Np	96 Samples = 24 hours
n	1
m	2
λ	0.996

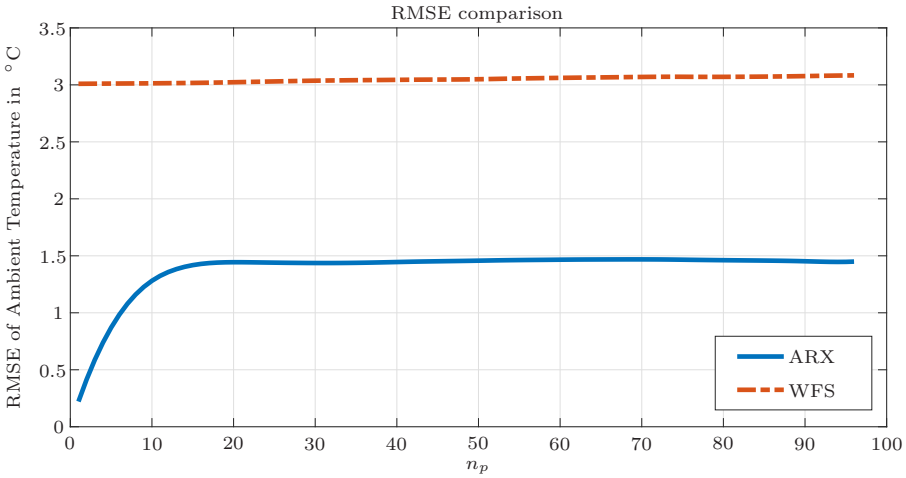


Fig. 2. RMSE between the ARX-Model output and the WFS for ambient temperature forecasting

In Fig. 2 the root mean square error (RMSE) between ϑ_{WFS} and ϑ_{amb} are shown. The RMSE for the WFS is fairly constant over the whole prediction horizon with 3°C, while the RMSE between $\hat{\vartheta}_{\mathrm{amb}}$ and ϑ_{amb} is only 1.5°C at the most and significant lower for short term predictions. This yields a forecast skill [9] ranging from 52% to 92% over the available weather predictions.
In this dataset the ARX model is primarily acting as offset compensation. The ARX model output and the WFS forecast can be seen in Fig. 3. Note that the depicted interval requires a sufficient run-in period for the parameters to settle.

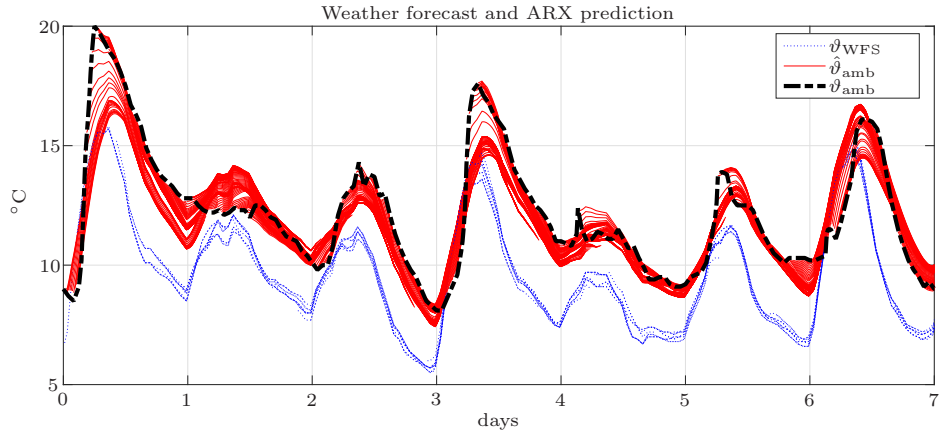


Fig. 3. Sample 7 day interval of WFS and ARX-Model output for ambient forecasting

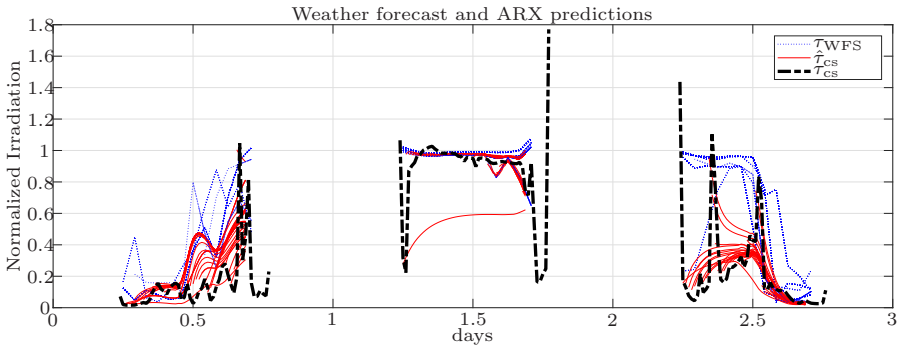


Fig. 4. Comparison of normalized WFS and ARX-Model output for solar forecasting for 3 selected days

4.2 Solar Irradiation

As seen in Sec. 4.1, the optimal values for n and m were evaluated by examining the parameter significance and the global model significance. The chosen simulation parameters can be found in Tab. 2 alongside with the parameter K which is used for the overnight prediction.

Table 2. Parameters used for the solar forecast

Variable	Value
Np	till sunset
n	1
m	3
λ	0.98
K	5

In Fig. 4 the ARX model output $\hat{\tau}_{cs}$ and the WFS predictions τ_{WFS} can be seen for 3 selected days. The black dash-dotted line represents the real (measured) clear sky index τ_{cs}. During the start/end of the sunny days the value for the clear sky index τ_{cs} shows erratic changes. This is due to the small magnitudes of solar irradiation measured and also small magnitudes of clear sky solar irradiation G_{cs} which leads to ill-conditioned normalization. During the day, when the clear sky index settles, the predictions are reliable. In Fig. 5 the outputs have been converted to GHI in W/m^2 and the clear sky global horizontal irradiation G_{cs} is plotted for reference.

Figure 6 gives a better overview on the accuracy of the ARX model compared to the predictions provided by the WFS. The development of the RMSE error over the prediction horizon is plotted for the ARX model output and the WFS predictions. Both predictions have an increasing RMSE over the prediction horizon with the ARX predictions displaying a significantly better RMSE for short

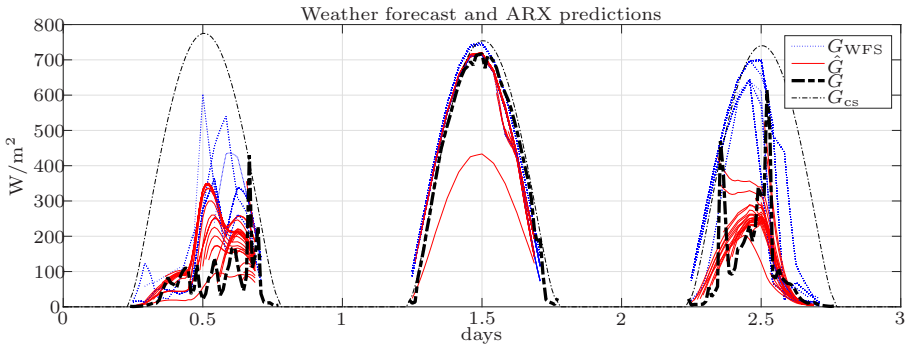


Fig. 5. Comparison of WFS and ARX-Model output for solar forecasting for 3 selected days

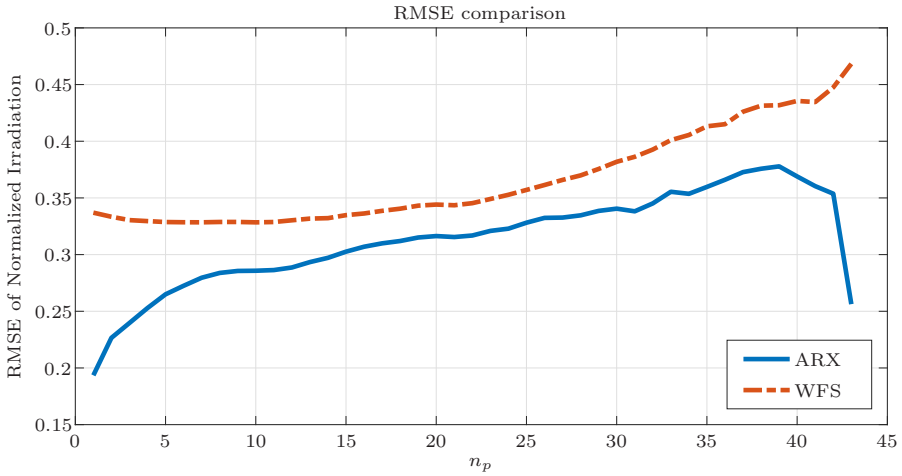


Fig. 6. RMSE between the ARX-Model output and the WFS for solar forecasting

term predictions and at the end of n_p. The forecast skill over the WFS model is 8% to 42% for the solar irradiation forecast.

5 Conclusion

Inspired by previous studies a forecasting method for ambient temperature and solar irradiation has been developed. The proposed approach localizes the numerical weather prediction provided by WFS to increase the local accuracy and reduce the forecasting errors. The overnight adaption of the ARX model parameters allow the model to accommodate to unmeasured changes and ensures a well tuned start into the next day.
Simulation results with real data showcase the benefits of the proposed methods. The method is not computational intensive and can easily be run online

on low-cost CPUs, for example in home automation systems. There the localized forecasts could provide better predictions for smart home controllers and therefore increase the comfort, save money and energy.

References

1. Bacher, P., Madsen, H., Nielsen, H.A.: Online short-term solar power forecasting. Solar Energy **83**(10), 1772–1783 (2009)
2. Chow, C.W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B.: Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. Solar Energy **85**(11), 2881–2893 (2011)
3. Fernndez-Peruchena, C.M., Gastn, M.: A simple and efficient procedure for increasing the temporal resolution of global horizontal solar irradiance series. Renewable Energy **86**, 375 – 383 (2016). https://doi.org/https://doi.org/10.1016/j.renene.2015.08.004, http://www.sciencedirect.com/science/article/pii/S0960148115302044
4. Giorgi, M.D., Malvoni, M., Congedo, P.: Comparison of strategies for multi-step ahead photovoltaic power forecasting models based on hybrid group method of data handling networks and least square support vector machine. Energy **107**, 360 – 373 (2016). https://doi.org/https://doi.org/10.1016/j.energy.2016.04.020, http://www.sciencedirect.com/science/article/pii/S0360544216304261
5. Killian, M., Zauner, M., Kozek, M.: Comprehensive smart home energy management system using mixed-integer quadratic-programming. Applied Energy **222**, 662 – 672 (2018). https://doi.org/https://doi.org/10.1016/j.apenergy.2018.03.179, http://www.sciencedirect.com/science/article/pii/S0306261918305282
6. Larson, D.P., Nonnenmacher, L., Coimbra, C.F.: Day-ahead forecasting of solar power output from photovoltaic plants in the american southwest. Renewable Energy **91**, 11 – 20 (2016). https://doi.org/https://doi.org/10.1016/j.renene.2016.01.039, http://www.sciencedirect.com/science/article/pii/S0960148116300398
7. Marquez, R., Coimbra, C.F.: Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the nws database. Solar Energy **85**(5), 746–756 (2011)
8. Stein, J.S.: The photovoltaic performance modeling collaborative (pvpmc). 2012 38th IEEE Photovoltaic Specialists Conference (2012). https://doi.org/10.1109/pvsc.2012.6318225, https://pvpmc.sandia.gov/
9. Wang, Z., Tian, C., Zhu, Q., Huang, M.: Hourly solar radiation forecasting using a volterra-least squares support vector machine model combined with signal decomposition. Energies **11**(1), 68 (2018)

Storm characterization using a BME approach

Manuel Cobos[1], Andrea Lira-Loarca[1], George Christakos[2], and Asunción Baquerizo[1]

[1] Andalusian Institute for Earth System Research. University of Granada. Av. del Mediterráneo s/n, 18006 Granada, Spain.
mcobosb@ugr.es, aliraloarca@ugr.es, abaqueri@ugr.es
[2] Geography Department, San Diego State University, San Diego, CA, USA.
gchrista@sdsu.edu

Abstract. The storm that occurred at the Spanish coast of the Mediterranean sea during the end of January of 2017 produced severe coastal floods. This event is analyzed in space and time using a combination of the spatio-temporal random field (S/TRF) theory and the Bayesian maximum entropy (BME) method. The results include sea wave heights statistics, which can be used to assses the uncertainty of flooding risk at coastal zones, a key question for coastal area management purposes.

Keywords: Storm characterization, flood risk, spatiotemporal random field, Bayesian maximum entropy

1 Introduction

It is widely recognized by the scientific community that climate change presents an unstable and uncertain picture [1]. It is expected that climate-driven effects can cause an increase of the mean sea level during the current century, but also, an increase of the frequency and severity of exceptional events like storms. This situation increases the vulnerability of large coastal regions with house properties, infrastructures and industrial instalations. This is an issue in Spain where a high percentage of its more than 4000 kilometers of coastline are occupied and are very vulnerable to climate-driven hazards.

A new approach is used to describe the space-time distribution of sea-wave heights during storms based on the knowledge-bases (KB) proposed by [2, 3] and the BME theory that is widely applied in Earth Sciences [4]. Some quantitative site-indicators are also provided to further characterize the spatiotemporal attributes of sea-wave heights during storm propagation.

2 Methodology

The BME approach involves three main stages [2, 3]. First, an adequate space-time geometry and rules of KB integration are selected. These relations are established in terms of a set of moment equations $g_\alpha(\chi_m)$, $\alpha = 1, .., N$; g_α that represent known statistics of the S/TRF $X(p)$; N denotes the number of

stochastic moments; and $\chi_m = \chi_s \cup \chi_k$ are the union of specificatory data and the unknown values at the estimation space-time points. A mean estimator based on the optimization of the information criteria was used. Therefore, the general knowledge equations are expressed as:

$$\overline{h_\alpha(p_m)} = \overline{g_\alpha(x_m)} = \int g_\alpha(\chi_m) f_G(\chi_m; p_m) d\chi_m. \quad (1)$$

At this stage, the general knowledge was defined in terms of equations and the form of the general KB-based pdf is assumed $f_G(\chi_m) = e^{\mu_0 + \mu'_\alpha g_\alpha}$, where μ_0, μ_α is a vector of coefficients associated with $g_\alpha(\chi_m)$; usually $\overline{g_0} = 1$ and μ_0 accounts for the normalization constraint.

Second, hourly-data from monitoring stations (χ_h) and simulations (χ_s) are obtained using data provided by Puertos del Estado (Spain). Sea wave heights vary slightly in space and time, therefore, the physical process is assumed to be homostationary [3], and the covariance is a function of the distance h and the time-lag τ. Then, the covariance of specificatory data can be written as $c_\chi(h,\tau) = [\chi_S(s,t) - \overline{\chi_S(s,t)}][\chi_S(s+h,t+\tau) - \overline{\chi_S(s+h,t+\tau)}]$, where $c_\chi(h,\tau)$ is the covariance of soft-data.

Finally, the general KB-based probability density funcion (pdf) (eq. 1) is updated with the specificatory data as $P_K[x_k \leq \chi_k] = P_G[x_k \leq \chi_k \vee \chi_{data}]$, yielding:

$$\overline{h_\alpha(p_m)} = \int g_\alpha(\chi_m) f_K(\chi_m; p_m) d\chi_m, \quad (2)$$

where $f_K(\chi_k) = A^{-1} \int_B f_G(\chi_m) d\chi_s$; $A^{-1} = \int_B f_G(\chi_{data}) d\chi_s$ is a normalization constant; and, B is the domain given by the data.

3 Results

The methodology is applied to the storm that ocurred between 20th and 24th of January of 2017. The severity of this storm led to several floods in the coastal area of Valencia (north-west Meditarrean Sea, Spain).

The first and second centered and single point moments were used in space-time estimation (eq. 1). An isotropic space-time covariance model was fitted to the values of different data types. Based on 1440 spacetime data points distributed in 10 locations, we made a linear regression of pair-points where simulated and observed data coexisted. The computation of the BME posterior PDFs provides the statistical description of the mean and standard deviation of seawave heights.

As it is observed in figure 1, the storm is generated at the Strait of Sicilia and travels over the north basin of the Mediterranean sea. Southeastern winds favour the growth and propagation of waves from the south of Sardinia (Jan, 20th 21.00). The fecth is maximum and persistent winds drive higher sea waves. The wave front reaches the Balearic Islands, and then is divided into two fronts that propagate over the lower and upper parts of the basin (Jan, 21th 06.00).

During this stage maximum wave height ($\sim$ 7 meters) are observed. The bottom front reaches quickly the Gulf of Valencia and is dispersed downward. The upper front propagates until the Gulf of Lion, and then it decays too. The same pattern is observed at the video included as supplementary material where SIMAR hindcast data (Puertos del Estado) was used.

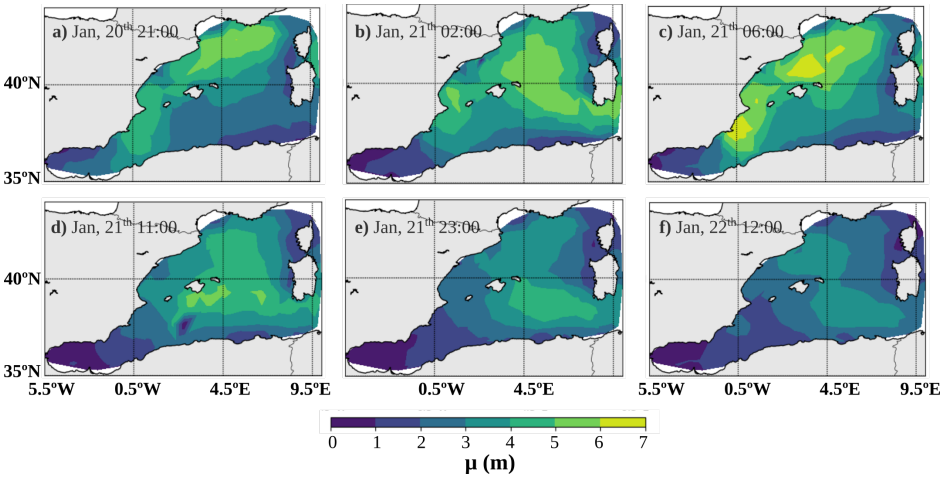


Fig. 1. Estimated mean wave height at the study area for several relevant dates.

The accuracy of the methodology was assessed in terms of a 10-fold cross-validation technique. The computed mean absolute error and root mean square error were lower than 1.2 and 0.15 meters, respectively. An step-by-step application of the spacetime methodology to seawave height evolution during storms and an in-depth analysis of results and site-characterization will be provided in the presentation.

Acknowledgments. This research was partially funded by the Campus de Excelencia Internacional del Mar (Cei-MAR) and the program of Precompetitive Research projects for young researchers of the UGR plan (PPJI_B-06). It was also supported by AQUACLEW. Project AQUACLEW is part of ERA4CS, an ERA- NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Comission. M.C. acknowledges the mobility support received for the research stay at the San Diego University State and grateful to the Department of Geography (SDSU) for kindly hosting him during the analysis of the research.

References

1. Impacts, I. P. C. C. (2014). adaptation, and vulnerability. Part A: global and sectoral aspects. Contribution of Working Group II to the fifth assessment report of the Intergovernmental Panel on Climate Change. Climate change.
2. Christakos, G. (2000). Modern spatiotemporal geostatistics (Vol. 6). Oxford University Press.
3. Christakos, G. (2017). Spatiotemporal random fields: theory and applications. Elsevier.
4. He, J., & Kolovos, A. (2017). Bayesian maximum entropy approach and its applications: a review. Stochastic Environmental Research and Risk Assessment, 1-19.

Air Pollution Forecasting using Machine Learning Techniques

Marijana Cosovic[1], Emina Junuz[2]

[1]Faculty of Electrical Engineering, University of East Sarajevo, Istocno Sarajevo, BiH
[2]Faculty of Information Technology, Dzemal Bijedic University, Mostar, BiH

marijana.cosovic@etf.ues.rs.ba, emina@edu.fit.ba

Abstract. Air pollution in Sarajevo, the capital of Bosnia and Hercegovina, is an ever-increasing problem. Very little research is being conducted in the domain of air pollutants' impact on public health in B&H. Recently, there has been an initiative of several parties for deployment of a particle matter ($PM_{2.5}$) sensor network in Sarajevo, and based on this valuable data, future research can be conducted.

Weather behavior can be analyzed as time series data with hourly, daily, weekly, and seasonal periodicities. Machine learning techniques are proving useful for air pollution forecasting. In this paper we evaluate the performance of several machine learning algorithms applied to air quality and meteorology datasets.

The network architecture based on multilayer perceptron (MLP), which is a feed-forward type of artificial neural network, was used for air pollution forecasting. We also used Long Short-Term Memory (LSTM) units for building a Recurrent Neural Network capable of learning data long-term dependencies. We compare prediction accuracies of urban air quality, as this is of significant importance to the public.

Keywords: Machine learning·Air pollution forecast·MLP·LSTM·

1 Introduction

The most important effects of air pollution are on human health [1], the ecosystem [2] and the human built environment [3]. The Federal Ministry of Environment and Tourism B&H issued a rule book on the method of air quality monitoring, defining of pollutant types, limit values and other air quality standards [4] that is in accordance with USA and EU standards regarding ambient air pollutants. Even though we recognize effects of indoor air quality on human health, outdoor air quality was the primary focus in this study.

Air pollutants considered in this study are: PM_{10}, SO_2, NO_2, CO and O_3. In addition, we used atmospheric pressure, temperature, relative humidity and wind speed. We aim to predict PM_{10} concentration, based on air pollutant concentrations and major meteorological data, by using machine learning techniques. By exploring a multilayer percep-

tron (MLP) feedforward neural networks and Long Short-Term Memory (LSTM) recurrent neural networks we choose a suitable approach to deal with nonlinear systems such as air pollution and aim at next-day particulate matter forecast.

2 Datasets

The datasets are obtained courtesy of the Federal Hydrometeorological Institute (FHMZ) BH [5]. There are four measuring stations in the vicinity of the greater Sarajevo area (Bjelave, Otoka, Vjecnica, and Ilidza) located and operated by either the Federal hydrometeorological Institute or the Cantonal Institute for Public Health. Location data of the measuring stations is given in Table 1. The datasets report continuous measurements (averaged hourly values) of the temperature, relative humidity, pressure, average wind speed, PM_{10}, SO_2, NO_2, CO and O_3 during 2017.

Table 1. Station location: latitude, longitude and altitude

Station	latitude (°,',")	longitude(°,',")	altitude (m)
Bjelave	43 52 03 N	18 25 23 E	631
Otoka	43 50 54 N	18 21 49 E	512
Vijecnica	43 51 33 N	18 26 04 E	554
Ilidza	43 49 40 N	18 18 49 E	499

We consider a one-year period with hourly averaged values for air pollutants and major meteorological variables. Choosing these variables (date, time, type of day (working week day, Saturday, and Sunday), temperature, relative humidity, pressure, average wind speed, PM_{10}, SO_2, NO_2, CO and O_3) was shown suitable for detecting/forecasting elevated values of particle matter in the ambient air. The dimension of the feature matrix is 8760x13. Hourly average values for PM_{10} concentration during the year 2017 are shown in Fig. 1.

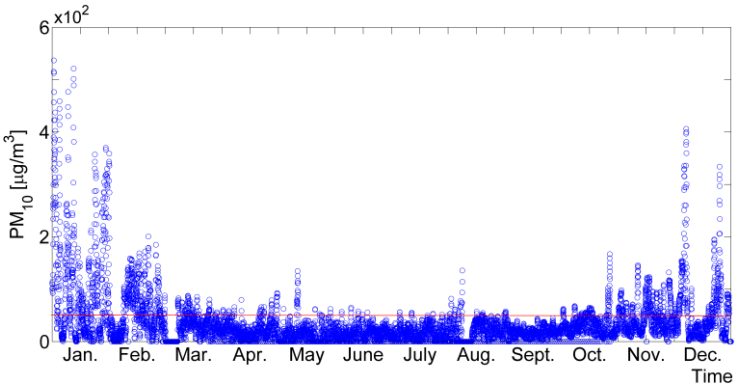


Fig. 1. PM_{10} concentrations measured at Bjelave station during 2017.

The red line in the figure represents a guideline value of 50 $\mu g/m^3$ suggested by the World Health Organization, and that is the maximum tolerated value at any given time. During the winter season that value is exceeded frequently as shown in Fig. 1. The aim of this paper is to structure a forecasting problem in which the pollution of the next day is determined based on previous data.

2.1 Missing data

Ambient air pollutant data was continuously measured during 2017 at Bjelave monitoring station, an urban part of Sarajevo city. Only data that has passed FHMZ control and validation is reported in this study. The table below shows available and missing hourly average air pollutant data. Daily average values are also computed based on hourly average values for those days that are not missing more than six hourly average values. For example, 40 hourly average values are missing for PM_{10} pollutant during the fall season, but since those missing values are spread out among different days, a few values at a time, the daily average value is computed for all of the days in the fall of 2017.

Missing hourly average values for all five air pollutants considered are in the range of 8-20 percent and are due to analyzers not performing the measurements. Equipment failure, need for calibration of analyzers and unforeseen problems that are either difficult to recognize at their occurrence or resolve in a short amount of time are some of the reasons for missing data. In general, calibration of the equipment is done during the summer season. The increased number of missing values during that period could be contributed to the scheduled outages.

Table 2. Available and missing hourly average values (total and separated by seasons) for five pollutants during 2017

Pollutants	Available hourly average values	Missing hourly average values	Missing data by season (Spring, Summer, Fall, Winter)
PM_{10}	7761	999	363, 384, 40, 212
SO_2	6996	1764	528, 711, 226, 299
NO_2	7278	1482	220, 995, 128, 139
CO	8056	704	187, 452, 12, 53
O_3	7785	975	280, 445, 39, 211

As the data from various sources are becoming abundant, the problem of missing data needs to be addressed accordingly. Within our datasets, sporadic points were missing, along with longer gaps of up to several days. Firstly, we considered deletion of the entire row (time instance) in which certain features have missing values, but that failed as a plausible option due to the limited number of instances in the dataset (one calendar year of observations). Generation of methods and tools for effective data assertion were explored. Hence, we addressed data imputation and looked at statistical and machine learning approaches [6] for those. Computation and insertion of overall mean, although

a fast method for data imputation, introduces side effects such as dataset variance reduction.

We considered seasonal adjustment since our data has seasonal variation. Time-series specific methods of data imputation, such as last observation carried forward (LOCF) and next observation carried backwards (NOCB), were used. Authors in [6] presented imputation model based on machine learning techniques (LASSO regression and Bagging Ensemble). Results reported show improvements in suggested data imputation method in comparison to hot deck methods. We have used the root mean squared error (RMSE) to evaluate different methods of data imputation. Performance of LASSO-Bagging method has shown reduction in RMSE values on our datasets.

3 Air Pollution Forecasting Models

We have considered four seasons for air pollution forecasting: spring, summer, fall, and winter. The dataset for the winter season was collected every hour from January 1, 2017 until March 21, 2017 as well as December 22, 2017 until December 31, 2017. The spring season dataset included data from March 21, 2017 until June 21, 2017. The summer season dataset was collected from June 21, 2017 until September 23, 2017, while the fall dataset was collected from September 23, 2017 until December 22, 2017. The datasets contain 2208 instances for spring (92 days observed), 2256 instances for summer (94 days observed), 2160 instances for fall (90 days observed) and 2136 instances for winter (89 days observed). We considered nineteen features for each of the datasets: date, time, type of the day (working week day, Saturday, and Sunday), PM_{10}, SO_2, NO_2, CO, O_3, previous-days PM_{10} (up to seven features for seven previous days), atmospheric pressure, temperature, relative humidity and wind speed. Hence, the dimensions of the spring, summer, fall, and winter feature matrices are 2208x19, 2256x19, 2160x19, and 2136x19 respectively.

3.1 Correlation

Correlation among meteorological parameters and air pollutants can be measured with the use of different coefficients: Pearson's, Spearman's and Kendall's correlation coefficients. The Pearson correlation coefficient is a measure of linear correlation between normally distributed variables. When the variables are not normally distributed or the relationship between the variables is not linear, it is more appropriate to use the Spearman rank correlation coefficient [7]. The Kolmogorov-Smirnov test was used to determine the type of distribution of our data. Hence, Spearman's correlation coefficient (SCC) was used. SCC assesses the monotonic relationship between the variables and has a value between –1 and +1. Results of SCC for all year data are shown in Table 3. Spearman's correlation coefficient amongst selected features was computed for all four seasons, as well, and shows stronger correlation during the winter season when the air pollution is at its maximum. In addition, positive correlation is observed between PM_{10}, CO, NO_2, and SO_2, while negative correlation is observed between those air pollutants and O_3. We also observed negative correlation of all air pollutants and temperature apart from O_3. Atmospheric pressure and relative humidity also have positive

correlation coefficient with PM_{10}, CO, NO_2, and SO_2 and a negative correlation with O_3. The largest correlation coefficients are present between PM_{10}, CO, NO_2, and SO_2, confirming the fact that these air pollutants are originating from the same sources.

Table 3. Spearman's correlation coefficient amongst selected features

	PM_{10}	**CO**	**NO_2**	**SO_2**	**O_3**	**Press**	**Temp**	**Hum**	**Wind**	**Class**
PM_{10}	1	**0.55**	**0.49**	**0.36**	-0.25	**0.2**	-0.39	**0.1**	-0.22	**0.58**
CO		1	**0.7**	**0.24**	-0.39	**0.18**	-0.4	**0.26**	-0.29	**0.34**
NO_2			1	**0.41**	-0.26	**0.2**	-0.28	**0.11**	-0.38	**0.33**
SO_2				1	-0.03	**0.17**	-0.2	**0.06**	-0.19	**0.33**
O_3					1	-0.13	**0.55**	-0.51	**0.28**	**0.11**
Press						1	-0.19	**0.08**	-0.12	**0.16**
Temp							1	-0.59	**0.18**	-0.09
Hum								1	-0.23	-0.1
Wind									1	-0.14
Class										1

We also observe that temperature and wind speed have a negative correlation coefficient with PM_{10}, CO, NO_2, and SO_2 and a positive correlation with O_3. Daily temperatures, relative humidity and wind have effects on O_3 formation. Generally, more beneficial meteorological circumstances for ozone formation are higher temperatures with lower relative humidity, as opposed to lower temperatures with higher relative humidity. Also, depending on wind speed (high/light), we could have dilution or building up of ozone concentration, hence positive correlation of wind speed and ozone concentrations.

3.2 Methodology

We used the Keras [8] deep learning Python library with the Theano backend for development and evaluation of deep learning models. Two models were developed for air pollution forecast: MLP and LSTM models. Model, based on a backpropagation algorithm for training of fully connected multiplayer perceptron (MLP) neural network is similar to our previous work [9] developed for short-term load forecasting. It is defined as a sequence of layers: an input layer, hidden layers and an output layer. The shape of the input data needs to be specified only for the first layer in the sequence. We devised MLP models with 13-19 features for the input layer. In forecasting PM_{10} concentration we opted to generate PM_{10} values up to seven days prior. A model that minimizes performance measures presented in 3.3 will be chosen. In Keras, using Dense class is one of the ways to define fully connected layers. Network weights were initialized to random numbers using either uniform or Gaussian distribution. Use of appropriate activation function allows for better training of the network [10]. Traditionally, sigmoid and tanh activation functions are used, but the authors in [10] have shown that better performance can be achieved using a rectifier activation function.

We use 10-fold cross validation for determining accuracy on the test dataset, and as we increase the number of hidden layers beyond two, classification accuracy decreases,

as noticed in [9]. In our case, a neural network with two hidden layers is the optimal model for air pollution forecasting. Using either too few or too many neurons in the hidden layers may result in problems of underfitting and overfitting, respectively. General guidelines for determining the number of neurons within each hidden layer are used. We selected neural network architecture based on trial and error, but in accordance with the following general guidelines: the number of neurons in hidden layers should be between the sizes of input and output layers, and they should be the sum of 2/3 of the input layer neurons and output layer neurons. Hence, we trained the neural network with two dense hidden layers with 15 and 10 neurons, respectively [9].

While in the feed-forward neural networks information travels in forward direction only, recurrent neural networks (RNN) can maintain information from computation of an earlier input, hence having memory capabilities. RNN performance degrades when long-term dependencies between previous inputs and present targets occur. Implementation of a LSTM (Long Short-Term Memory) cell allowed for better tradeoff concerning RNN performance at one side and lapsed time between previous inputs and present targets on the other side. A LSTM network is RNN composed of LSTM cells. LSTM solve the vanishing gradient problem of RNN by updating the state of each cell in an additive way. We developed a LSTM model for air pollution forecasting in the Keras deep learning library such that given the meteorological conditions and concentration of pollution of prior days, as well as expected air pollution for the next day, we can forecast air pollution for the next day. All the features are normalized with a zero mean and standard deviation of one. Datasets are split into training and testing datasets, and we fit our LSTM model on 80% of the data and evaluate it on the remaining 20%. We trained the LSTM model with 50 neurons in the first hidden layer. A neuron in the output layer enabled prediction of air pollution. We used an Adam optimization algorithm [11] instead of stochastic gradient descent because of its forthright implementation, computation efficiency and small memory requirements.

3.3 Performance Measures

The mean absolute error (MAE) is the sum of absolute differences between the actual value and the forecast, divided by the number of observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| f_i - y_i \right| \tag{1}$$

Hence, the mean absolute error is an average of the absolute errors where f_i is the prediction and y_i the actual value as shown in (1) where all individual differences have equal weight. The mean absolute percentage error (MAPE) is another measure of prediction accuracy of a forecasting model, and it is shown in (2). It is the average of absolute errors divided by actual observation values.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left| \frac{f_i - y_i}{f_i} \right| \tag{2}$$

The mean squared error (MSE) shown in (3) represents the sum of the squared errors divided by the number of observations.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \tag{3}$$

The mean square error (MSE) is probably the most commonly used error metric. It penalizes larger errors because squaring larger numbers has a greater impact than squaring smaller numbers.

The root mean squared error (RMSE) shown in (4) represents the sample standard deviation of the differences between predicted and true values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{4}$$

4 Results and Conclusion

We have forecasted air pollution in the last week of all four season datasets. Those periods were not included during the training step. The air pollution forecast based on the MLP model for the winter season was performed from March 15 until March 21, and its comparison to real air pollution is shown in Fig. 2. Elevated values of particulate matter can be observed during that period as well as increase in the evening hours mostly due to coal burning for heating. Air pollution forecast for the spring season was performed from June 15, until June 21.

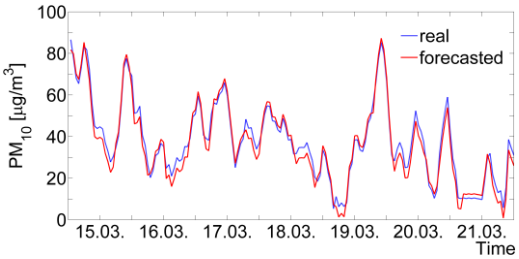


Fig. 2. MLP model Real/Forecasted PM_{10} Concentration from March 15, until March 20, 2017 at Bjelave Station

A comparison of real and forecasted air pollution concentration on March 20, 2017 (winter season) and on June 20, 2017 (spring season) using MLP model is shown in Fig. 3., on the left and right respectively. Those are the best prediction results that were obtained with one day prior information, hence the performance measure values were the smallest in that case. All of the performance measures are negatively-oriented scores, which means that lower values are better.

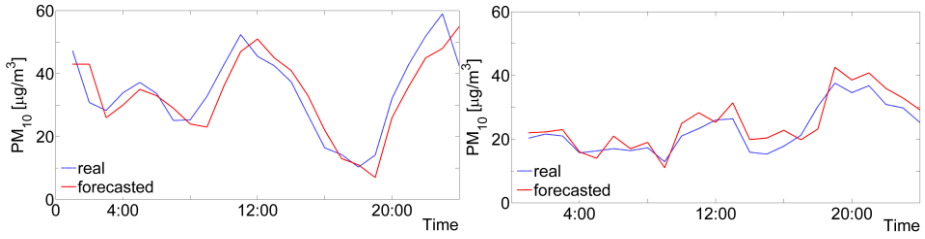


Fig. 3. MLP model Real/Forecasted PM_{10} Concentration on March 20, 2017 (Winter season) and June 20, 2017 (Spring season) at Bjelave Station

Summer season air pollution forecast was performed from September 17 until September 23, and for the fall season from December 16 until December 22. A comparison of real and forecasted air pollution concentration on September 22, 2017 (summer season) and on December 21, 2017 (fall season) using LSTM model is shown in Fig. 4., on the left and right respectively. Those are the best prediction results that were obtained with one day prior information, hence the performance measure values were the smallest in that case.

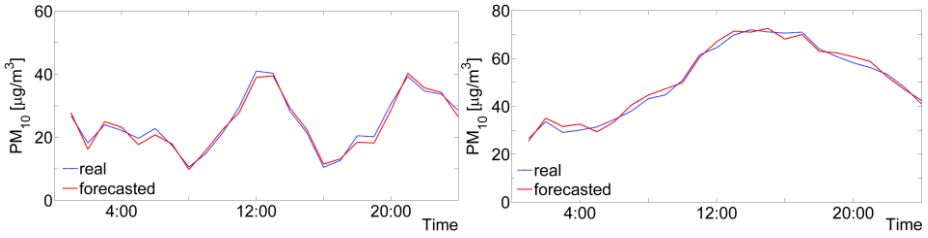


Fig. 4. LSTM model Real/Forecasted PM_{10} Concentration on September 22, 2017 (Summer season) and December 21, 2017 (Fall season) at Bjelave Station

LSTM model, when compared to MLP model, provides reduced performance measures on all datasets considered. The largest positive correlation is observed between particulate matter and carbon monoxide as shown in Table 1. We observed that during spring and summer season the most missing data comes from PM_{10} and CO pollutants. We conclude that due to this phenomenon performance measures shown in Table 4. are slightly increased but nevertheless smaller than in case of MLP model. Air pollution forecast based on LSTM model for the remaining seasons were performed for the same dates as in MLP forecasting model.

Table 4. MAPE and RMSE values for Bjelave station using MLP and LSTM models for Winter, Spring, Summer and Fall season

Station/ Season		MLP		LSTM	
		MAPE	RMSE	MAPE	RMSE
Bjelave	Winter	0.019	0.027	0.018	0.023
	Spring	0.023	0.037	0.019	0.022
	Summer	0.027	0.039	0.021	0.031
	Fall	0.015	0.026	0.012	0.024

We have developed two models for air pollution forecasting based on artificial neural networks: feed-forward MLP and recurrent LSTM neural networks. In this paper we analyzed the performance of those models applied to datasets collected in one calendar year from one location in the Sarajevo city area. Since we encountered missing data problems, we explored statistical and machine learning methods for data imputation. We have considered four seasons for air pollution forecasting: spring, summer, fall, and winter. The MLP model with two hidden layers was optimal since choosing additional hidden layers caused performance indices to deteriorate. The LSTM model used one hidden layer. We used a cross-validation technique to determine the number of neurons in each of the layers. We concluded that some features had greater effect than others on the forecast, as well as that performance measures were the best for forecast done based on the previous day's information. By using more prior information, performance indices worsened. LSTM model performed slightly better than MLP model for all seasons considered. For future work we will explore LSTM model with additional layers as well as other methods of data imputation in search of reducing performance measures further.

References

1. G. S. Martinez, J.V. Spadaro, D. Chapizanis, V. Kendrovski, M. Kochubovski, P. Mudu, "Health Impacts and Economic Costs of Air Pollution in the Metropolitan Area of Skopje," International Journal of Environmental Research and Public Health 2018, vol. 15, no. 4: 626 [Online]. Available: https://doi.org/10.3390/ijerph15040626
2. L. Jones, G. Mills, A. Milne, "Assessment of the impacts of air pollution on ecosystem services – gap filling and research recommendations," (Defra Project AQ0827), Final Report, July 2014.
3. M. C. Rodriguez, L. Dupont-Courtade, W. Oueslati, "Air pollution and urban structure linkages: Evidence from European cities," Renewable and Sustainable Energy Reviews, vol. 53, pp.1-9, 2016. [Online]. Available: https://doi.org/10.1016/j.rser.2015.07.190
4. Rule book on the method of air quality monitoring. [Online]. Available: http://extwprlegs1.fao.org/docs/pdf/bih149115.pdf
5. Federal Hydrometeorological Institute (FHMZ) Bosnia and Herzegovina. [Online]. Available: http://www.fhmzbih.gov.ba/latinica/index.php
6. Rosati, G. (2017). Construcci´on de un modelo de imputaci´on para variables de ingresos con valores perdidos a partir de Ensamble Learning. Aplicaci´on a la Encuesta Permanente de Hogares, Revista SaberES, 9 (1), 91-111.

7. Hauke J, Kossowski T (2011) Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae 30(2):87–93
8. F. Chollet, Keras, (2015), GitHub repository. [Online]. Available5: https://keras.io
9. Bećirović E, Ćosović M (2016) Machine learning techniques for short-term load forecasting. In: 4th International symposium on environmental friendly energies and applications. Serbia, Belgrade
10. Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In Proceedings of ICML. 807-814.
11. D. P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs.LG], December 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

Forward Regression with Discrete and Continuous Wavelet Time-Frequency Window

-An application to the Market Line-

Roman MESTRE Michel TERRAZA

University of Montpellier - MRE

Abstract.

The Rolling-Regression are currently used to study the parameters stability over time. In finance, we can analyse the time evolutions of systematic risk relaxing the constant-Beta hypothesis. This method can be associated with a wavelet decomposition of the variables in order to the parameters stability of frequency regression. Then, we compare continuous and discrete wavelets methodologies of rolling regression with the standard rolling regression. The discrete methods are based on time-frequency window but we compare if we have to use it on the wavelets filter output or directly on the series and realize the wavelet decomposition at each step of the window. The continuous method is based on wavelets coherence-phase. We use daily data of AXA returns and the CAC 40 index from 2005 to 2015. We show that the differences between discrete methods are more important at Low-Frequencies and we compare the results with the Continuous Time-Frequency Betas. JEL codes G00 G11 G12.

Keywords: Time-Frequency Rolling Regression, Wavelets, Time-Frequency Betas

1 Introduction

Frequency analysis or spectral analysis of time series in economics, developed from Fourier's work and harmonic analysis, has several limits (choice of the window) because its imply an arbitrage between Time and Frequency information. Timelessness is one of these for its, despite the work of Priestley on the evolutionary spectrum.

The Time-Frequency Analysis, elaborated by Haar in 1909 and Gabor (1946), is an important advance in this field to overcome the Timelessness. Popularized under the name of wavelets, this methodology separates the frequency components of a time series and allows a time representation of it. Consequently, the wavelets are suitable to take into consideration the hypothesis about agents' homogeneous behaviour during the portfolio/ investment choices (as example, the Wavelets Value-at-risk). This is a univariate approach of wavelets but there is bivariate case. Gençay et al (2003) apply wavelets to the CAPM theory and the market line in order to estimate the systematic risk according to the behaviour of agent illustrated by heterogeneous investment time horizon.

Many studies as those of Black,Jensen et Scholes (1972), Fama et McBeth (1973) focused on CAPM limits and the instability of its main parameter the Beta. The rolling windows are developed to overcome this problem as showed Faff et al (1992, 1998), Brooks et al (1992, 1998). Groenewold and Fraser (1997) and Yeo (2001) compare the use of GARCH processes, rolling and recursive windows and Kalman-Bucy filter to estimate a time varying beta. The Rolling Forward window associated with the Wavelets is then the theoretical extension of these works within the CAPM framework. Mestre and Terraza (2017) showed that the beta value (of equity) is strongly volatile but it is differentiated across frequencies. To highlight this result, we use the rolling forward regression with a fixed size window associated with discrete wavelets. This method is equivalent to consider a unique discrete wavelets decomposition of the interest series. Conceptually this is contestable because the length of the series is an important parameter to determine the wavelet decomposition order operated by successive filtering and subsampling. In this case, if we use traditionally the rolling window on Wavelets outputs, the results may be wrong. Moreover, the discrete wavelets transformation is a practical simplification of Continuous Wavelets because its is based on dyadic scale regrouping the frequencies in bands. The Low-frequencies bands is related to large investment horizon, which can limit its use.

In this paper, we compare two discrete wavelets methodologies to estimates a time-varying beta: a time rolling window applied on wavelets outputs estimating the beta or the time-frequency window decomposing (with wavelets) the series and estimating the beta simultaneously at each step of the rolling window. Also, we developpe a extension based on continuous wavelets to estimate more precisely the beta at each time and frequency. We use the daily returns of AXA and the CAC40 index from 2005 to 2015 in order to estimate the Beta of it market line.

2 Theoretical aspect of Wavelets applied to CAPM

Our paper is based on the Capital Asset Pricing Model (CAPM) of Sharpe (1962) and its relations between variables: the market line.

$$y_t = \alpha + \beta . x_t + \varepsilon_t \quad (1)$$

$y_t \; and \; x_t$are stationary stochastic processes illustrating respectively the assert returns and market returns and ε_t is a i.i.d.(0,σ_ε) process.

The parameter β estimated by OLS is BLUE estimators by hypothesis and represents the systematic risk of the asset. To establish this relationship, we use the returns of AXA (French listed insurance companies) and the CAC40 index as market reference for the daily period from 2005 to 2015.

The CAPM model supposes that the risk estimated by Beta is stable and constant across time and whatever the investment horizon. It doesn't consider the time variability (or changings) of systematic risk and the behavioral hypothesis about heterogeneity of agents. By using rolling regression, we can appreciate the risk across time and thanks the wavelets we are able to estimate it on several frequencies.

The wavelets method is mathematical tool developed to reduce the time or frequency arbitrage imposed by the Fourier Transform. Wavelets decomposition is based on a particular function $\psi(t)$ called Wavelet-Mother with zero-mean et energy/variance preserving property.

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{s}} \psi \left(\frac{t-\tau}{s} \right) \quad (2)$$

The wavelet-mother is shifted by τ and dilated by s to provide the wavelet family $\psi_{\tau,s}(t)$ (equivalent to the filter).

The wavelets coefficients is calculated by the following equation:

$$W(s,\tau) = \int_{-\infty}^{+\infty} x_t \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-\tau}{s} \right) dt = \langle x_t, \psi_{\tau,s} \rangle \quad (3)$$

$\psi^* \left(\frac{t-\tau}{s} \right)$ is the complex conjugate of $\psi_{\tau,s}(t)$.

It is continuous wavelets decomposition (CWT) but the discrete decomposition (DWT) is easiest to implement because it is based on dyadic scale (the length of series N must be o multiple of 2). In this paper, we use a particular version of DWT called

Maximal Overlap Discrete Wavelet Transform (MODWT) having supplementary properties as a better variance analysis without the dyadic constraint. The decomposition is realized by a successive filtering of the series and wavelets coefficients with a rescaling (and subsampling). This is the Cascade Algorithm of Mallat (1989, 2009). After the decomposition, we have a frequency repartition of the series on frequency bands D_j for j=1,….J linked with a time horizon (see appendix table 1). The maximal number of frequency bands J is calculated by the following formula:

$$J = \frac{Ln(N)}{Ln(2)} \quad (4)$$

We can reconstruct the original series by summing the frequency bands:

$$x(t) = S_J + \sum_{j=1}^{j=J} D_j \quad (5)$$

The series is a sum of Trend Approximation S_J and frequency bands with different resolution level (different investment horizon). The discrete decomposition provides sub-chronic for which it is possible to estimate the Beta and apply rolling window. Then, we use the Forward Rolling Regression with a constant window with a length equal to 260 corresponding to one trading years. We define the following Betas:

- β_r represents the Rolling Betas estimated without the frequency decomposition.
- β_F illustrates the Rolling Frequency Betas estimated in the rolling window applied on the frequency bands.
- β_{TF} are the Time-Frequency Rolling Betas estimated with wavelets at each sliding of the window. We verify that there is a concordance between the size of the regression and the number of points used in the decomposition. In the previous methods, this property is not totally respected, so it can bias the estimators.

The use of discrete wavelets associated with rolling windows presents two majors constraints: the loss of observations related to the size of the window (260 in our case) and the loss of frequency information related to the discretization of the frequency bands. The wavelets gather the frequencies in bands so it is equivalent to aggregate different behaviors within this one. On the other hand, Continuous wavelet decompositions cover a wider frequency range and treat them separately.

Regression between frequency bands is possible in the discrete case because they are assimilated as time series. At the opposite, in the continuous case, the process is complexified by the different filters, but many tools are developed, as the wavelets coherence-phase, to measure the squared-correlation, its sign and sense. These tools are a

translation of the spectral analysis methods to the wavelets framework. In this case, the cross-wavelets spectrum defines the covariance between two time series and the auto-power spectrum represents a wavelets variance. It is also relatively easy to define the wavelet coherence as the ratio of the cross-spectrum on the auto-power spectra. The phase is the arc-tangent of the real and imaginary part ratio of the cross-spectrum. For more details about these tools, we can see the papers of Auth (2013), Bekiros et al (2016) and Mestre and Terraza (2017).

From these tools we develop a time-frequency estimator of the Beta $\beta_{\tau,s}$:

$$\beta_{\tau,s} = \theta_{\tau,s} * \sqrt{WQ_{\tau,s}} * \frac{G(s^{-1}|SW_x|^2)^{1/2}}{G(s^{-1}|SW_y|^2)^{1/2}} \quad (6)$$

$WQ_{\tau,s}$ is the wavelets coherence, $SW_i \; for \; i = x, y$ the power spectrum of the series i and $\theta_{\tau,s}$ a phase parameter indicating the sign of the correlation. G is a time-frequency smoothing operator used to have real values.

3 Time-Frequency Beta Estimation:

In this part, we use the previous methodologies based on discrete and continuous wavelets to estimate different kind of betas across time and frequency and compare them each other.

3.1 Static OLS Beta estimates

Previously, we check the stationary of the returns (see Appendixes A2) and then we estimate the Beta by OLS. The OLS beta is called ‘’Classic Beta’’ because it is non-varying parameter supposed constant over the period. We also discreetly decompose our series with discrete wavelets and we estimate an OLS Beta for each frequency bands. Results are recorded in Table 1. The estimated parameters confirm the nullity of the constant and we note that all Betas and R^2 are significantly different to zero.

However according to the tests, the residuals are not a white-noise process. Consequently, the Minimal Variance property of BLUE estimator is not respected in our case, so the beta could be instable over time.

Table 1: Static Beta Estimates

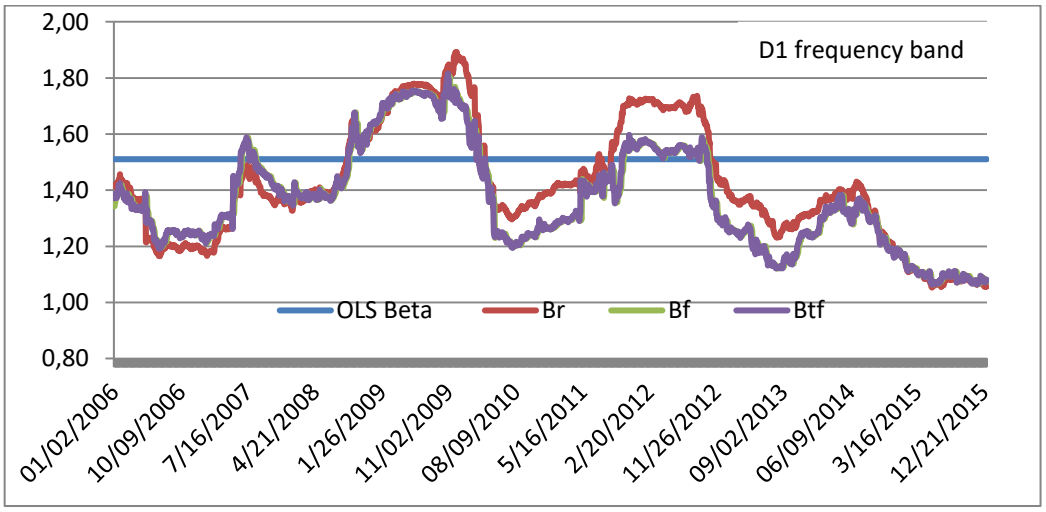
Axa	D1	D2	D3	D4	D5	D6	Overall Period
Beta	1.45	1.56	1.54	1.61	1.52	1.77	1.506
T-stats	78.36	80.39	75.8	79.97	60	85.1	77.98
Constant	0	0	0	0	0	0	0.000425
T-stats	0	0	0	0	0	0	1.54
R^2	0.68	0.69	0.67	0.69	0.56	0.72	0.68
LB	1050	2043	4678.9	5801	10905	13225	21.068
ARCH	621	792	1547	2664	2850	2864	62.132
JB	11765	10145	3695	1192	6617	2374	41993.2

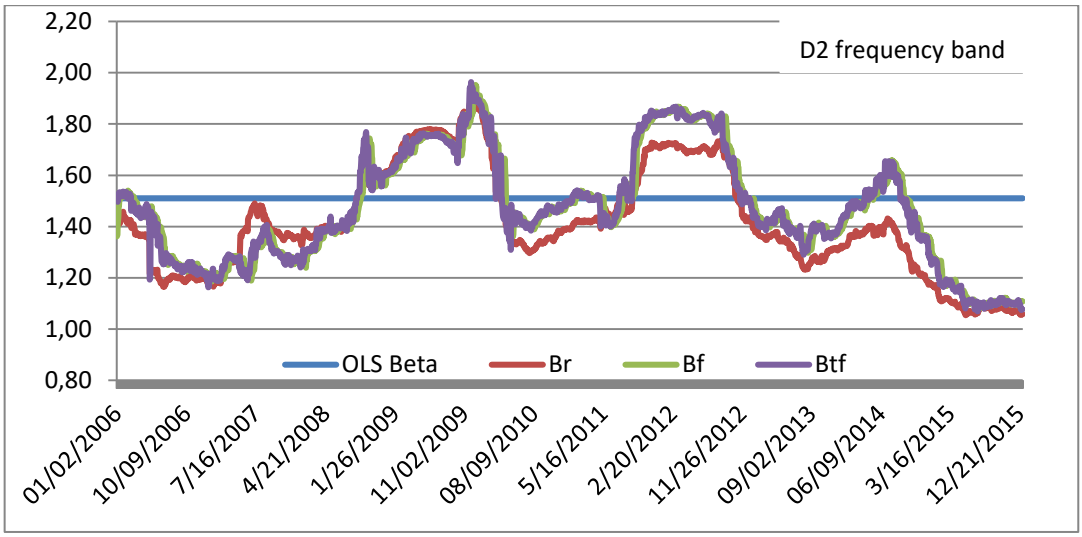
At 5% risk level, Colum LB (Ljung-Box test): $\chi^2(5)=11.1$; Colum ARCH (ARCH-LM test): $\chi^2(2)=5.99$; Colum JB (Jarque-Bera normality test): $\chi^2(2)=5.99$

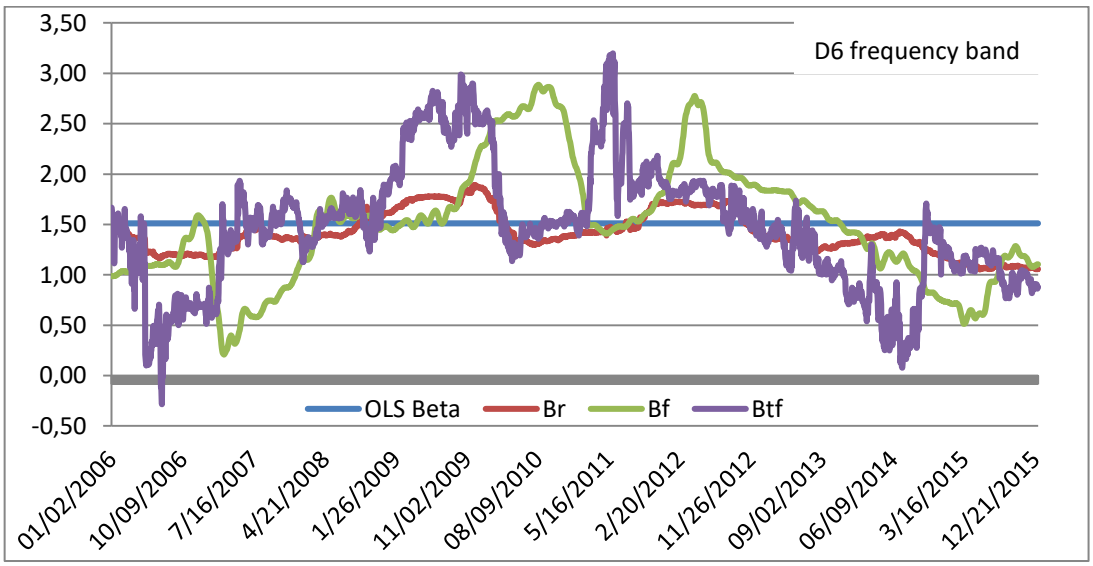
3.2 Rolling OLS Beta estimates β_r, β_F, β_{TF}

We apply the three previous methodologies in order to estimates a time-varying Betas and modelize it instability. Graphics 1 illustrates our results. We highlight the Beta volatility around its static estimator and we note a frequency differentiation of the beta dynamics.

Graphs 1: Rolling Betas of AXA stock







We notice that the short-run Beta (D1 and D2 bands) are volatile around 1.51 (OLS Beta) but they are always greater than one so the initial risk-profile is preserved across time only the intensity of the risk is varying. At the opposite, the long-run betas (D6 band) are more volatile than short-run Betas. Consequently, we note that the initial risk-profile is not the same across time.

We also observe a scale increasing gap between β_F and β_{TF}. The differences between these Betas increase with the frequency scale (time horizon larger). To confirm these observations, we test if the betas differences are significant (or not) and we calculate its Absolute Mean (see Table2).

Table 2: Percentages of Beta significantly different and Absolute Mean of Errors.

% of differents betas	D1	D2	D3	D4	D5	D6
β_r and β_F	7.93	6.82	32.39	43.5	64.66	67.27
β_r and β_{TF}	7.47	4.14	34.88	42.62	59.91	66.27
β_F and β_{TF}	0.04	0.5	6.13	26.14	48.37	87.27

MAE	D1	D2	D3	D4	D5	D6
β_r and β_F	0.07	0.08	0.13	0.18	0.33	0.39
β_r and β_{TF}	0.07	0.08	0.11	0.17	0.28	0.35
β_F and β_{TF}	0.01	0.03	0.06	0.13	0.3	0.59

($MAE = \frac{1}{T}\sum_{i=1}^{T}|\widehat{\beta}_i - \widehat{\beta}_j|$ for i = r, F, TF and j = r, F, TF with i ≠ j T = number of betas)

We note that the differences between β_F and β_{TF} (the two discrete methods) are more important at long-run (low-frequencies bands) whereas there are no significant differences at short-run (high-frequencies). We observe similar results for β_r and β_F and β_r and β_{TF} differences.

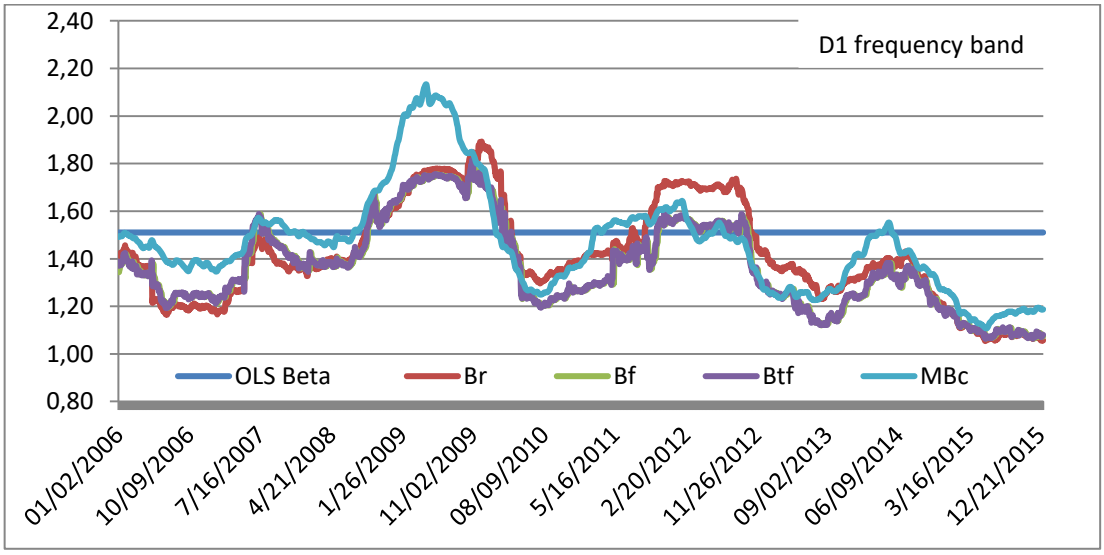
Globally, the differences become relatively important starting D3 bands, these results confirm our previous graphical observations: the estimation gap between β_F and β_{TF} increases as the frequency scale increases. This result is explained by Rolling Windows applied on discrete wavelet bands. This method doesn't consider the adequation between the window size/ length (260 points) and the number of point used during the decomposition (2868 points in our case), so it supposes a unique wavelet decomposition. In this case, the Window cuts the wavelet outputs. At the opposite the Time-Frequency window realize a wavelet transform inside the window (at each iteration) so the window length and the number of point used in the decomposition are equal. In this case, the wavelets coefficients are rescaled according to the window. Moreover, with our results, we can't fully determinate the best methods because we don't have a frequency benchmark. It is that why, we use a third method in order to find similarities between them.

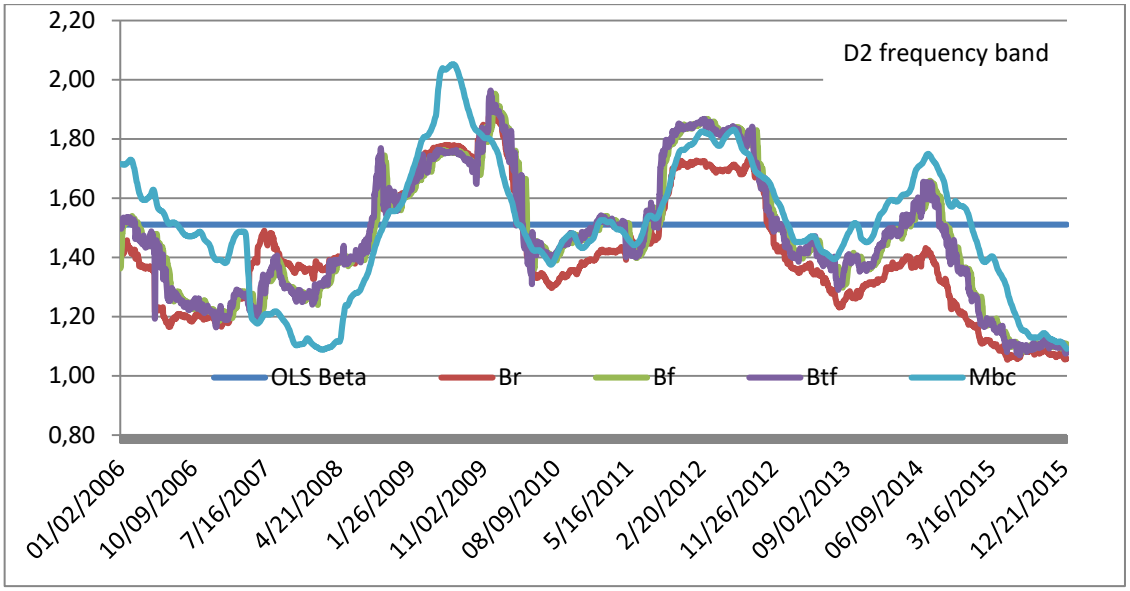
3.3 Time-Frequency Betas estimates by continuous wavelets.

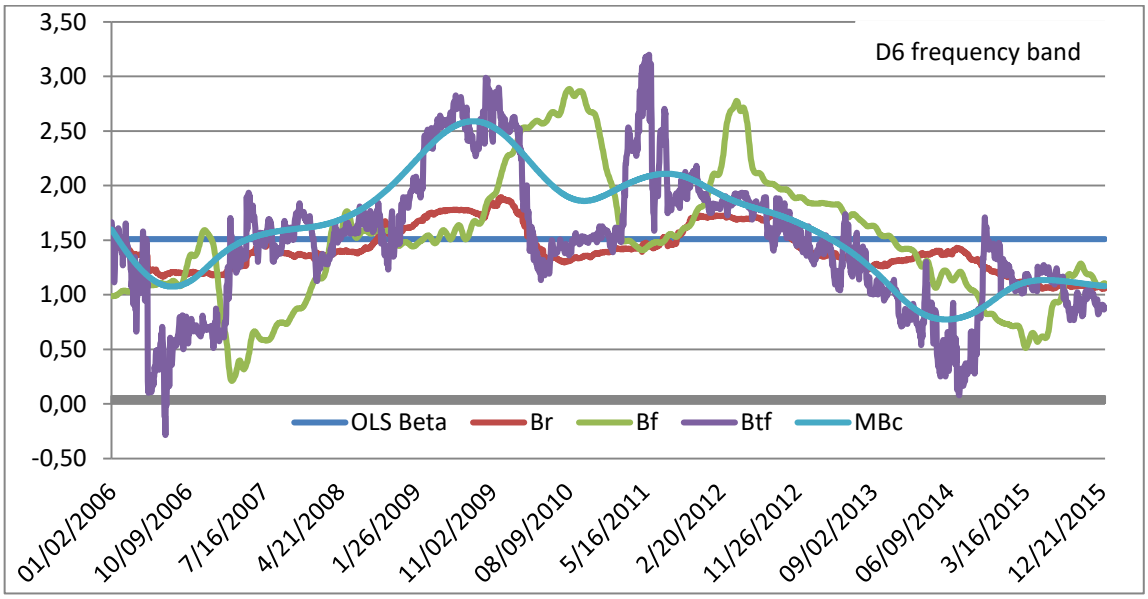
The betas estimates across time and frequency by using continuous wavelets don't directly use rolling window. By decomposing the series with CWT (Continuous Wavelets Transforms), we can estimate 2868 betas for each frequency noted $\beta_{\tau,s}$. τ represents the time localisation and s the fréquence. In order to compare the $\beta_{\tau,s}$ with β_F and β_{TF}. To simplify the analysis, we consider only the D1, D2 and D6 frequency bands corresponding respectively to short-run investment (2 – 4 days), medium-short run (1- 2 weeks) and medium-long run (3 – 6 months).

To compare the betas, we need to realize some adjustments to have a same comparative base. β_F and β_{TF} are estimated for a frequency bands regrouping several frequencies (discrete case) whereas $\beta_{\tau,s}$ is estimated for a specific frequency s (continuous case). Consequently, we calculate the $\beta_{\tau,s}$ frequency Mean (noted MBc) for the frequencies included in the bands (D1, D2 and D6). Another method is to consider the frequency s corresponding to the middle of the frequency bands (3 for D1, 6 for D2 and 96 for D6). Moreover, we have β_F and β_{TF} estimated in window of 260 points and $\beta_{\tau,s}$ estimated at each time, so there is no time adequation. Consequently, we calculate the time rolling mean of $\beta_{\tau,s}$ with 260 lengths to guarantee the time coherence of the comparison. The results are illustrated by the graphics 2.

Graphs 2: Time-Frequency Betas of AXA stock




D2 frequency band
2,20
2,00
1,80
1,60
1,40
1,20
1,00
0,80
OLS Beta
Br
Bf
Btf
Mbc
01/02/2006
10/09/2006
7/16/2007
4/21/2008
1/26/2009
11/02/2009
08/09/2010
5/16/2011
2/20/2012
11/26/2012
09/02/2013
06/09/2014
3/16/2015
12/21/2015


D6 frequency band
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
-0,50
OLS Beta
Br
Bf
Btf
MBc
01/02/2006
10/09/2006
7/16/2007
4/21/2008
1/26/2009
11/02/2009
08/09/2010
5/16/2011
2/20/2012
11/26/2012
09/02/2013
06/09/2014
3/16/2015
12/21/2015

On low-frequencies (D1-D2 bands), $\beta_{\tau,s}$ are not totally similar to β_F and β_{TF} but it has the same dynamics (similar evolutions). On high-frequencies, $\beta_{\tau,s}$ are less volatile and erratic than β_F and β_{TF} but its dynamics is similar to $\beta_{\tau,s}$. To improve our analysis, we calculate the Absolute Mean of the errors between the different betas estimators (see Table 3).

Table 3: Absolute Mean of Estimations Errors (MAE)

MAE between	$\beta_{\tau.s}$ D1	$\beta_{\tau.s}$ D2	$\beta_{\tau.s}$ D6
β_r	0.118	0.159	0.327
β_F	0.110	0.108	0.454
β_{TF}	0.109	0.111	0.252

We note that the differences are similar at low-frequencies, but at high-frequencies, the gap between $\beta_{\tau.s}$ and β_{TF} are lesser than the $\beta_{\tau.s} - \beta_F$ differences. This result confirms our hypothesis: at low-frequencies the two discrete methods and the continuous wavelets estimations are similar whereas at high-frequencies the differences are more important. The time-frequency window and the continuous wavelets method provide Betas with similar dynamics even if significant differences are observed. These differences could be explained by the adjustments used to compare Betas. Our results indicate that the Time-Frequency Rolling Window is more appropriate (than it intuitive approach) to estimate a time-varying parameter according to the investment horizon. The Continuous Wavelets Methods is more accurate because it considers more frequencies (opposite to a large scale) and it provide Betas at each time but the computational effort is greater.

4 Conclusion

In this paper, we want associate the wavelets with Forward Rolling Estimation. We consider an intuitive approach: apply the rolling window on the discrete wavelets filter outputs. Our hypothesis is that this method is not fully compatible with wavelets properties so it is necessary to realize the DWT inside the window. To test this hypothesis, we use the CAPM framework and the AXA stock in order to estimate its rolling beta (time-varying parameter).

Our results show that there are no significant differences between the short-run betas of the two discrete methods, but the errors increase when the frequency scale increase (from low-frequencies to high-frequencies). The intuitive method implies a unique wavelet decomposition based on 2868 points whereas the Beta estimation is realized on 260 points (window length). Consequently, the rescaling of wavelets coefficient is operated on 2868 points and it doesn't consider the window size. At the opposite,

time-frequency window decomposes the series inside the window and estimate the beta simultaneously, so the rescaling is effectuated at each iteration.

Without frequency benchmark, it is impossible to determinate the most adapted method. In this case, we developpe a third method based on continuous wavelets decomposition not totally based on a rolling window. By comparing the results, we remark that the $\beta_{\tau.s}$ and the β_{TF} share the same evolutions (similar dynamics). We can conclude, that the three methods have similar results if we consider short-run investment, but if we consider long-run investment, the time-frequency window and the continuous approach provide more relevant results.

5 Appendixes

Table A1: Frequency bands and time horizon

	Frequency days		Months	
Resolution	Inf border	Sup border	Inf border	Sup border
D1	2	4	0.09	0.18
D2	4	8	0.18	0.36
D3	8	16	0.36	0.73
D4	16	32	0.73	1.45
D5	32	64	1.45	2.91
D6	64	128	2.91	5.82
D7	128	256	5.82	11.64
D8	256	512	11.64	23.27
D9	512	1024	23.27	46.55
D10	1024	2048	46.55	93.09
D11	2048	4096	93.09	186.18
S11	4096	-	186.18	-

Table A2: Stationary tests

	Philips-Perron test	KPSS-test
AXA	-51.2187	0.074
CAC	-56.1127	0.106
	Critical Value	
1% level	-3.43	0.216
5% level	-2.86	0.146
10% level	-2.57	0.119

6 References

Auth C., “Continuous Wavelet Transform and Wavelet Coherence-Implementation and Application to the Diversification Analysis of Hedge Funds Returns”, 2013.

Bekiros S., Nguyen D.K. and Uddin G.S, “On time scale behavior of equity commodity links: implications for portfolio management”, Journal of International Financial Market Institutions and Money, vol 41, 2016, pp 79-121.

F.Black, M. Jensen, M. Scholes, “The Capital Asset Pricing Model: Some Empirical Test ; Studies in the Theory of Capital Markets” in M. Jensen ed., New York: Praeger Publishers, 1972 , pp. 79–12.

Brooks R.D., R.W, Faff and J.H.H Lee, “The form of time variation of systematic risk: some Australian evidence”, Applied financial econometrics, Vol 2, 1992, pp 191-198.

Brooks, R.D., R.W. Faff, and M.D. McKenzie, “Time-varying Beta Risk of Australian Industry Portfolios: A Comparison of Modelling Techniques”, Australian Journal of Management 23(1), 1998, pp 1–22.

R.W. Faff, J.H.H Lee and T.R.L Fry, “Stationarity of systematic risk : some Australian evidence”, Journal of Finance and Accounting, Vol 19, Issue 2, 1992 ; pp 253-270.

R.W. Faff, R.D. Brooks, “Time-varying risk for Australian industry portfolios, an exploratory analysis”, Journal of Business Finance and Accounting, Vol 25, Issue 5, 1998, pp 721-74.

E. Fama; J. MacBeth, “Risk, Return, and Equilibrium: Empirical Tests” Journal of Political Economy , Vol 81, 1973, pp. 607–636.

R. Gençay, F. Selçuk, B. Whitcher, “Systematic Risk and Timescales “, Quantitative Finance, Vol 3 issue 2, 2003,pp 108-116.

N. Groenewold and P. Fraser, “Time Varying Betas and Macroeconomic influences”, discussion paper, departement of economics, University of Western Australia, 1997.

S. Mallat, “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation” ;IEEE Transactions on Pattern Analysis and Machine Intelligence Vol 11 num 7 July 1989.

S. Mallat, Une exploration des signaux en ondelettes, Ecole polytechnique, 2009.

S. Mallat, Wavelet tour of signal processing: the sparse way, Academic Press, 2009.

Mestre R. and Terraza M., "Time-frequency analysis of CAPM-Application to the CAC 40", Managing Global Transitons Internationnal Research Journal, 2018.

Mestre R. and Terraza M., "Multidimensional Time-Frequency Analysis of the CAPM", ITISE 2017 Conference Proceedings, Issue 2, pp 1187-1200, 2017.

Mestre R. and Terraza M., "Time-Frequency varying Beta Estimation-A continuous wavelets approach-", MPRA Paper No. 86335 December 2017.

W. Sharpe, "Capital Asset Prices: a Theory of Market Equilibrium under risk"; Journal of Finance, Vol. 19, No. 3 (Sep., 1964), 1964, pp. 425-442.

J.Yeo, "Modelling Time-Varying Systematic Risk in Australia"; Proceedings of the International Congress on Modelling & Simulation. ed. / F. Ghessemi; M. McAleer; L. Oxley; M. Scoccimarro. Vol. 3, Canberra. ed. Western Australia : Modelling and Simulation Society of Australia and New Zealand Inc., 2001, p. 1565-1570.

Using subspace methods to model long memory processes

Dietmar Bauer[1]

Econometrics, Bielefeld University, Universitätsstrasse 25, 33615 Bielefeld, Germany,
Dietmar.Bauer@uni-bielefeld.de

Abstract. Subspace methods have been shown to be remarkably robust procedures providing consistent estimates of linear dynamical state space systems for multivariate time series in different situations including stationary and integrated processes without the need for specifying the degree of persistence.
Fractionally integrated processes bridge the gap between short-memory processes corresponding to stable rational transfer functions and integrated processes such as unit root processes. Therefore it is of interest to investigate the robustness of subspace procedures for this class of processes.
In this paper it is shown that a particular subspace method called canonical variate analysis (CVA) that is closely related to long vector autoregressions (VAR) provides consistent estimators of the transfer function corresponding to the data generating process also for fractionally integrated processes of the VARFIMA or FIVARMA type, if integer parameters such as the system order tend to infinity as a suitable function of the sample size. Additionally also almost sure orders of convergence are derived that depend on the degree of persistence.
The results are based on analogous statements for the consistency of long VAR modelling.

Keywords: long memory, fractional integration, subspace algorithms, long VAR models

1 Introduction

For the estimation of linear dynamic models of the VARMA (vector autoregressive moving average) type for multivariate times series the so-called subspace methods have been shown to provide attractive features in many different settings (for a review see e.g. [1]): In the stationary case a particular type of algorithm termed *canonical variate analysis* (CVA) [9][1] provides estimators of rational transfer functions that are asymptotically equivalent to quasi-maximum likelihood estimators. Hence in the case of Gaussian innovations efficient estimators are obtained for known order of the data generating system. Consistency for this

[1] This algorithm has also been called *canonical correlation analysis* (CCA) in the literature.

class of estimators has also been established for integrated processes with a unit root at $z = 1$, that is for $I(1)$ processes (consistency for an adaptation is shown in [2], consistency of `CVA` is claimed in [1], the corresponding result has not yet been published).
Beside the asymptotic properties the main appeal of subspace methods lies in their conceptual simplicity: they can be interpreted as applying model reduction to an initial high order vector autoregression (in the following 'long VAR') estimate. Thus their properties are in many cases easy to understand, problems in the data such as missing values (for example due to eliminated outliers) can be handled easily, the effects of demeaning and detrending are in most cases straightforward to understand. Additionally also numerically efficient implementations are heavily used.
As a final advantage within the algorithm also information on the suitability of the system order used for estimation is obtained from the reduced rank regression step. In this step the order can be estimated or tested for.
As `CVA` provides consistent transfer function estimates for stationary short-memory processes as well as for integrated processes, it appears tempting to investigate the asymptotic properties of system estimators using `CVA` in the setting of fractionally integrated processes. In this respect the main contribution of this paper is to establish consistency of the transfer function estimators when the order of the system tends to infinity at an appropriate rate and the data generating process is a fractionally integrated process of the VARFIMA (vector autoregressive fractionally integrated moving average) or FIVARMA (fractionally integrated vector autoregression moving average) type. The order of the fractional integration here may differ for each component of the process in the range $0 \leq d_j < 0.5$ or $-0.5 < d_j \leq 0$. Based on consistent estimation (including upper bounds for the order of convergence) for the underlying transfer function initial guesses for subsequent maximum likelihood estimation can be obtained.
Due to the intimate relation between `CVA` and long VAR modelling this paper is largely based on and slightly extends the knowledge on the properties of long VAR estimators in the setting of vector processes with different fractional integration orders d_j.
The organisation of the paper is as follows: In the next section a brief description of the `CVA` algorithm is provided. The main results of the paper are contained in Section 3. Section 4 concludes the paper. Proofs are delegated to an appendix.

2 The CVA subspace algorithm

Subspace algorithms are used for the identification of linear dynamic systems for multivariate time series $(y_t)_{t\in\mathbb{Z}}, y_t \in \mathbb{R}^s$ in state space representation:

$$x_{t+1} = Ax_t + K\varepsilon_t \ , \ y_t = Cx_t + \varepsilon_t \tag{1}$$

where $(x_t)_{t\in\mathbb{Z}}, x_t \in \mathbb{R}^n$ denotes the unobserved state process and the matrices $A \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{s\times n}, K \in \mathbb{R}^{n\times s}$. The process $(\varepsilon_t)_{t\in\mathbb{Z}}$ denotes the s-dimensional innovation process which is a white noise. In this paper we consider $(\varepsilon_t)_{t\in\mathbb{Z}}$ as

an independent identically distributed (iid) sequence.
The stability assumption $|\lambda_{max}(A)| < 1$ (the largest modulus of all eigenvalues of A is smaller than 1) implies that there exists a stationary solution to (1):

$$y_t = \sum_{i=0}^{\infty} k_i \varepsilon_{t-i} = \varepsilon_t + \sum_{i=1}^{\infty} CA^{i-1}K\varepsilon_{t-i}.$$

The impulse response coefficients $k_i = CA^{i-1}K, i \in \mathbb{N}$ corresponding to the system (A, C, K) are related to the rational transfer function

$$k(z) = I_s + \sum_{i=1}^{\infty} CA^{i-1}Kz^i = a^{-1}(z)b(z)$$

for some polynomial matrices $a(z) \in \mathbb{R}^{s\times s}, b(z) \in \mathbb{R}^{s\times s}$ building a VARMA representation of the state space system. Without restriction of generality we will only consider minimal state space representations (cf. chapter 2 of [6]).
The main idea for estimation using subspace methods is to note that the state equations (1) are linear in the system matrices for known state. Thus if an estimate of the state is available, the system can be estimated using least squares. An estimate for the state is obtained from the equation (see [1] for details):

$$Y_{t,f}^+ = \mathcal{O}_f x_t + \mathcal{E}_f E_{t,f}^+ = \mathcal{O}_f \mathcal{K}_p Y_{t,p}^- + \mathcal{O}_f \underline{A}^p x_{t-p} + \mathcal{E}_f E_{t,f}^+ = \beta_1 Y_{t,p}^- + N_{t,f}^+. \quad (2)$$

Here $Y_{t,f}^+ := [y_t', y_{t+1}', \ldots, y_{t+f-1}']', E_{t,f}^+ := [\varepsilon_t', \varepsilon_{t+1}', \ldots, \varepsilon_{t+f-1}']'$ for some integer f, $Y_{t,p}^- := [y_{t-1}', \ldots, y_{t-p}']', \mathcal{K}_p := [K, \underline{A}K, \underline{A}^2K, \ldots, \underline{A}^{p-1}K]$ for $\underline{A} := A - KC$ and $\mathcal{O}_f := [C', A'C', \ldots, (A^{f-1})'C']'$. Expressions for $\mathcal{E}_f$ are given in [1].
In the following we use $\langle a_t, b_t\rangle := T^{-1}\sum_{t=p+1}^{T-f+1} a_t b_t'$ for sequences $(a_t)_{t\in\mathbb{N}}$ and $(b_t)_{t\in\mathbb{N}}$. Then `CVA` proceeds as follows:

1. Specify the integer values f, p.
2. Solve the rank restricted regression problem $Y_{t,f}^+ = \beta_1 Y_{t,p}^- + N_{t,f}^+$ under the rank constraint $\text{rank}(\beta_1) = n$ (to be specified in this step) to obtain an estimate $\hat{\mathcal{O}_f}\hat{\mathcal{K}_p} := [(\hat{\Xi_f})^{-1}\hat{U}_n\hat{S}_n][\hat{V}_n'\hat{\Xi_p^-}]$ of β_1 using the SVD (singular value decomposition)

   $$\hat{\Xi_f}\hat{\beta}_1\hat{\Xi_p^-} = \hat{U}\hat{S}\hat{V}' = \hat{U}_n\hat{S}_n\hat{V}_n' + \hat{R}_n, \quad \hat{\Xi_f} := \langle Y_{t,f}^+, Y_{t,f}^+\rangle^{-1/2}, \quad \hat{\Xi_p^-} := \langle Y_{t,p}^-, Y_{t,p}^-\rangle^{1/2}.$$

   Here $\hat{\beta}_1 = \langle Y_{t,f}^+, Y_{t,p}^-\rangle\langle Y_{t,p}^-, Y_{t,p}^-\rangle^{-1}$, $\hat{U}_n \in \mathbb{R}^{fs\times n}$ denotes the matrix whose columns are the left singular vectors to the singular values which are the diagonal entries in[2] $\hat{S}_n := \text{diag}(\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_n), \hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_n > \hat{\sigma}_{n+1} > 0$ and $\hat{V}_n \in \mathbb{R}^{ps\times n}$ contains the corresponding right singular vectors as its columns. $\hat{R}_n$ denotes the approximation error.
3. Use the estimated state $\hat{x}_t := \hat{\mathcal{K}_p} Y_{t,p}^-, t = p+1, \ldots, T+1$ to obtain estimates $(\hat{A}, \hat{C}, \hat{K})$ of the system matrices from regression in the system equations.

[2] In the unlikely case of identically estimated singular values $\hat{\sigma}_n = \hat{\sigma}_{n+1}$ the basis in the corresponding spaces is chosen randomly.

The choice of f and p influences the asymptotic properties of the estimators. For fractionally integrated processes we will let $f = p$ as well as the order n tend to infinity at a rate to be defined later on.
Note that for $f = p, n = sp$ typically the then square matrix $\hat{\mathcal{K}_p}$ will be regular and hence the estimated state equals $Y_{t,p}^-$ subject to a basis change. It then follows that the estimated transfer function coincides with the VAR(p) estimate. The `CVA` estimator for $n < sp$ then can be interpreted as a particular method of model reduction starting from the estimated long VAR model.

3 Results for fractionally integrated processes

In this paper we consider fractionally integrated processes of the VARFIMA or FIVARMA type (cf. e.g. [13]):

Definition 3.1 *The process $(y_t)_{t\in\mathbb{Z}}$ is called a* vector autoregressive fractionally integrated moving average (VARFIMA) *process if it is obtained as the stationary solution to the vector difference equation*

$$a(L)y_t = b(L)e_t, \qquad e_t = D(L)^{-1}\varepsilon_t$$

where $(a(z), b(z))$ is a stable and invertible VARMA system of left coprime polynomial matrices $a(z) = I_s + a_1 z + ... + a_p z^p, b(z) = I_s + b_1 z + ... + b_q z^q$, $D(L) = diag[(1-L)^{d_j}]_{j=1,...,s}$ is a diagonal matrix of fractionally integrating filters where $|d_j| < 0.5, max_{j=1,...,s}|d_j| > 0$ and where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an iid white noise sequence. L denotes the backward shift operator.

Thus a VARFIMA process applies a VARMA filter to a fractionally integrated noise sequence. The FIVARMA model reverses the order by filtering a VARMA process with a fractionally integrating filter:

Definition 3.2 *The process $(y_t)_{t\in\mathbb{Z}}$ is called a* fractionally integrated vector autoregressive moving average (FIVARMA) *process if it is obtained as the stationary solution to the vector difference equation*

$$y_t = D(L)^{-1}u_t, \qquad a(L)u_t = b(L)\varepsilon_t,$$

where $(a(z), b(z))$ is a stable and invertible VARMA system of left coprime polynomial matrices $a(z) = I_s + a_1 z + ... + a_p z^p, b(z) = I_s + b_1 z + ... + b_q z^q$, $D(L) = diag[(1-L)^{d_j}]_{j=1,...,s}$ is a diagonal matrix of fractionally integrating filters where $|d_j| < 0.5, max_{j=1,...,s}|d_j| > 0$ and where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an iid white noise sequence. L denotes the backward shift operator.

In the scalar case $s = 1$ the two definitions are identical as in this case the operators $D(z)$ and $a(z)^{-1}b(z)$ commute, whereas in the vector case they are different if not all $d_j = d$ are identical (compare [13]). Note also that $|d_j| < 0.5$ contains the case that some (but not all) d_j are zero. Therefore the processes may contain a mixture of long- and short-memory processes. We will always assume that there exists at least one integration order $d_j \neq 0$ to rule out the case of exclusively short-memory processes, for which the properties of `CVA` are well known. [7] derived a number of common properties:

Theorem 3.1. *Let $(y_t)_{t\in\mathbb{Z}}$ be a stationary FIVARMA or VARFIMA process with fractional integration orders $d_1, ..., d_s$. Then*

(i) $y_t = \sum_{j=0}^{\infty} k_j \varepsilon_{t-j}$ *where* $\|k_j\| \leq M_k j^{d_+ - 1}$ *where* $d_+ = \max(d_1, ..., d_s; d_j \neq 0) < 0.5$.

(ii) $y_t = \varepsilon_t + \sum_{j=1}^{\infty} \phi_j y_{t-j}$ *where* $\|\phi_j\| \leq M_\phi j^{-1-d_-}$ *where* $d_- = \min(d_1, ..., d_s; d_j \neq 0) > -0.5$.

Thus both an AR(∞) and an MA(∞) representation with square summable coefficients exist.

The theorem shows that the two orders d_- and d_+ determine the rate of decay of the coefficients in the AR and MA representations.

The processes included in the above setup typically do not allow for a state space representation with finite state dimension. It has been shown by [3] that fractionally integrated processes correspond to infinite state dimension while at the same time approximations with finite state dimension exist in the sense that allowing the state dimension to grow, a convergent sequence of transfer functions can be obtained. It is the main contribution of this paper to show that this holds for transfer functions estimated using `CVA` (for the proof see section A.2 in the appendix):

Theorem 3.2. *Let the process $(y_t)_{t\in\mathbb{Z}}$ be a VARFIMA or FIVARMA process generated according to Definition 3.1 or Definition 3.2 with d_- and d_+ defined in Theorem 3.1 where the white noise process $(\varepsilon_t)_{t\in\mathbb{Z}}$ is independent identically distributed with zero mean, non-singular variance Ω and finite fourth moments. Let $(\hat{A}_n, \hat{C}_n, \hat{K}_n)$ denote the* `CVA` *estimates using $f = p = p(T) \to \infty$ choosing a system order n. Further let $\phi(z) = I_s - \sum_{j=1}^{\infty} \phi_j z^j = k^{-1}(z), \hat{k}_n(z) = I_s + z\hat{C}_n(I_n - \hat{A}_n z)^{-1}\hat{K}_n, \hat{\phi}_n(z) = \hat{k}^{-1}(z)$.*

(I) If $0 < d_- \leq d_+ < 0.5$ and if $d_- > 2d_+ - 0.5$ then for $p = p(T) \to \infty$ such that $pR_T(d_+) \to 0$ (where $R_T(d) := (T/\log T)^{d-1/2}$) then there exists a choice $n = n(T)$ such that

$$\|\hat{\phi}_n(z) - \phi(z)\|_2 = o(R_T(d_+)p) + O(p^{2d_+ - d_- - 0.5}) = o(1).$$

where $\|.\|_2$ denotes the two norm in the space of functions defined on the unit circle.

(II) If $-0.5 < d_- \leq d_+ < 0$ and $2d_+ < 0.5 + 3d_-$ and if $p = p(T) \to \infty$ such that $p = o((T/\log T)^{1/(2-8d_-)})$ then there exists a choice $n = n(T)$ such that

$$\|\hat{\phi}_n(z) - \phi(z)\|_2 = o((\log T/T)^{1/2} p^{1-4d_-}) + O(p^{2d_+ - 3d_- - 0.5}) = o(1).$$

The theorem allows for different d_- and d_+, however both must have the same sign. Additional short-memory components of the VARMA type are allowed for (that is some of the d_j but not all can be zero; of course the case that all d_j are zero is well covered in the literature). The proof of the theorem extends the current knowledge on long VAR approximation in the long memory setting as it implies that for appropriate increase of p as a function of the sample size autoregressions can be used in order to approximate the data generating process in the FIVARMA and VARFIMA case for different values d_j.

4 Conclusions

This paper shows that (under appropriate assumptions on the integration orders) the transfer function corresponding to a VARFIMA or FIVARMA process can be estimated consistently using the `CVA` approach when the maximal lag order p tends to infinity as a function of the sample size. This adds fractionally integrated processes to the list of settings in which `CVA` provides useful outcomes.
In the rational case, where the data generating process corresponds to finite n, choosing the correct order reduces the noise and leads to better estimators. Also in the fractionally integrated case one might suspect that using a smaller n is beneficial in terms of accuracy. Preliminary simulations verify this in simple settings. Unfortunately the proof does not give much intuition on how to best choose n in this setting. This is left for future research.

Acknowledgments

The author is grateful to Philipp Sibbertsen and Christian Leschinski for discussions and comments on the contents of the paper.

References

1. Bauer, D.: Estimating Linear Dynamical Systems Using Subspace Methods. Econometric Theory 21, 181–211 (2005)
2. Bauer, D., Wagner, M.: A Canonical Form for Unit Root Processes in the State Space Framework. Econometric Theory 6, 1313–1349 (2012)
3. Chan, N.H., Palma, W.: State Space Modeling of Long-Memory Processes. The Annals of Statistics 26, 719–740 (1998)
4. Dahlen, A., Scherrer, W.: The Relation of the CCA Subspace Method to a Balanced Reduction of an Autoregressive Model. Journ. of Econometrics 118, 293–312 (2004)
5. Galbraith, J.W., Zinde-Walsh, V.: Autoregression-Based Estimators for ARFIMA Models. CIRANO Working Papers 2001s-11 (2001)
6. Hannan, E.J., Deistler, M.: The Statistical Theory of Linear Systems. John Wiley, New York (1988)
7. Hosking, J.R.M.: Fractional Differencing. Biometrika 68(1), 165–176 (1981)
8. Hosking, J.R.M.: Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. Journal of Econometrics 73(1), 261–284 (1996)
9. Larimore, W.E.: System Identification, reduced order filters and modeling via canonical variate analysis. pp. 445–451. Proceedings of the 1983 American Control Conference, Piscataway, NJ (1983)
10. Lewis, R., Reinsel, G.: Prediction of Multivariate time Series by Autoregressive Model Fitting. Journal of Multivariate Analysis 16, 393–411 (1985)
11. Palma, W., Bondon, P.: On the Eigenstructure of Generalized Fractional Processes. Statistics and Probability Letters 65, 93–101 (2003)
12. Poskitt, D.S.: Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible case. Annals of Institute of Statistical Mathematics 59, 697–725 (2006)
13. Sela, R.J., Hurvich, C.M.: Computationally efficient methods for two multivariate fractionally integrated models. Journal Time Series Analysis 30(6), 631–651 (2009)

A Appendix

A.1 Auxiliary Lemmas

Lemma A.1. *Let* $y_t = \sum_{j=0}^{\infty} k_j \varepsilon_{t-j}$ *where* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is an i.i.d. sequence of random variables having zero mean and finite fourth moments. Let* $\hat{\gamma}_j = T^{-1}\sum_{t=1+j}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y})'$ *where* $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t, \tilde{\gamma}_j = T^{-1}\sum_{t=1}^{T} y_{t+j}y_t' - \bar{y}\bar{y}'$ *and* $\gamma_j = \mathbb{E}y_t y_{t-j}'$. *Assume that* $k_j = O(j^{d-1})$ *where* $-0.5 < d < 0.5$. *Then:*
(i) $\|\gamma_j\| \le \kappa j^{2d-1}$.
(ii) $\mathbb{E}\bar{y}\bar{y}' = O(T^{2d-1})$ *and* $\mathbb{E}\|\bar{y}\|^4 = O(T^{4d-2})$.
(iii) $\mathbb{E}\|\hat{\gamma}_j - \tilde{\gamma}_j\| \le \tau j/T$ *where* $0 < \tau < \infty$ *does not depend on* j.
(iv) $\hat{\gamma}_j - \gamma_j = O_P(j/T) + (\tilde{\gamma}_j - \gamma_j)$ *where* $\mathbb{E}vec(\tilde{\gamma}_j - \gamma_j)vec(\tilde{\gamma}_k - \gamma_k)' = O(Q_T(d)^2)$ *with*

$$Q_T(d) = \begin{cases} T^{2d-1} & , for \quad 0.25 < d < 0.5 \\ T^{-1/2}\log T & , \quad d = 0.25, \\ T^{-1/2} & , \quad -0.5 < d < 0.25. \end{cases}$$

(v) Let $R_T(d) := (T/\log T)^{d-1/2}$. *For* $H_T R_T(\tilde{d}) \to 0$ *for* $\tilde{d} = \max(0, d)$ *we have*

$$\max_{0\le j\le H_T} \|\hat{\gamma}_j - \gamma_j\| = O(R_T(\tilde{d})), \quad \max_{0\le j\le H_T} \|\tilde{\gamma}_j - \gamma_j\| = O(R_T(\tilde{d})).$$

Proof. The proof uses the results of Theorem 1, 3 and 5 of [8] and Theorems 1 and 2 of [12].
(i) Using $\Omega = \mathbb{E}\varepsilon_t\varepsilon_t'$ we have for some constant $0 < \kappa < \infty$ not depending on j

$$\|\gamma_j\| = \|\sum_{i=j}^{\infty} k_i \Omega k_{i+j}'\| \le \mu \sum_{i=0}^{\infty} \|k_i\|\|k_{i+j}\| \le \mu\bar{\mu}^2 \sum_{i=0}^{\infty} i^{d-1}(i+j)^{d-1} \le \kappa j^{2d-1}$$

since $\|\Omega\| < \mu, \|k_i\| \le \bar{\mu} i^{d-1}$ for some $\bar{\mu} < \infty$, $i < i + j$ and the techniques in the proof of Lemma 3.2 of [3].
(ii) The first part follows directly from (i) in combination with [8, Theorem 1] dealing with each coordinate separately.
Below we deal without restriction of generality with scalar processes. The vector case is merely notationally more complex. With respect to the second part note that $\mathbb{E}y_t y_s y_r y_0 = \gamma_{t-s}\gamma_r + \gamma_{t-r}\gamma_s + \gamma_t\gamma_{s-r} + \kappa_4(t, s, r)$ for $\kappa_4(t, s, r) := \sum_{a=-\infty}^{\infty} k_{a+t}k_{a+s}k_{a+r}k_a(\mathbb{E}\varepsilon_t^4 - (\mathbb{E}\varepsilon_t^2)^2)$ where for notational simplicity $k_a = 0, a < 0$ is used. Next

$$\mathbb{E}\bar{y}^4 = T^{-4}\sum_{t,s,r,u=1}^{T} \mathbb{E}y_t y_s y_r y_u = T^{-4}\sum_{t,s,r,u=1}^{T} \gamma_{t-s}\gamma_{r-u} + \gamma_{t-r}\gamma_{s-u} + \gamma_{t-u}\gamma_{s-r} + \kappa_4(t-u, s-u, r-u)$$

is the sum of four terms where the first three are identical:

$$T^{-4}\sum_{t,s,r,u=1}^{T} \gamma_{t-s}\gamma_{r-u} = \left(T^{-2}\sum_{t,s=1}^{T} \gamma_{t-s}\right)^2 = (\mathbb{E}\bar{y}^2)^2 = O(T^{4d-2}).$$

The last term equals the fourth term of (A.2) in [8]. Using Lemma 3.2. (i) of [3] in the fourth row of the equation on p. 277 of [8] we obtain

$$T^{-4}\sum_{t,s,r,u=1}^{T}\kappa_4(t-u,s-u,r-u)=O(T^{2d-2})$$

for $0.25\leq d<0.5$. By dominated convergence we have that this term is $O(T^{-1})$ for $-0.5<d<0.25$. Hence we obtain the bound $\mathbb{E}\bar{y}^4=O(T^{4d-2})$ for $d>0.25$ and $O(T^{-1})$ else.

$$\begin{aligned}(iii)\quad \hat{\gamma}_j &= \frac{1}{T}\sum_{t=1+j}^{T}(y_t-\bar{y})(y_{t-j}-\bar{y})' = \frac{1}{T}\sum_{t=1+j}^{T}(y_t y_{t-j}' - \bar{y}y_{t-j}' - y_t\bar{y}' + \bar{y}\bar{y}')\\ &= \frac{1}{T}\sum_{t=1+j}^{T}y_t y_{t-j}' - \bar{y}(\frac{1}{T}\sum_{t=1+j}^{T}y_{t-j})' - (\frac{1}{T}\sum_{t=1+j}^{T}y_t)\bar{y}' + \bar{y}\bar{y}'\frac{T-j}{T}\\ &= \frac{1}{T}\sum_{t=1}^{T}y_{t+j}y_t' - \frac{1}{T}\sum_{t=T-j+1}^{T}y_{t+j}y_t' - \bar{y}(\bar{y}-\frac{1}{T}\sum_{t=T-j+1}^{T}y_t)' - (\bar{y}-\frac{1}{T}\sum_{t=1}^{j}y_t)\bar{y}' + \bar{y}\bar{y}'\frac{T-j}{T}\\ &= \tilde{\gamma}_j - \frac{1}{T}\sum_{t=T-j+1}^{T}y_{t+j}y_t' + \frac{1}{T}\sum_{t=T-j+1}^{T}\bar{y}y_t' + \frac{1}{T}\sum_{t=1}^{j}y_t\bar{y}' - \frac{j}{T}\bar{y}\bar{y}'\end{aligned}$$

This shows the result.
(iv) From (iii) it follows that $\hat{\gamma}_j=\tilde{\gamma}_j+O_P(j/T)$. The rest follows from [8].
(v) Uniformity of a.s. convergence is a direct consequence of [12] Theorem 1 and 2 noting that these results only use the variance bounds derived above. □

Lemma A.2. *Let $(y_t)_{t\in\mathbb{Z}}$ be as in Lemma A.1. Define $k(z):=\sum_{j=0}^{\infty}k_jz^j$ and the spectrum $f(\omega)=\frac{1}{2\pi}k(e^{i\omega})\Omega k(e^{i\omega})^*$ where $\Omega:=\mathbb{E}\varepsilon_t\varepsilon_t'$.*
(i) Assume that there exist constants $0<a,b<\infty$ such that $aI_s\leq f(\omega)\leq bI_s\overline{f}(\omega)$ where $\overline{f}(\omega)=|1-e^{i\omega}|^{-2d}$ for $d\geq 0$. Then there exists constants $0<C_1<C_2<\infty$ such that $C_1I_{ps}\leq\mathbb{E}Y_{t,p}^-(Y_{t,p}^-)'\leq C_2I_{ps}p^{2d}$.
(ii) If there exist constants $0<a,b<\infty$ such that $aI_s\underline{f}(\omega)\leq f(\omega)\leq bI_s$ where $\underline{f}(\omega)=|1-e^{i\omega}|^{-2d}$ for $d<0$, then there exist constants $0<C_1<C_2<\infty$ such that $C_1I_{ps}p^{2d}\leq\mathbb{E}Y_{t,p}^-(Y_{t,p}^-)'\leq C_2$.

Proof. The proof is a straightforward generalization of the univariate result in Theorem 2 of [11], compare also Lemma 2 in [12]. Only (i) is proved, the remaining statements follow analogously. Let $\Gamma_p^-=\mathbb{E}Y_{t,p}^-(Y_{t,p}^-)'$. Note that the definition of $\overline{f}$ and $\underline{f}$ coincide with the definition given in (9), (10), p. 96 of [11]. The smallest eigenvalue of Γ_p^- equals the minimum of $x'\Gamma_p^-x$ for $x'x=1$ and the largest corresponds to the maximum. Since $(y_t)_{t\in\mathbb{Z}}$ is assumed to be stationary with spectral density $f(z):=k(z)\Omega k(z)^*/(2\pi)$ it follows that

$$x'\Gamma_p^-x=\int_{-\pi}^{\pi}(\sum_{j=1}^{p}x_je^{ij\omega})^*f(e^{i\omega})(\sum_{j=1}^{p}x_je^{ij\omega})d\omega$$

where $x' = [x_1', \ldots, x_p'], x_j \in \mathbb{R}^s$. If $d > 0$ then

$$\int_{-\pi}^{\pi} (\sum_{j=1}^{p} x_j e^{ij\omega})^* f(e^{i\omega}) (\sum_{j=1}^{p} x_j e^{ij\omega}) d\omega \geq \int_{-\pi}^{\pi} (\sum_{j=1}^{p} x_j e^{ij\omega})^* (\sum_{j=1}^{p} x_j e^{ij\omega}) d\omega a \geq 2\pi a.$$

Also

$$\int_{-\pi}^{\pi} (\sum_{j=1}^{p} x_j e^{ij\omega})^* f(e^{i\omega}) (\sum_{j=1}^{p} x_j e^{ij\omega}) d\omega \leq b \int_{-\pi}^{\pi} \| \sum_{j=1}^{p} x_j e^{ij\omega} \|^2 \overline{f}(\omega) d\omega.$$

It follows that the function $h(\omega) := \| \sum_{j=1}^{p} x_j e^{ij\omega} \|^2$ is in the set P_p (p. 97 of [11]). Hence the result holds. □

Lemma A.3. *Let $(y_t)_{t\in\mathbb{Z}}$ be as in Theorem 3.2. Then:*

- *For $0.25 < d_+ < 0.5$ we have $\|\mathbb{E}Y_{t,f}^+ (Y_{t,p}^-)'\|_{Fr} = O((f+p)^{2d_+})$, $\|\mathbb{E}Y_{t,p}^- (Y_{t,p}^-)'\|_{Fr} = O(p^{2d_+})$.*
- *For $d_+ = 0.25$ we have $\|\mathbb{E}Y_{t,f}^+ (Y_{t,p}^-)'\|_{Fr} = O((f+p)^{1/2})$, $\|\mathbb{E}Y_{t,p}^- (Y_{t,p}^-)'\|_{Fr} = O(\sqrt{p \log p})$.*
- *For $0 \leq d_+ < 0.25$ we have $\|\mathbb{E}Y_{t,f}^+ (Y_{t,p}^-)'\|_{Fr} = O((f+p)^{2d_+})$, $\|\mathbb{E}Y_{t,p}^- (Y_{t,p}^-)'\|_{Fr} = O(\sqrt{p})$.*
- *For $d_+ < 0$ we have $\|\mathbb{E}Y_{t,f}^+ (Y_{t,p}^-)'\|_{Fr} = O(1)$, $\|\mathbb{E}Y_{t,p}^- (Y_{t,p}^-)'\|_{Fr} = O(\sqrt{p})$.*

The lemma is an easy consequence of $\|\gamma_l\| \leq \mu l^{2d_+ - 1}$ as shown in Lemma A.1 in combination with $\sum_{j=1}^{m} j^{\beta-1} = O(m^\beta)$ for $\beta > 0$.

Lemma A.4. *Let the assumptions of Theorem 3.2 hold. Let $\tilde{d}_- = \min(d_-, 0), \tilde{d}_+ = \max(d_+, 0)$ and assume that $p^{2-4\tilde{d}_-} R_T(\tilde{d}_+)^2 \to 0$ and $pR_T(\tilde{d}_+) \to 0$. Then*

$$\begin{aligned} \|\langle Y_{t,f}^+, Y_{t,p}^- \rangle - \mathbb{E}Y_{t,f}^+ (Y_{t,p}^-)'\|_{Fr} &= O\left(\sqrt{fp}\left(R_T(\tilde{d}_+)\right)\right), \\ \|\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \Gamma_p^- \|_{Fr} &= O(pR_T(\tilde{d}_+)), \\ \|\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} - (\Gamma_p^-)^{-1}\|_2 &= O(p^{1-4\tilde{d}_-} R_T(\tilde{d}_+)) \end{aligned}$$

Proof. All three parts follow almost immediately from Lemma A.1 (v). With respect to the third statement using [10], p. 397, l. 11, we obtain:

$$\|\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} - (\Gamma_p^-)^{-1}\|_2 = F^2 Z_{p,T} / (1 - F Z_{p,T})$$

where $F := \|(\Gamma_p^-)^{-1}\|_2, Z_{p,T} := \|\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} - (\Gamma_p^-)^{-1}\|_2 / F(\|\langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} - (\Gamma_p^-)^{-1}\|_2 + F) \leq \|\langle Y_{t,p}^-, Y_{t,p}^- \rangle - \Gamma_p^-\|_2$. Then the upper bound on p implies that $Z_{p,T} = o(1)$. For $d_- > 0$ and hence $\tilde{d}_- = 0$ we have $F < \infty$ and therefore the result follows. For $d_- < 0$ we have $F = O(p^{-2d_-})$ and therefore the upper bound on p has to ascertain that $F Z_{p,T} \to 0$ where $Z_{p,T} = O(pR_T(\tilde{d}_+))$. □

Lemma A.5. *Let the assumptions of Theorem 3.2 hold and let $\phi_j(p)$ denote the coefficients of the long VAR approximation in (3). Then $\sum_{j=1}^p \|\phi_j - \phi_j(p)\|^2 = O(p^{4d_+-2d_--1})$ if $d_+ \geq d_- \geq 0$ which tends to zero if $d_- > 2d_+ - 0.5$. If $d_+ = d_- = d$ this always holds and the order equals $O(p^{2d-1})$.*
If $d_- \leq d_+ \leq 0$ then $\sum_{j=1}^p \|\phi_j - \phi_j(p)\|^2 = O(p^{4d_+-6d_--1})$ which tends to zero if $2d_+ < 0.5 + 3d_-$. For $0 \geq d_+ = d_- > -0.5$ this always holds and the rate equals $O(p^{-2d-1})$ in this case.

Proof. This result has already been obtained in the scalar case (where automatically $d_+ = d_-$) by [5]. Note that using $f = 1$ we obtain:

$$0 = \mathbb{E}Y_{t,1}^+(Y_{t,p}^-)' - \mathbb{E}Y_{t,1}^+(Y_{t,p}^-)' = \beta_{1,p}\Gamma_p^- - [I_s, 0]\beta\Gamma_\infty^-[I_{ps}, 0]' = [\Phi_p - [\Phi]_p]\,\Gamma_p^- - \Phi_p^-\tilde{\Gamma}_{2,p}^-$$

where Φ_p^- denotes the matrix Φ where the first p block columns are omitted. Further $\tilde{\Gamma}_{2,p}^- = \mathbb{E}Y_{t-p,\infty}^-(Y_{t,p}^-)'$. Therefore it is sufficient to compute the Frobenius norm of $\Phi_p^-\tilde{\Gamma}_{2,p}^-$ which contains as a typical element $\sum_{i=1}^\infty \phi_{i+p}\mathbb{E}y_{t-p-i}y_{t-j}'$, $j = 1, \ldots, p$. Using $\|\phi_l\| \leq M_l l^{-1-d_-}, \|\gamma_l\| \leq M_g l^{2d_+-1}$ we can bound the norm of this entry by $M_l M_g \sum_{i=1}^\infty (i+p)^{-1-d_-}(p+i-j)^{2d_+-1}$. For $d_- > 0$ this is of order $O(p^{-d_-}(p-j)^{2d_+-1})$. Summing the squares over $j = 1, \ldots, p$ shows that the squared Frobenius norm of $\Phi_p^-\tilde{\Gamma}_{2,p}^-$ in this case is of order $O(p^{4d_+-2d_--1})$. Since the smallest eigenvalue of Γ_p^- is bounded away from zero for $d_- > 0$ the result follows. The result for $d_+ = d_-$ is obivous.
For $d_+ < 0$ the norm of the typical entry is of order $O(p^{-1-d_-}(p-j)^{2d_+})$ and therefore the squared Frobenius norm of $\Phi_p^-\tilde{\Gamma}_{2,p}^-$ also in this case is of order $O(p^{4d_+-2d_--1})$. Here the smallest eigenvalue of Γ_p^- tends to zero as p^{2d_-} and hence the inverse adds a factor p^{-2d_-} to the Frobenius norm adding up to order $O(p^{-1+4d_+-6d_-}) = o(1)$ if $-0.5 + 2d_+ - 3d_- < 0$. □

A.2 Proof of Theorem 3.2

The main insight into the algorithm lies in the fact that for $n = ps$ the `CVA` estimate $(\hat{A}, \hat{C}, \hat{K})$ equals the long VAR approximation of $(y_t)_{t\in\mathbb{Z}}$ using lag order p [4], the properties of which follow along the lines of [12]:

Lemma A.6. *Let $(y_t)_{t\in\mathbb{Z}}$ be as in Theorem 3.2 denoting a long VAR approximation as*

$$y_t = \sum_{i=1}^p \phi_i(p)y_{t-i} + \varepsilon_t(p). \tag{3}$$

Further let the assumptions on d_-, d_+ be as in Theorem 3.2. Then the OLS estimates of the coefficients in this long VAR approximation fulfill uniformly in $1 \leq p \leq H_T$ where H_T is such that $H_T^{2-4\tilde{d}_-}R_T(\tilde{d}_+)^2 \to 0, R_T(\tilde{d}_+)P_T(d_-, H_T) \to 0$ where $P_T(d,p) = p$ for $d \geq 0$ and $P_T(d,p) = p^{1-4d}$ for $d < 0$.
Then

$$\max_{1\leq p\leq H_T}\left(\sum_{j=1}^p \|\hat{\phi}_j(p) - \phi_j(p)\|^2/P_T(d_-,p)^2\right)^{1/2} = O(R_T(\tilde{d}_+)).$$

Proof. The proof follows the arguments of [12]. The long AR-approximation can be written as $y_t = \beta_{1,p} Y_{t,p}^- + \varepsilon_t(p)$. It follows that $\hat{\beta}_{1,p} = \langle y_t, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} = \left[\hat{\phi}_1(p) \; \hat{\phi}_2(p) \; \dots \; \hat{\phi}_p(p) \right]$. Let

$$\beta_{1,p} = \mathbb{E} y_t (Y_{t,p}^-)' [\mathbb{E} Y_{t,p}^- (Y_{t,p}^-)']^{-1} = \mathcal{H}_{1,p} (\Gamma_p^-)^{-1} = \left[\phi_1(p) \; \phi_2(p) \; \dots \; \phi_p(p) \right].$$

Then Lemma A.5 implies that $\|\beta_{1,p} - [\Phi]_p\|_{Fr} = o(1)$ where $\sup_p \|[\Phi]_p\|_{Fr} < \infty$ due to the square integrability of the AR coefficients. This implies $\sup_p \|\beta_{1,p}\|_{Fr} < \infty$. Since $\|\gamma_l\|_{Fr} \leq \kappa l^{2d_+ - 1}$ it follows that $\sup_p \|\mathcal{H}_{1,p}\|_{Fr} < \infty$ for $d_+ < 0.25$. Further note that $\|(\hat{\Gamma}_p^-)^{-1}\|_2 \leq \|(\hat{\Gamma}_p^-)^{-1} - (\Gamma_p^-)^{-1}\|_2 + \|(\Gamma_p^-)^{-1}\|_2 = o(1) + \|(\Gamma_p^-)^{-1}\|_2$. Then

$$\begin{aligned} \hat{\beta}_{1,p} - \beta_{1,p} = \hat{\mathcal{H}}_{1,p} (\hat{\Gamma}_p^-)^{-1} - \mathcal{H}_{1,p} (\Gamma_p^-)^{-1} &= (\hat{\mathcal{H}}_{1,p} - \mathcal{H}_{1,p})(\hat{\Gamma}_p^-)^{-1} + \mathcal{H}_{1,p}[(\hat{\Gamma}_p^-)^{-1} - (\Gamma_p^-)^{-1}] \\ &= (\hat{\mathcal{H}}_{1,p} - \mathcal{H}_{1,p})(\hat{\Gamma}_p^-)^{-1} - \beta_{1,p}[\hat{\Gamma}_p^- - \Gamma_p^-](\hat{\Gamma}_p^-)^{-1}. \end{aligned}$$

From these equations (for $d_+ > 0$ we can use the third equation, else the second can be used) in combination with Lemma A.4 (where the norm bounds hold uniformly in the lag length) the result follows. □

Consequently $p \to \infty$ at the rate given in the theorem implies that for $\hat{\phi}(z) = \sum_{j=0}^p \hat{\phi}_j z^j$ and $\phi_p(z) = \sum_{j=0}^p \phi_j(p) z^j$ it follows that $\|\hat{\phi}(z) - \phi_p(z)\|_2 \to 0$. Furthermore the norm bound implies that

$$\sum_{j=p+1}^{\infty} \|\phi_j\|^2 \leq M_\phi \sum_{j=p+1}^{\infty} j^{-2-2d_-} = O(p^{-1-2d_-})$$

and hence the Fourier series $\sum_{j=0}^p \phi_j z^j$ converges in L_2 to $\phi(z)$, that is $\|\phi_p(z) - \phi(z)\|_2 \to 0$ and thus $\|\hat{\phi}(z) - \phi(z)\|_2 \to 0$.
These two lemmas show that subspace methods with the maximal choice of the order $n = ps$ deliver consistent estimates of the inverse transfer function $\phi(z)$ and thus also of the transfer function. This extends theory for long VAR estimation from the univariate setting to the multivariate setting allowing for different indices d_j in each component.
The final lemma provides some facts on the approximation error from using $n < ps$.

Lemma A.7. *Let $(\hat{A}, \hat{C}, \hat{K})$ denote a system with corresponding state $\hat{x}_t$ such that $\langle \hat{x}_t, \hat{x}_t \rangle = I_N$ and innovation noise variance $\hat{\Omega}$. Then consider the partitioning of the system as (where $\hat{A}_{11} \in \mathbb{R}^{n \times n}, \hat{K}_1 \in \mathbb{R}^{n \times s}, \hat{C}_1 \in \mathbb{R}^{s \times n}$)*

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}, \quad \hat{K} = \begin{bmatrix} \hat{K}_1 \\ \hat{K}_2 \end{bmatrix}, \quad \hat{C} = \left[\hat{C}_1 \; \hat{C}_2 \right].$$

(i) The system $(\hat{A}_n, \hat{C}_n, \hat{K}_n)$ obtained from using $\hat{x}_{t,n} = [I_n, 0]\hat{x}_t$ in the regressions in the last step of the `CVA` *algorithm fulfills*

$$\hat{A}_n = \hat{A}_{11}, \quad \|\hat{K}_n - \hat{K}_1\|_{Fr} = O(\|\hat{C}_2\|_{Fr}(\sqrt{n}\|\hat{C}_2\|_{Fr} + \|\hat{A}_{1,2}\|_{Fr})), \quad \hat{C}_n = \hat{C}_1.$$

(ii) With $\hat{C}_{:,j}$ denoting the j-th column of $\hat{C}$ we obtain ($\tilde{d}_+ = \max(d_+, 0)$) $\|\hat{C}_{:,j}\|_2 = O\left(p^{\tilde{d}_+}\hat{\sigma}_j\right)$.
(iii) Furthermore let $\|f(z)\|_\infty = \sup_{\omega\in[0,2\pi]} \|f(e^{i\omega})\|_2$ and use the notation $\hat{k}(z) = I_s + z\hat{C}(I - z\hat{A})^{-1}\hat{K}, \hat{k}_{n1}(z) = I_s + z\hat{C}_n(I - z\hat{A}_n)^{-1}\hat{K}_1$. Then $\|\hat{k}(z) - \hat{k}_{n1}(z)\|_\infty$

$\leq \|\hat{C}_2 + \hat{C}_1(zI_n - \hat{A}_{11})^{-1}\hat{A}_{12}\|_\infty \|(zI - \hat{A}_{22} - \hat{A}_{21}(zI_n - \hat{A}_{11})^{-1}\hat{A}_{12})^{-1}(\hat{K}_2 + \hat{A}_{21}(zI_n - \hat{A}_{11})^{-1}\hat{K}_1)\|$.

(iv) Consequently using $\hat{k}_n(z) = I_s + z\hat{C}_n(I - z\hat{A}_n)^{-1}\hat{K}_n$ we obtain

$$\|\hat{k}(z) - \hat{k}_n(z)\|_\infty \leq \|\hat{k}(z) - \hat{k}_{n1}(z)\|_\infty + \|\hat{C}_n(I - z\hat{A}_n)^{-1}\|_\infty \|\hat{K}_n - \hat{K}_1\|.$$

Proof. (i) [4] show that $C_n = C_1, \Omega_n = \Omega + C_2C_2', A_n = A_{11}, K_n = (M_1 - A_{11}C_1')\Omega_n^{-1}$ where $M_1 = K_1\Omega + [I_n, 0]AC'$. Thus $\|\Omega_n - \Omega\| = \|C_2\|^2$. Therefore

$$K_n\Omega_n - K_1\Omega = A_{1,2}C_2' \Rightarrow (K_n - K_1)\Omega_n = A_{1,2}C_2' + K_1(\Omega_n - \Omega).$$

Since $\|\Omega^{-1}\| \leq M, \|K_1\| \leq \sqrt{n}, \|A_{1,2}\| \leq \sqrt{n}$ and $\Omega_n \geq \Omega$ it thus follows that

$$\begin{aligned}\|K_n - K_1\| &= \|[K_1(\Omega_n - \Omega) + A_{1,2}C_2']\Omega_n^{-1}\| \leq \|K_1C_2\|\|\Omega^{-1}\|\|C_2\| + \|C_2\|\|A_{1,2}\|\|\Omega^{-1}\| \\ &= \|\Omega^{-1}\|\|C_2\|(\|K_1C_2\| + \|A_{1,2}\|) \leq \|\Omega^{-1}\|\|C_2\|(\|C_2\|\sqrt{n} + \sqrt{n}).\end{aligned}$$

It is straightforward to show that this also holds for the estimates as only orthogonality relations are used here.
(ii) Consider the estimation of the state as $\hat{x}_t = \hat{V}'(\hat{\Gamma}_p^-)^{-1/2}Y_{t,p}^-$ which implies that $\langle\hat{x}_t, \hat{x}_t\rangle = I_{ps}$. According to Lemma A.4 $\|\hat{\Gamma}_f^+ - \Gamma_f^+\|_{Fr} \to 0$ if $pR_T(\tilde{d}_+) \to 0$. Then $\|\hat{\Gamma}_f^+\|_2 \leq \|\hat{\Gamma}_f^+ - \Gamma_f^+\|_2 + \|\Gamma_f^+\|_2 = O(p^{2\tilde{d}_+})$. Furthermore

$$\hat{C}_2 = \hat{\beta}_{f,p}\hat{V}_2 = [I_s, 0](\hat{\Gamma_f^+})^{1/2}\hat{U}\hat{\Sigma}\hat{V}'\hat{V}[0, I]' = [I_s, 0](\hat{\Gamma_f^+})^{1/2}\hat{U}_2\hat{\Sigma}_2,$$

showing that the 2-norm of the j-th column of $\hat{C}_2$ is of order $O(p^{\tilde{d}_+}\hat{\sigma}_{n+j})$.
(iii) and (iv) follow from straightforward computations. □

Changepoints to Improve Forecasts

Jamie-Leigh Chapman*[1], Rebecca Killick[1] and Idris Eckley[1]

[1]Lancaster University, UK
*Email: j.chapman1@lancaster.ac.uk

Abstract: Forecasting time series accurately is challenging, particularly so when the structure of the data changes through time. In recent years, research has focused on forecasting for non-stationary time series, particularly in a local evolutionary setting. However many time series encountered in practice may have a piecewise stationary form. In this talk we introduce a novel approach to improve forecasting using changepoints and apply it to Gross Domestic Product data.

Introduction

Forecasting models try to capture the behaviour of a time series through various components, including outliers, level, trend, volatility, seasonality, dependence and explanatory variables. In practice, these components are often assumed to be stationary, but the reality is that these properties may vary over time in a piecewise way. Consider for example the UK's Gross Domestic Product (GDP) data in Figure 1, the volatility is higher prior to 1981 and in general the behaviour is not constant over time.

Methods

Suppose, by way of introduction, we wish to obtain forecasts for the UK's GDP. As a precursor to formal forecasting, one might perform changepoint analysis to identify regions of piecewise stationary structure. More formally, suppose our time series $(y_1, \ldots, y_n)$ contains m changepoints with positions $\tau = (\tau_1, \ldots, \tau_m)$. Each changepoint position, τ_i, is an integer between 1 and $n-1$ and we define: $\tau_0 = 0$ and $\tau_{m+1} = n$. The changepoints are ordered such that: $\tau_i < \tau_j \iff i < j$. Consequently the m changepoints split the time series data into $m+1$ segments with the i^{th} segment containing $y_{(\tau_{i-1}+1):\tau_i}$. Each segment i is summarized by a set of parameters $\{\theta_i, \phi_i\}$. Within this set, ϕ_i is a set of nuisance parameters and θ_i is a set of parameters that may contain changes (Eckley *et al.* 2011).

One way to detect multiple changepoints is to minimize: $\sum_{i=1}^{m+1} \left[\mathcal{C}(y_{(\tau_{i-1}+1):\tau_i})\right] + \beta f(m)$, where $\mathcal{C}$ is some cost function for a segment and $\beta f(m)$ is a penalty based

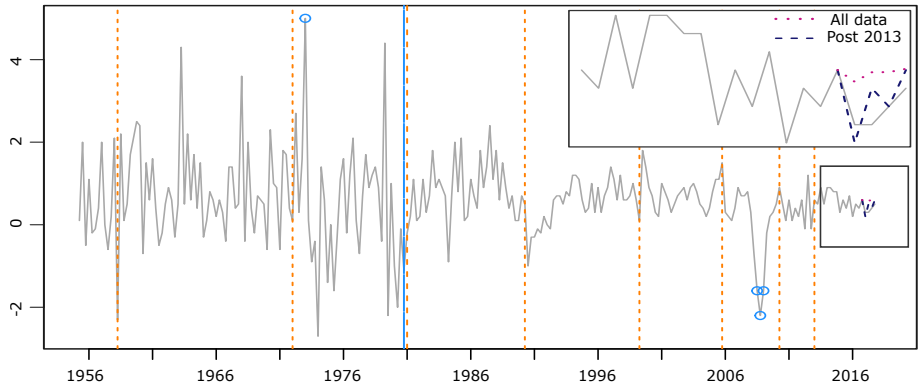


Figure 1: The United Kingdom's Gross Domestic Product (GDP) quarter on quarter growth. The vertical (blue) solid line is a change in variance and the (orange) dashed lines are changes in ARIMA structure. The enclosed box shows forecasts using all of the data and using only data after the most recent change.

on the number of changepoints m which prevents over-fitting (Killick *et al.*, 2012). For forecasting GDP, we can use the negative log-likelihood for an autoregressive integrated moving average (ARIMA) model as the cost function. In particular, we can use a dynamic cost function which searches over different models, fits the best one, and returns the log likelihood. This allows for both changes in coefficients and model order. If we locate a significant change in the ARIMA model, then we should forecast using only the data after that change. Figure 1 shows the detected changes in the ARIMA model for the UK's GDP and forecasts using all of the historical data, and only using the data after 2013.

Results and Conclusions

Changepoint analysis is not only useful for the retrospective understanding of data, but it can also be used to improve future forecasts. It's integration into existing time series forecasting models can be straight forward and we often achieve better out-of-sample forecasts based on less historical information.

References

Eckley, I. *et al.*. (2011). Analysis of changepoint models. *Bayesian Time Series Models*. Cambridge University Press, Cambridge.

Killick, R. *et al.*. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, pp. 1590 – 1598.

ProMoBed: a forecasting and simulation model for estimating future hospital bed capacity

Marlies Van der Wee[1], Timo Latruwe[1], Sofie Verbrugge[1], Pieter Vanleenhove[2], Henk Vansteenkiste[2], Sebastiaan Vermeersch[2]

[1] IDLab, Ghent University – imec, Ghent, Belgium
[2] hict, Bruges, Belgium

Corresponding author: marlies.vanderwee@ugent.be

Healthcare is an everchanging environment, where currently the most important trends are an ageing population and a shift towards chronic diseases. These trends result in growing costs and have forced governments to cut in healthcare budgets and initiate reformations such as a reduction of length of stay (LOS) in hospitals and a transfer towards home care, care hotels or elderly homes. Hospitals play a central role within this landscape as they cover 25-45% of the total healthcare budget within Belgium (CM, 2015) and Europe (Eurostat, 2012). Hence, hospitals are largely affected by these budget cuts and reformations. Novel legislation (will) force(s) hospitals to form networks and reorganize the care offer and a recent study from the Belgian Healthcare Knowledge Centre (KCE) stated that it is possible to reduce the amount of hospital beds across Belgium with 10% (Belgian Healthcare Knowledge Centre, 2017).

A tool which predicts the required bed capacity over long-term periods would help hospitals take strategic decisions. This is the case with the upcoming network formation, but also with the – expected to be continuously evolving – shift towards home care, the reducing length of stay for most treatments or the changes in population and epidemiology. Moreover, since the role of hospitals will change, other instances will take over some roles (elderly homes, care hotels, home care nurses etc.). Those parties could similarly be helped by a tool that predicts the required (bed, chair, revalidation room...) capacity. Besides these long-term evolutions, short-term fluctuations (summer/winter, outbreak of infections) pose operational challenges. Being able to predict such changes would help in re-scheduling nurse shifts within hospitals, improving the work-load and service quality within hospitals.

A market survey performed by the authors shows that most hospitals base their predictions on comparisons with other hospitals (benchmarks such as Forcea (2018) or-Belfius (2017)) or simple business intelligence-dashboards (e.g. 3M (2018), internal tools). These predictions however lack demographic evolutions, changes in epidemiology, regional parameters, governmental decisions or network dynamics. Additionally, the feedback data from governmental instances are generally outdated (>2-3 years). Hence, hospitals frequently lack detailed and correct information to which they can base their policy decisions.

The KCE recently published a study in which the future need for hospital beds is modelled and calculated (Belgian Healthcare Knowledge Centre, 2017). This study is of great relevance and can be seen as 'state of the art' for modelling hospital bed capacity in Belgium. Although the model is of great importance for the Belgian healthcare context, several aspects are lacking from the model, which leads to a difficult translation into practice. These are (1) a fixed prediction horizon (2025) instead of a flexible horizon that allows both short- and long-term predictions, (2) nation-wide aggregated data instead of hospital-specific inputs, (3) geographically broad demographic forecasts, not taking into account regional differences and (4) no room for simulating a large variety of alternative ("what-if") scenarios.

Internationally, there are also models that aim at predicting future bed capacity in hospitals. First of all, there is literature that studies which parameters influence the required numbers of hospital beds, e.g., discharge rates, distribution of patients across diseases, patient length of stay, demographics and seasonal & epidemic scenarios (Barnett et al, 2012, Cavanagh and Chadwick, 2005). Next, there are also specific statistical prediction models. Boyle at al. (2016), for example, present a regression model based on patient influx in the emergency department of one specific hospital. Thirdly, discrete event simulation (DES) models are used. One example is presented by Devapriya et al (2015), who developed a decision support tool based on a DES that receives patient arrival and discharges as inputs. Finally, there are combinations of statistical predictions and simulations. Demir et al. (2017) present such a tool and applied it to the National Health Service hospitals in England, be it on a limited timeframe of 5-10 years. They however do allow hospital decision makers to run multiple scenarios, varying different parameters.

A bed capacity simulation model that combines statistical prediction models with the simulation of different alternative scenarios going beyond the DES input parameters (such as new network formations, governmental decisions, changes in outbreaks of infections, etc.) has – to the best of the author's knowledge – not yet been developed. Such a model would be of greater relevance to hospital decision makers, as it combines the strengths of the different approaches described above:

- The trends in the model are based on statistically validated data and hence reliable.
- The DES incorporated in the model can be used to compare different scenarios and evaluate the impact of the variation of certain parameters.
- The knowledge of domain experts can be inputted in the simulator to allow for defining and simulating alternative scenarios going beyond the DES input parameters.
- The flexibility of the model allows forecasting and simulating for multiple time horizons, ranging from short-term operational impacts to long-term strategic decision making.

Our innovative solution, ProMoBed, tries to overcome subjectivity and simplifications and enables accurate long and short-term predictions by using a simulation

approach where multiple parameters, identified to be relevant in determining the bed capacity and the bed occupancy within hospitals, will be used and refined to get a solid, predictive model. The model is able to forecast both in the long and short run, changes in market shares are explicitly modelled and specific regional demographic changes are included.

More specifically, ProMoBed aims at solving the bed capacity estimation problem by combining the strengths of structured literature research, parameter correlation studies and simulation models. This approach will lead to data-supported results, while still allowing the flexibility of including expert knowledge, situation-specific policies or constraints (such as market share changes), and sensitivity analysis:

1. Structured literature research and statistical parameter correlation studies will generate more precise and valid disease predictions on patient, national or European level. These studies are based on detailed per-hospital data combined with regional, national and European trends.
2. The statistical regression model will combine this input with historic and demographic data to generate reliable estimates of general trends on e.g. patient arrival dates, regional demographics and length of stay.
3. The discrete event simulation model will be able to model situation-specific policies or constraints (for example changes in market shares), to achieve a forecast on needed bed count.

This paper will present the conceptual ProMoBed model, which is still work-in-progress.

In summary, ProMoBed is a huge improvement with respect to the current way of working: transition from subjective expert opinion on a limited set of parameters to up-to-date datasets, basing decisions on the combination of data forecasts and expert knowledge.

References

1. CM (2015) Twaalfde CM-ziekenhuisbarometer. Retrieved from https://www.cm.be/media/CM-Ziekenhuisbarometer-2016_tcm47-19246.pdf
2. Eurostat (2012) Healthcare expenditure by function, 2012. Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Healthcare_expenditure_by_function,_2012_(%25_of_current_health_expenditure)_YB15.png
3. Belgian Healthcare Knowledge Centre (2017) Required Hospital Capacity in 2025 and criteria for rationalization of complex cancer surgery, radiotherapy and maternity services. Retrieved from https://kce.fgov.be/sites/default/files/atoms/files/Download%20het%20rapport%20in%20het%20Engels%20%28550%20p.%29.pdf
4. Forcea (2018) Forcea Health report. http://www.forcea.be/en/solutions/forcea-healthreport-0
5. Belfius (2017) Maha-study 2017. Algemene ziekenhuizen. Retrieved from https://www.belfius.be/publicsocial/NL/Media/MAHA%202017%20persconferentie%20NL_tcm_31-142132.pdf

6. 3M (2018) 3M Health information systems. https://www.3m.com/3M/en_US/health-information-systems-us/
7. Barnett, K. (2012). Best practices for community health needs assessment and implementation strategy development: A review of scientific methods, current practices, and future potential. *Atlanta, Georgia: The Public Health Institute & The Centre for Disease Control and Prevention*.
8. Cavanagh, S. and Chadwick, K. (2005) *Health needs assessment: a practical guide*. National Institute for health and Clinical Excellence.
9. Boyle, J., Ireland, D., Webster, F., & O'Sullivan, K. (2016, February). Predicting demand for hospital capacity planning. In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on* (pp. 328-331). IEEE.
10. Devapriya, P., Strömblad, C. T., Bailey, M. D., Frazier, S., Bulger, J., Kemberling, S. T., & Wood, K. E. (2015). StratBAM: A discrete-event simulation model to support strategic hospital bed capacity decisions. *Journal of medical systems*, *39*(10), 130.
11. Demir, E., Gunal, M. M., & Southern, D. (2017). Demand and capacity modelling for acute services using discrete event simulation. *Health Systems*, *6*(1), 33-40.

Panel Data Unit Root Tests on the Income-Health Relationship of the Mexican States

Vicente German-Soto[1] [2]
Autonomous University of Coahuila, Faculty of Economics
Unidad Camporredondo, Edificio "E", Planta Baja, entre las calles Av. De los Maestros y David Berlanga, C.P. 25000, Saltillo, Coahuila, México. Tel.: +52 844 412 87 82, Fax: +52 844 410 26 79.
E-mail: vicentegerman@uadec.edu.mx

Martha Elena Fuentes Castillo
Antonio Narro Agrarian Autonomous University, Sociology Department
Calzada Antonio Narro No. 1923, Buenavista, Saltillo, Coahuila, Mexico. C.P. 25315. Tel: +52 844 411 02 00.
E-mail: mefuecas@uaaan.edu.mx

Abstract. This work analyses the equilibrium between income and health expenditure across the Mexican states along 1990-2011. The methodology consists in stationarity tests for heterogeneous panel data that first consider the presence of structural breaks in the series, in both intercept and slope, and then evaluate the unit root. Consequently, distortions in size and location of the breaks are avoided. After the cross-section dependence is accounted for, most of the economies can demonstrate stationary processes evolving around a broken trend, reinforcing the theoretical expectative on the income-health link. Location of the breakpoints constitutes meaningful information about how reforms can promote and extend the health coverage.

Keywords: panel data, unit roots, structural breaks, health economics, income

1 Introduction

Income level and health are quite related in the long-run. The idea is that to higher available income, higher health care expenditure will be. Recently, this relationship has attracted the attention of scientists and researchers since the seminal publications of Mushkin [34], Fuchs [15] and Grossman [18]. New empirical evidence for developed economies has argued that not only a strong positive correlation exists between income and health but also the former explains a high proportion of the health variation.

A review on this line suggests that exists much more empirical evidence for developed countries in comparison to developing countries [24, 26, 39]. Even for the regional scale the evidence is still scarcer (some exceptions are [2, 3], for Indian states, and [8], for U.S. states).

[1] Orcid: 0000-0001-5844-1296.
[2] Corresponding author.

The aim of this work is testing the theory through stochastic panel data at the regional level of Mexico. The hypothesis is that effectively an equilibrium relationship exists between income and health, but in a context where stochastic processes evolved around a broken trend that changed the equilibrium relationship. Some reforms and economic events, such as crises, seem to justify the consideration of the structural breaks in the estimates. Methodology is supported on Im, Lee and Tieslau [20, 21] and Lee, Strazicich and Meng [27] in the case of panel data that allows heterogeneous structural breaks in both intercept and slope.

The paper contributes in three ways. Firstly, we apply the two-step procedure of Lee, et al. [27] to test stationarity in panel data, a more flexible deterministic specification because it incorporates both level and slope shifts. Secondly, our analysis of stationarity is applied on the 'gap' between income and health public expenditure and, therefore, causality is not a concern, as is the case in methodologies that should to estimate the cointegration vector. Finally, the paper extends the body of knowledge focused on the health-GDP relationship of Mexico and its regions because there are not works at this context with the long-run perspective.

2 The income-health connection

Theoretical discussions on this link appeared in Mushkin [34], Fuchs [15] and Grossman [18]. The first work postulated that healthiest people contributes to better production and so the revenues are reinforced. The second work postulates that the relationship between them is because persons require adequate levels of health and, given those levels, they can reach better increases in production. Meanwhile, the third work emphasized the importance of the health as consumption good and as an investment. As consumption good, health attends the necessity to have medical services, while as an investment it allows to improve health.

Subsequent contributions emphasized the construction of theoretical models where health is a fundamental factor, as in studies of the economic growth. On the one hand, it allowed expanding the focus of neoclassical economic growth (Barro [5]; Bloom, Canning, and Sevilla [6]). On the other hand, the endogenous economic growth expanded the theoretical contributions on this field (Zon and Muysken [40], about human capital, and Howitt [19], from the perspective of creative destruction).

More recently, this connection is analyzed as a link that would satisfy an equilibrium condition, mainly in the long-run. The idea is that if persons (and economies) in fact are disposal to invest more in health when they increase their income, then a sustained equilibrium should prevail. This result can be assessed through analyzing the statistical properties of the variables.

In the income-health relationship the causality may arise from health to the income, vice versa, or it goes in both ways. At this respect, it does not exist a consensus. In this investigation, we emphasize that none consideration about the direction of causality be necessary if the income-health relationship is treated as a gap evolving along time. So, our estimates are concerned only with the reviewing about whether this series is constant in time (in equilibrium) or, on the contrary, it is changing (out of equilibrium).

3 Empirical evidence, a review

One line of investigation is focused on economic growth models including health among the factors. This is the case of Chang and Yung-Hsiang [10], for countries of the Organization for Economic Cooperation and Development (OECD), and Adeniyi and Abiodun [1], for Nigeria. In both studies, it was found that health expenditure had a positive effect on economic growth. Other line considers analyzing the stochastic properties of the variables, in two ways: with and without structural changes.

Evidence when none structural change is considered is wide. Cointegration analyses find a positive relationship in the long-run (Dreger and Reimers [12], Baltagi and Moscone [4], Wang [38], Pekkurnaz [36], among others). Dreger and Reimers [12] evaluated 21 countries between 1975 and 2001 and incorporated into their analysis the life expectancy and the mortality rates as proxies of medical and technological progress. Baltagi and Moscone [4] found that health expenditure is a 'necessary good' in the 20 OECD countries of their sample, during the period 1971-2004. Wang [38] analyzed 31 countries for the period 1986-2007, and his study emphasized notable income levels differences between the countries. Finally, the study of Pekkurnaz [36] was mainly focused on the convergence process. Only a few countries converge. Jaunky and Khadaroo [24] explored this relationship for 28 African countries differentiating among public and private health expenditure. Meanwhile, a study for oil producing countries, including Mexico, was provided by Mehrara and Musai [31] for the period 1971-2009. Elmi and Sadeghi [13] studied 20 developed countries and the results also were positive.

The incorporation of structural changes in the analysis is crucial because it allows knowing the effect of the most representative economic events. Jewell, Lee, Tieslau and Strazicich [25] and Carrion-i-Silvestre [9] analyzed 20 OECD countries during the period 1960-1997. In both investigations, a non-linear relationship of equilibrium between health expenditure and per capita gross domestic product (GDP) was reported. On the other hand, Narayan, Narayan and Mishra [35] showed evidence on two structural changes in five Asian countries.

The sub-national level analysis has been less explored. One example in a regional context is the work of Chou and Wang [11] for 28 Chinese provinces along 1978-2004. Other example is Bukenya [7], who analyzed the southern states of U.S. and the findings showed that there was a relationship between health expenditure and income only for Georgia. Also, Moscone and Tosetti [33] included 49 states, and their evidence was favorable for cointegration; however, a strong heterogeneity across states was found.

In the case of Mexico, only very few studies outstand from the national perspective. Mayer, Mora, Cermeño, Barona and Duryeau [30] show a relationship between economic growth and life expectancy between 1965 and 1995. Other studies have found an increase in health expenditure that improves the health conditions of the population (Lustig [29] and Pinzon Florez, Reveiz, Idrovo and Reyes [37]). German-Soto and Fuentes [17] use methods of unit roots with structural breaks to analyze the relationship between income level and life expectancy, child mortality rates, and public health expenditure for the period 1940-2011.

4 Methodology: panel data unit root tests with structural breaks

For a panel data structure without structural break there is a number of tests to check stationarity. Some of them are Levin, Lin and Chu [28], Im, Pesaran and Shin [22]. For contexts of structural break are the statistics proposed by Im, Lee and Tieslau [21, 22] and Lee, Strazicich and Meng [27], among others. These lasts tests are based on a Lagrange Multiplier (LM) approach that first considers the possible presence of structural break and then evaluates the relevance of unit root.

The test of Im et al. [20] seeks to detect breakpoints in the level of the series; lately, Im et al. [21] incorporated two structural changes (in both slope and level). In short, consider a data generating process (DGP) as follows:

$$Y_t = \delta' Z_t + \varepsilon_t, \text{ where } \varepsilon_t = \beta\varepsilon_{t-1} + \upsilon_t \tag{1}$$

where Z_t contains the definition of the DGP (breaks in level, slope, or both), δ is a vector with until *R* structural breaks. The procedure in (1) is de-trended Y_t, so the measures are taken to correct serial correlation and heterogeneity. The de-trending procedure is expressed by $\tilde{S}_t$, which is given as $\tilde{S}_t = Y_t - \tilde{\psi} - Z_t\tilde{\delta}$ and $\tilde{\psi} = Y_t - Z_t\tilde{\delta}$. However, given the inclusion of structural breaks, the equation (1) is expressed as:

$$\Delta Y_t = \delta' Z_t + \phi\tilde{S}_{t-1} + \sum_{j=1}^{k} d_j \Delta\tilde{S}_{t-j} + \varepsilon_t \tag{2}$$

In spite that specification (2) corrects by serial correlation, the dependency on nuisance parameters is not removed with this de-trending procedure when trend breaks are present in the model (Lee et al. [27]), so an option is transforming the model as:

$$\tilde{S}_t^* = \begin{cases} \dfrac{T}{T_{B1}}\tilde{S}_t & \text{for} \quad t < T_{B1} \\ \dfrac{T}{T_{B2} - T_{B1}}\tilde{S}_t & \text{for} \quad T_{B1} < t < T_{B2} \\ \dfrac{T}{T - T_{BR}}\tilde{S}_t & \text{for} \quad T_{BR} < t < T \end{cases}$$

Therefore, $\tilde{S}_{t-1}$ is replaced with $\tilde{S}_{t-1}^*$ and now the regression (2) takes the form:

$$\Delta Y_t = \delta' Z_t + \phi\tilde{S}_{t-1}^* + \sum_{j=1}^{k} d_j \Delta\tilde{S}_{t-j} + \varepsilon_t \tag{3}$$

From (3), the extension to a data panel structure is immediate:

$$\Delta Y_{i,t} = \delta' Z_{i,t} + \phi_i\tilde{S}_{i,t-1}^* + \sum_{j=1}^{k} d_{ij} \Delta\tilde{S}_{i,t-j} + \varepsilon_{i,t} \tag{4}$$

The correct number and location of breaks are necessary to avoid a loss of power. However, determining of these two conditions are difficult due to the nuisance dependency problem. A solution by Lee et al. [27] is to adopt a two-step procedure. In the first step, significance of possible breaks is tested by using an *F*-test. Under the null hypothesis that no break exists, Lee et al. [27] estimate the following *F*-statistic:

$$F \text{ or } F^* = \frac{\left(SSR_0 - SSR_1(\lambda)\right)/k}{SSR_1(\lambda)/(T-q)} \tag{5}$$

where k is the number of restrictions, λ denotes the break, q is the number of regressors in (3), and SSR_0 (SSR_1) is the squared sum of residuals from the LM unit root test without breaks (with breaks).

In the second step, an exogenous LM unit root test with one break is used if the null of no break is rejected. The LM statistic and the null hypothesis to be tested are from (4), that is, $H_0 : \phi_i = 0$ for all i, while $H_1 : \phi_i < 0$ for some i. Meanwhile, the calculus of LM statistic is as follows (Lee et al. [27]):

$$\tilde{\tau}^* = -\frac{1}{2}\left[\sum_{i=1}^{R+1}\int_0^1 V_i(r)^2\, dr\right]^{-1/2} \tag{6}$$

where V_i corresponds to the limit of the partial sum of residuals, $\tilde{S}_t$. Lee et al. [27] emphasize the importance of firstly locate and determine the size of the break. Such procedure is performed considering the statistical contrast *MaxF**, which tries to examine the significance of the dummy variable. Once the *MaxF** test demonstrates the significance of the location of the break, then it should be proved if the series has a unit root under the test defined in (6). Because of break (or breaks) is unknown, its identification is through performing the following proposal:

$$\text{Max}F^* = \max_{\lambda} \frac{1}{8}\left(\sigma_\varepsilon^2/\sigma^2\right)\left[\left(\sum_{i=1}^{R+1}\int_0^1 V_i(r)^2\, dr\right)^{-1} - \left(\int_0^1 V_0(r)^2\, dr\right)^{-1}\right] \tag{7}$$

The distribution of the *F*-statistic in (7) is non-standard, so Lee *et al.* [27] develop the critical values of (7) for R = 1, 2. Simulation results suggest applying the *F*-test to a stationary or non-stationary process.

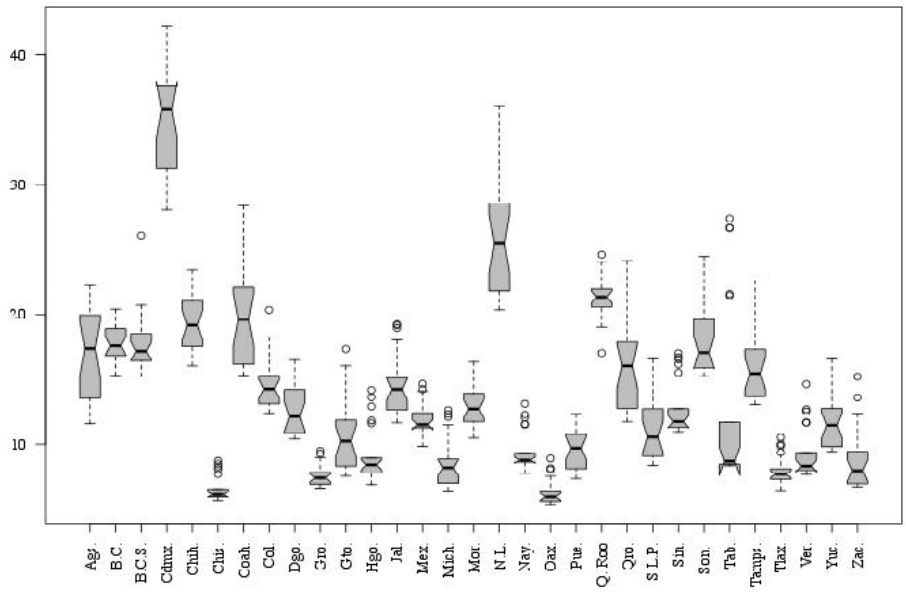
5 Database and statistical summary

Database covers the 1990-2011 period and it has three main series. The per capita gross state product (GSP), in real values of 1993, as proxy of income. As health indicators, we consider the public health expenditure in both per capita terms and as share of the GSP. The source for health indexes is the national institute of statistics (INEGI [23]), while for GSP series is the database published by German-Soto [16], also based on INEGI, but with the advantage that it is homogeneous in both time and space.

The Mexican regional system integrates 32 federal entities. As exploratory analysis, Fig. 1 shows the box plots of the per capita GSP along the period. By excluding Campeche state, we can have a better idea of the behavior of states.[3] Favorable performance is seen for Nuevo Leon (NL) and Ciudad de Mexico (Cdmx), while weak performance is appreciated on Guerrero (Gro), Oaxaca (Oaxaca) and Chiapas (Chis).

[3] Campeche is a potential outlier because of its high oil production, biasing our perception on income and its spatial distribution.

Fig. 1. Mexican states: per capita gross state product, 1990-2011.



Source: own elaboration.

In terms of health, the Mexican state performance is not less heterogeneous. Initially, the federal government headed the state health system; however, the public administration implemented a great reform along 1985-1987, which consisted on the decentralization of the medical and financial resources to states. At the start of the reform, only fourteen entities signed the pact of decentralization. The agreement was signed by all state only until 1996, when they joined the decentralization reform (Merino [32]). A new health sector reform took place by the 2000's decade: the pilot program of 'Seguro Popular' (SP: popular security), which started by 2001. The purpose of this program has been to provide universal health coverage, especially for that economically disadvantaged population sector. By 2004 all states were integrated into the SP (Frenk, Sepulveda, Gomez, Dantés and Knaul [14]).

The process of decentralization implied to transfer the financial autonomy to the states, but also led to a significant increase in the state health expenditure along 1996-1999.[4] Meanwhile, at the time of the SP program was launched, the health expenditure suffered a significant reduction.[5] To carry out these reforms the Mexican government obtained financing from the World Bank.

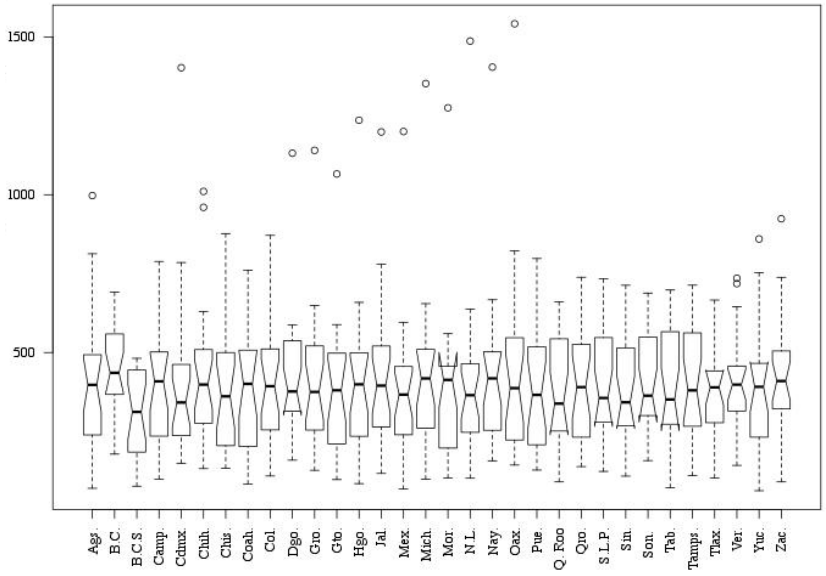
Fig. 2 shows the box plots of public health expenditure, it is possible to observe a low dispersion. Differently, the indicator of health expenditure as a share of GSP presents higher dispersion. According to Fig. 3 major volatility is shown across states,

[4] In average, the absolute increment across the states was of 34 percent.

[5] Around of 12 percent for each year of the period 2001 to 2004.

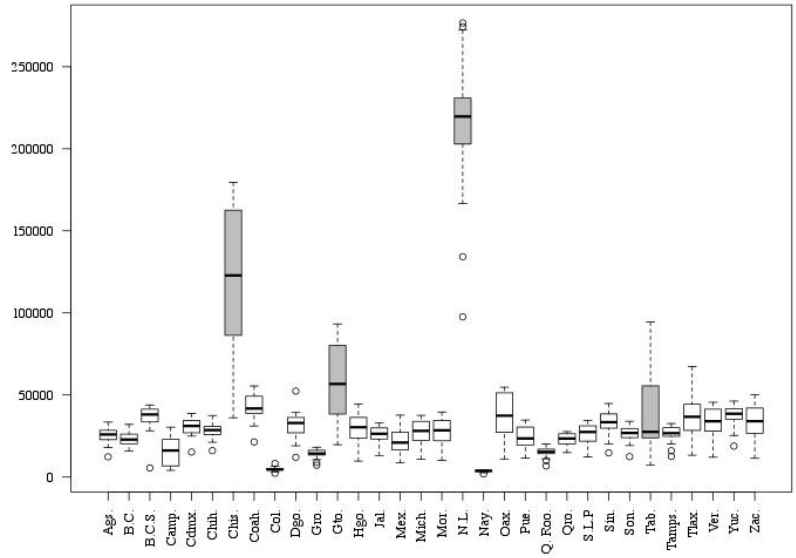
making clearer the performance of entities such as Nuevo Leon (NL), Chiapas (Chis), Guanajuato (Gto), and Tabasco (Tab).

Fig. 2. Mexican states: per capita health public expenditure, 1990-2011.



Source: own elaboration.

Fig. 3. Mexican states: health public expenditure (% of GSP), 1990-2011.



Source: own elaboration.

6 Estimates and results

The statistical tests were executed according to Im et al. [21] and Lee et al. [27]. Routines were programmed in Gauss software to consider the several alternatives of estimation implied by the contrasts. They are estimates through fixed effects and without fixed effects, as well as options of one and two breaks in both level and slope. We use the Bayesian information criterion to select among them.

Firstly, the MaxF* is estimated. If it is significant (at the 10 percent of significance), then we evaluate the possible rejection of unit root (see Lee et al. [27]). Only the results for those states that satisfy these two conditions are reported. Table 1 is for income and public health expenditure in per capita terms (HPEpc). Table 2 is for public health expenditure as a share of GSP (HPEgsp).

Results with HPEpc confirm the existence of a relationship in almost all Mexican states. Exceptions are Aguascalientes, Baja California, Baja California Sur, San Luis Potosi, Sonora, and Tamaulipas. A majority of states selected the fixed effects model with two structural changes in the slope as the best option. Some estimated breakpoints highlight the presence of structural breaks in those years linked to the reforms and economic crises. For example, some states were affected by two health reforms, the first in 1997-1999, due to the decentralization of health services of the federation to the states, and the second, during 2004-2006, due to SP program. The states that experienced the effect of these reforms were Guanajuato, Tlaxcala, Coahuila, Michoacan, Colima, Veracruz, Yucatan, and Campeche (see Table 1).

Table 1. Results between income and per capita health (significant cases only)

Federal entity	TB1	TB2	LM ($\tilde{\tau}$*)	*p*	Federal entity	TB1	TB2	LM ($\tilde{\tau}$*)	*p*
Campeche	1997	2005	-1289.0*	2	Morelos	2006		-3837.9*	3
Coahuila	1998	2006	-10.35*	3	Nayarit	2006		-4.817*	4
Colima	1998	2005	-9.777*	4	Nuevo León	2000	2008	-5.807*	4
Chiapas	1997	2001	-1783.2*	4	Oaxaca	2009		-6.287*	3
Chihuahua	1997	2001	-3149.2*	1	Puebla	2009		-4.202*	4
Cd. de México	1998	2002	-1720.2*	3	Querétaro	2006	2009	-4791.2*	0
Durango	1999	2003	-1498.9*	3	Quintana Roo	2008		-3.742*	1
Guanajuato	1999	2005	-10.49*	4	Sinaloa	1997	2009	-5.227*	1
Guerrero	2000	2006	-6.322*	4	Tabasco	1997	2009	-4.757*	4
Hidalgo	2002	2005	-6.996*	2	Tlaxcala	1999	2006	-9.455*	4
Jalisco	2003	2007	-684.4*	4	Veracruz	1997	2004	-5.829*	3
México	2004	2008	-4112.7*	1	Yucatán	1997	2004	-7.216*	4
Michoacán	1997	2005	-8548.8*	4	Zacatecas	2001	2008	-6.010*	0

Notes: the suscript * indicates significant at 10%. TB1 and TB2 denote the estimated breakpoints. *p* is the selected optimal lag. The critical values of the LM-test are: for one break -4.604 (1%), -3.95 (5%), and -3.635 (10%); for two breaks -5.365 (1%), -4.661 (5%), and -4.338 (10%) –see Im et al. [21].

Other states, as Sinaloa and Tabasco, reported structural breaks during the decentralization process and along the international crisis (2008-2009). Others more reported structural breaks due to SP and economic crisis, as Mexico and Queretaro. Table 1 also reflects a set of reasons by the which states select different structural changes. Some Mexican states exhibited one break, at least, during the period of decentralization. They

are Chiapas, Chihuahua, Durango, and Mexico City. Other are by SP reform as Guerrero, Hidalgo, Morelos, Nayarit, and Jalisco. Finally, a third group of states highlight the effects of the economic crisis. This is the case of Nuevo Leon, Zacatecas, Quintana Roo, Oaxaca, and Puebla. It is possible to identify other unexplained breaks, but they seem to be consequence of proper and particular situations into each state.

About the second relationship (income-HPEgsp), Table 2 presents the main results. There you can see that seven states do not have a balanced relationship. They are Aguascalientes, Baja California, Baja California Sur, Quintana Roo, Tabasco, Tamaulipas, and Veracruz. The remaining states confirm the existence of a relationship between income and HPEgsp. In this case, most of states selected the model of two structural breaks occurring in the slope without fixed effects.

An analysis of the breakpoints indicates that some states were doubly affected by the health system reforms that have taken place around 1997-1999 and 2005-2007. They are Campeche, San Luis Potosi, Sonora, Michoacan, Guanajuato, and Morelos. Other states were also affected by at least one reform. For example, Ciudad de Mexico, Yucatan, Chiapas, Chihuahua, Durango, and Zacatecas are states impacted by the first reform (decentralization of health), while Guerrero, Nayarit, Sinaloa, and Mexico are examples where the second break, due to SP, was clearer. In change, only a few federal entities reported a structural break that can be explained by the international crisis. They are Queretaro, Puebla, Oaxaca, and Nuevo Leon.

Table 2. Results between income and health as % of the GSP (significant cases only)

Federal entity	TB1	TB2	LM ($\tilde{\tau}$*)	*p*	Federal entity	TB1	TB2	LM ($\tilde{\tau}$*)	*p*
Campeche	1997	2006	-2125.6*	2	Morelos	1999	2006	-11.83*	4
Coahuila	1997		-6.337*	3	Nayarit	2000	2007	-6.299*	4
Colima	2001		-1071.5*	4	Nuevo León	2008		-757.1*	1
Chiapas	1997	2000	-660.0*	4	Oaxaca	2009		-14.96*	3
Chihuahua	1997	2001	-1046.4*	3	Puebla	2000	2009	-6.052*	2
Cd. de México	1998	2002	-743.1*	3	Querétaro	2004	2009	-1241.5*	0
Durango	1997	2000	-6.625*	4	San Luis P.	1997	2006	-7.036*	4
Guanajuato	1999	2005	-10.64*	4	Sinaloa	2000	2007	-4.416*	3
Guerrero	2000	2006	-4.670*	4	Sonora	1997	2007	-9.481*	0
Hidalgo	2002		-5.118*	4	Tlaxcala	1999		-4.998*	1
Jalisco	2003		-6.938*	4	Yucatán	1997	2002	-6.147*	1
México	2003	2006	-2525.5*	0	Zacatecas	1998	2003	-4.604*	4
Michoacán	1998	2005	-3742.6*	4					

Notes: the suscript * indicates significant at 10%. TB1 and TB2 denote the estimated breakpoints. *p* is the selected optimal lag. The critical values of the LM-test are: for one break -4.604 (1%), -3.95 (5%), and -3.635 (10%); for two breaks -5.365 (1%), -4.661 (5%), and -4.338 (10%) –see Im et al. [21].

As we can see both indicators of health expenditure sustain an equilibrium relationship in the long-run, but it was interrupted by reforms and crises occurred in the period. So, the results highlight two conclusions, at least. First, the equilibrium path exists also for the Mexican regional case, according to the theory, but in a context where macroeconomic events broke the balance. The second conclusion affirms that it is not possible to arrive at this equilibrium if the presence of structural break is not considered, therefore the implications of policy recommendations could be biased.

7 Conclusions

Once the cross-section dependence is accounted for, the study finds that the gap between real per capita income and public health expenditure can be a stationary process evolving around a broken trend with until two structural changes. It means three regime shifts. This result reinforces the theoretical expectative on the income and health relationship.

From the stationarity result, meaningful advice arises for investigators and policy makers when the structural breaks are considered. For researchers, our results indicate that the standard panel regression estimation is biased if the presence of structural breaks is not considered in the model. For policy makers, it is evidence of how reforms can promote and extend the health coverage.

The Mexican economy has exhibited very low rates of growth in the last twenty years. This performance is reproduced at the state level. Also, the public health expenditure has seen meaningful reductions. Consequently, both variables are cointegrated. It means that they evolve in the same direction. We sustain that it should concern because if population income level is not growing, then one way to reduce this adverse result could be improving the health conditions, and this last not seem to be happening. Besides, some regions are lagging regarding income, while others enhance their performance. So, heterogeneity is augmenting, a not desired result by the government because the costs implied by a major inequality.

The non-stationaries cases are material for future investigation. Some possible explanations indicate that more than two structural breaks could characterize them, although it is also possible that reductions in public health expenditure were stronger in comparison to those effectively stationaries states, and so they are widening the gap. Besides, it is possible that reductions in the public health expenditure obey to augments in the private health expenditure. Some features of the current economic model seem to indicate that this is so, but private health expenditure is another line of investigation for which we have not sufficient data to test it. Therefore, it was left as future research line.

8 References

1. Adeniyi O and Abiodun L (2011) Health Expenditure and Nigerian Economic Growth. European Journal of Economics, Finance and Administrative Sciences 30(30):125–129.
2. Apergis N (2013) Health Expenses: Evidence from the Club Clustering Approach. International Advances in Economic Research 19(4):399–407.
3. Apergis N and Padhi P (2013) Health Expenses and Economic Growth: Convergence Dynamics across the Indian States. International Journal of Health Care Finance and Economics 13(3-4):261–277.
4. Baltagi BH and Moscone F (2010) Health Care Expenditure and Income in the OECD Reconsidered: Evidence from Panel Data. Economic Modelling 27(4):804–811.
5. Barro R (1996) Health and Economic Growth, Harvard University Press, Cambridge.
6. Bloom D, Canning D and Sevilla J (2004) The Effect of Health on Economic Growth: A Production Function Approach. World Development 32(1):1–13.

7. Bukenya J (2009) Do Fluctuations in Health Expenditure Affect Economic Growth? The Open Economics Journal 2:31–38.
8. Caporale GM, Cunado J, Gil-Alana L and Gupta R (2015) The Relationship between Health Care Expenditures and Diposable Personal Income in the US States: A Fractional Integration and Cointegration Analysis. Department of Economics Working Paper Series 2015-32, University of Pretoria.
9. Carrion-i-Silvestre JL (2005) Health Care Expenditure and GDP: Are They Broken Stationary? Journal of Health Economics 24(5):839–54.
10. Chang K and Yung-Hsiang Y (2006) Economic Growth, Human Capital Investment, and Health Expenditure: A Study of OECD Countries. Hitotsubashi Journal of Economics 47(1):1–16.
11. Chou WL and Wang Z (2009) Regional Inequality in China's Health Care Expenditures. Health Economics 18(S2): S137-S146.
12. Dreger C and Reimers HE (2005) Health Care Expenditures in OECD Countries: A Panel Unit Root and Cointegration Analysis. IZA D.P. No. 1469, January 2005.
13. Elmi ZM and Sadeghi S (2012) Health Care Expenditures and Economic Growth in Developing Countries: Panel Co-integration and Causality. Middle-East Journal of Scientific Research 12(1):88–91.
14. Frenk J, Sepúlveda J, Gómez-Dantés O and Knaul F (2003) Public Health Evidence-Based Health Policy: Three Generations of Reform in Mexico. The Lancet. 362(9396):1667–1671.
15. Fuchs VR (1972) Front Matter, Essays in the Economics of Health and Medical Care. In: VR Fuchs (ed) Essays in the Economics of Health and Medical Care, NBER: Columbia University Press.
16. German-Soto V (2005) Generación del Producto Interno Bruto Mexicano por Entidad Federativa, 1940-1992. El Trimestre Económico 72(3):617–653.
17. German-Soto V and Fuentes Castillo ME (2016) The Long-Run Relationship among Health and Income in Mexico, 1940-2011. Journal of Reviews on Global Economics. 5:131–144.
18. Grossman GM (1972) On the Concept of Health Capital and the Demand for Health. The Journal of Political Economy 80(2):223–255.
19. Howitt P (2005) Health, Human Capital, and Economic Growth: A Schumpeterian Perspective. Seminar Paper for the Pan American Health Organization.
20. Im KS, Lee J and Tieslau M (2005) Panel LM Unit Root Tests with Level Shifts. Oxford Bulletin of Economics and Statistics 67(3):393–419.
21. Im KS, Lee J and Tieslau M (2010) Panel LM Unit Root Tests with Trend Shifts. FDIC Center for Financial Research Working Paper No. 2010-1, March 2010.
22. Im KS, Pesaran MH and Shin Y (2003) Testing for Unit Roots in Heterogeneous Panels. Journal of Econometrics 115(1):53–74.
23. Instituto Nacional de Estadística y Geografía (INEGI) (2013) Banco de Información Económica, www.inegi.org.mx. Accessed 15 June 2017.
24. Jaunky VC and Khadaroo AJ (2008) Health Care Expenditure and GDP: An African Perspective. Applied Econometrics and International Development 8(1):131–146.
25. Jewell T, Lee J, Tieslau M and Strazicich MC (2003) Stationarity of Health Expenditures and GDP: Evidence from Panel Unit Root Tests with Heterogeneous Structural Breaks. Journal of Health Economics 22(2):313–323.
26. Kiymaz H, Akbulut Y and Demir A (2006) Tests of Stationarity and Cointegration of Health Care Expenditure and Gross Domestic Product: An Application to Turkey. European Journal of Health Economics 7(4):285–289.
27. Lee J, Strazicich MC and Meng M (2012) Two-step LM Unit Root Tests with Trend-Breaks. Journal of Statistical and Econometric Methods 1(2):81–107.

28. Levin A, Lin CF and Chu J (2002) Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. Journal of Econometrics 108(1):1–24.
29. Lustig N (2007) Investing in Health for Economic Development: The Case of Mexico. In: G Mavrotas and Shorrocks A (eds) Advancing Development: Core Themes in Global Economics. Palgrave Macmillan, London.
30. Mayer D, Mora H, Cermeño R, Barona AB and Duryeau S (2003) Salud, Crecimiento y Distribución del Ingreso en América Latina y el Caribe: Estudio de los Factores Determinantes y del Comportamiento Regional y Local. In: Organización Panamericana de la Salud (ed) Invertir en Salud: Beneficios Sociales y Económicos. Organización Panamericana de la Salud, Washington.
31. Mehrara M and Musai M (2011) Granger Causality between Health and Economic Growth in Oil Exporting Countries. Interdisciplinary Journal of Research in Business 1(8):103–108.
32. Merino G (2003) Descentralización del Sistema de Salud en el Contexto del Federalismo. In: Knaul FM and Nigenda G (eds) Caleidoscopio de la Salud. De la Investigación a las Políticas y de las Políticas a la Acción. Funsalud, México.
33. Moscone F and Tosetti E (2010) Health Expenditure and Income in the United States. Health Economics 19(12):1385–1403.
34. Mushkin SJ (1962) Health as an Investment. Journal of Political Economy 70(5):129–157.
35. Narayan S, Narayan PK and Mishra S (2010) Investigating the Relationship between Health and Economic Growth: Empirical Evidence from a Panel of 5 Asian Countries. Journal of Asian Economics 21(4):404–411.
36. Pekkurnaz D (2015) Convergence of Health Expenditure in OECD Countries: Evidence from a Nonlinear Asymmetric Heterogeneous Panel Unit Root Test. Journal of Reviews on Global Economics 4:76–38.
37. Pinzon-Florez CE, Reveiz L, Idrovo AJ and Reyes H (2014) Gasto en Salud, la Desigualdad en el Ingreso y el Índice de Marginación en el Sistema de Salud de México. Revista Panamericana de Salud Pública 35(1):1–7.
38. Wang KM (2011) Health Care Expenditure and Economic Growth: Quantile Panel-Type Analysis. Economic Modelling 28(4):1536–1549.
39. Xu K, Saksena P and Holly A (2011) The Determinants of Health Expenditure. A Country-Level Panel Data Analysis. World Health Organization, Working Paper December 2011.
40. Zon AH van and Muysken J (2003) Health as a Principal Determinant of Economic Growth. MERIT–Infonomics Research Memorandum Series, 2003-021, Maastricht.

Forecasted trends for cardiovascular disease in England and Wales to 2040 and impact of reduction in smoking prevalence: a Markov modelling study

Authors: **Sara Ahmadi-Abhari[1], Piotr Bandosz[2,3], Maria Guzman Castillo[2], Hannah Whittaker[4], Martin J Shipley[1], Mika Kivimäki[1,5], Simon Capewell[2], Martin O'Flaherty[2], Eric J Brunner[1]**

Affiliations:

1. Department Epidemiology & Public Health, University College London, United Kingdom
2. Department of Public Health and Policy, University of Liverpool, United Kingdom
3. Department of Prevention and Medical Education, Medical University of Gdansk, Poland
4. Imperial College London, United Kingdom
5. Clinicum, Faculty of Medicine, University of Helsinki, Finland

Background:

Despite the decline in cardiovascular disease incidence and mortality, this condition remains a major cause of morbidity and mortality across the globe. Forecasts for prevalence of cardiovascular disease are vital in defining future healthcare needs. Continuing increase in life-expectancy results in a growing number of people at risk of cardiovascular disease. On the other hand however, risk of cardiovascular disease at any given age is declining. Given the opposing effects, forecasts for prevalence of cardiovascular disease can be improved using a dynamic modelling approach that integrates calendar trends in incidence of cardiovascular disease and mortality and can be used to estimate the impact of primary prevention interventions on future numbers with this condition.

Methods:

A dynamic multi-state Markov model IMPACT-BAM (Figure 1) was developed and validated to track the progression of the population of England and Wales aged 35+ through states of cardiovascular disease, cognitive impairment, dementia, and disability to death, enabling the burden of disease and life-expectancy to be forecasted to 2040 (*1,2*). Population estimates and mortality data were obtained from the UK Office for National Statistics (ONS). Data from the English Longitudinal Study of Ageing (ELSA), a representative panel study with 6 waves of data across 2002-2013, was used to inform the baseline prevalence estimates and transition probabilities in the model. In this abstract we present forecasts of cardiovascular disease and mortality according to various scenarios of reduction in prevalence of smoking at population level.

Cardiovascular disease in ELSA was ascertained by self-reported doctor diagnosis of myocardial infarction, stroke, angina, coronary artery bypass grafting, or death from cardiovascular causes. Incidence of cardiovascular disease was defined as first ever record of disease or intervention for each participant. Cognitive impairment was ascertained by low scores on the cognitive function tests or based on the Informant Questionnaire for Cognitive decline (IQCODE). Functional impairment (disability) was defined as inability to independently perform one or more basic activities of daily living including getting in or out of bed, walking across a room, bathing or showering, using the toilet, dressing, cutting food and eating. Dementia was defined as the co-occurrence of cognitive and functional impairment (according to DSM-IV criteria) or self-reported doctor diagnosis of dementia. Date and cause of death for ELSA study participants were obtained by data linkage with the UK Office for National Statistics (ONS).

IMPACT-BAM was initially populated using age- and sex-specific prevalence estimates for each of the eight health states in the model. The year 2006 (mid-point of the ELSA data collection period) forms the baseline of the model. One-year transition probabilities are applied at each 1-year iteration to predict number of deaths and prevalence of each health state at the next calendar year. The age- and sex-specific transition probabilities were obtained from 2-year epochs of ELSA. All transition probabilities entered in IMPACT-BAM are calendar-time specific. To obtain calendar trends for mortality rates, P-spline smoothed lines were fitted to CVD and non-CVD mortality rates in each 5-year age band from 1990 to 2016 by sex and projected to the future. The calendar trend in cardiovascular incidence was assumed to be parallel to that for cardiovascular mortality as observed in ELSA. The calendar trend in cognitive impairment and dementia was calculated in 2-year epochs in ELSA and corrected for the selective dropout of participants with dementia using the joint modelling of longitudinal and time to event method. The effect of calendar time for each of the transition probabilities was imposed on the transition probabilities for the year 2006 to obtain transition probabilities for future years. In each 1-year iteration of the model, numbers of men and women reaching age 35 are obtained from ONS predictions and are entered into the model.

The impact of smoking on cardiovascular disease and mortality to 2040 was estimated under three scenarios: S1. Prevalence of smoking remains at 2017 levels of 19% in men and 17% in women to 2040 (baseline scenario); S2. Prevalence of smoking is reduced progressively to 9% by 2020 and 5% by 2025; S3: hypothetical scenario in which smoking is eliminated by 2020 (prevalence zero). To estimate the impact of change in smoking prevalence, transition probabilities in IMPACT-BAM were modified according to the change in Population Attributable Risk Fraction (PARF), a composite measure of the prevalence of smoking and the relative risk of smoking with each of the health conditions in the model. Systematic reviews and meta-analyses of published literature provided relative risks of CVD, dementia, disability, and mortality among smokers by age and sex.

Monte-Carlo simulation (1000 iterations) was used to obtain the impact of parameter uncertainty on model outputs, presented as 95% uncertainty intervals. To validate methods, definitions, and assumptions, we ran

the model starting in 2006 to predict prevalence of cardiovascular disease in 2011, and compared that with the prevalence of CVD observed in the Health Survey for England (HSE)-2011.

The IMPACT-BAM model was implemented in R statistical software and a package specifically written for it by author PB. Stata-14 (StataCorp 2015. College Station, TX: StataCorp LP) was used for data management and regression analysis to derive model inputs. The R package "JM" was used for joint modelling of longitudinal and time to event data.

Results:

Age-specific cardiovascular disease incidence in ELSA and mortality rates declined over the past decade in both men and women. Using 2006 data, predicted prevalence of cardiovascular disease using IMPACT-BAM was consistent with observed prevalence of cardiovascular disease observed in HSE in 2011 (Figure 2). If trends in incidence of disease and mortality continue to the future as observed over the past two decades, the prevalence and numbers of people living with cardiovascular disease will decline in both men and women to 2040. Forecasted numbers of men and women living with cardiovascular disease (prevalence %) in 2017, 2020, 2030 and 2040 are 3.2 Million (9.6%), 3.0 M (8.9%), 2.2 M (5.8%), and 1.4 M (3.5%) respectively (Figure 3).

If prevalence of smoking remains at the current levels of 19% in men and 17% in women (S1), there will be an estimated 952,000 (95% Uncertainty Interval 808,000-1,153,000) deaths from cardiovascular causes and 10,624,000 (95%UI 9,562,000-11,776,000) deaths from non-cardiovascular causes between 2017 and 2040. 35,000 (95%UI 29,000-43,000) CVD deaths and 408,000 (95%UI 373,000-443,000) non-CVD deaths can be prevented or postponed if prevalence of smoking is reduced to 9% by 2020 and 5% by 2025 (S2). Corresponding figures are 84,000 (95%UI 68,000-105,000) and 843,000 (95%UI 764,000-907,000) respectively if smoking is eliminated by 2020 (S3).

Under the stable smoking prevalence scenario (S1) life expectancy at 35 (LE) will increase from 52.2 years (95% UI 49.8, 54.1) in 2017 to 56 years (95%UI 52.8, 59.2) by 2040. Life expectancy at 35 in 2040 is estimated to increase by an additional 1.1 year under S2 (LE 57.1 (95%UI 51.7, 62.7)) and by 2.3 years under S3 (LE 58.3 (95%UI 51.6, 65.2)). The increase in numbers of people living with cardiovascular disease due to increased life expectancy (i.e., higher numbers of people at risk of disease and higher survival with cardiovascular disease) was offset by lower risk of cardiovascular disease due to reduced prevalence of smoking. Therefore, little change in the numbers of people living with cardiovascular disease or prevalence of the condition was observed under the different scenarios for smoking prevalence at population level.

Conclusion:

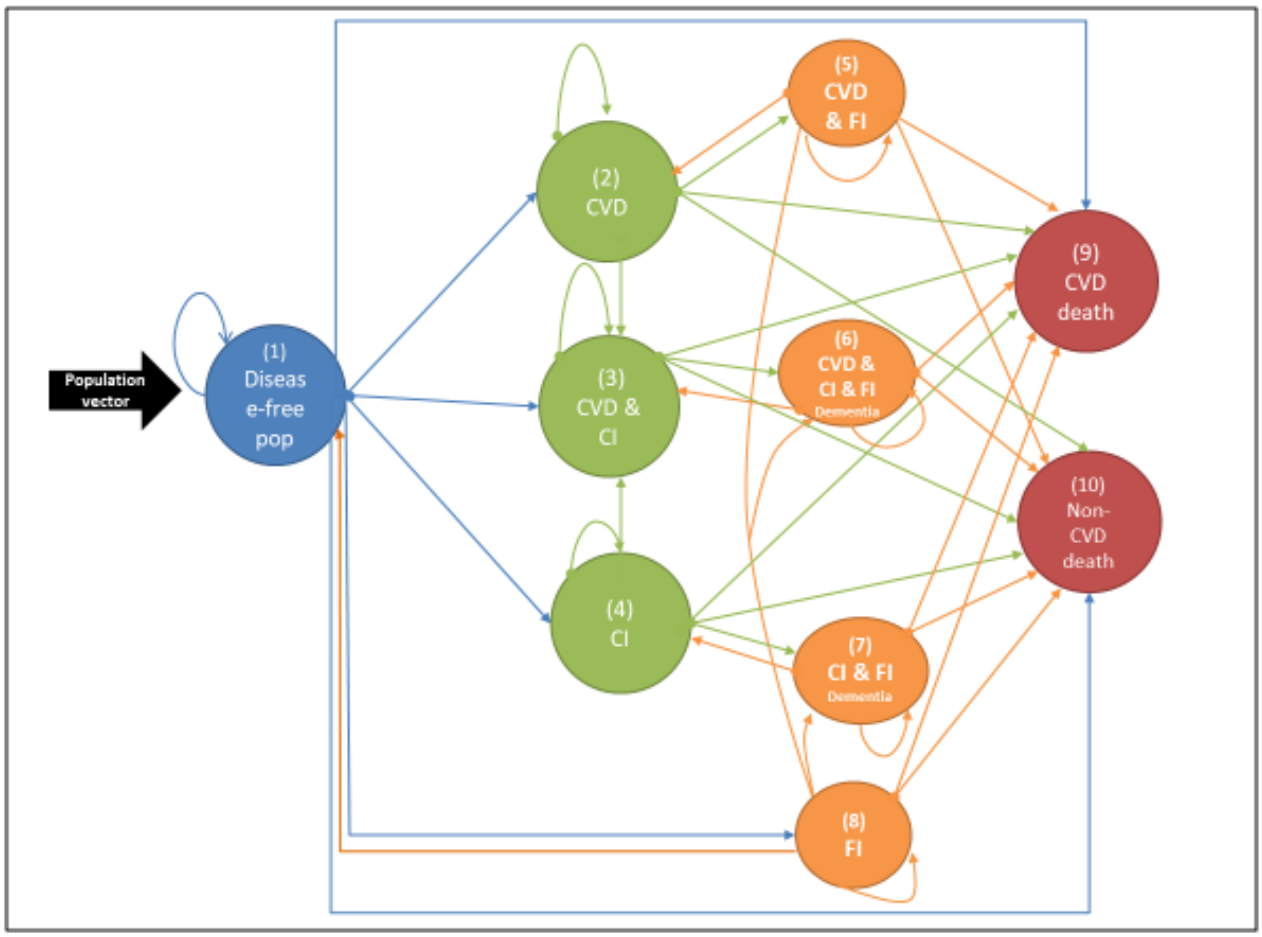
If public health efforts continue to the future and cardiovascular disease incidence and mortality continue to decline, the numbers of people living with, and needing care for, cardiovascular disease will half in the next two decades. If prevalence of smoking is reduced further there will be a remarkable increase in life-expectancy without an increase in prevalence of cardiovascular disease, compatible with the notion of compression of morbidity.

References:

1. Ahmadi-Abhari S, Guzman Castillo M, Bandosz P, Shipley MJ, Muniz-Terrera G, Singh-Manoux A, Kivimäki M, Steptoe A, Capewell S, O'Flaherty M, Brunner EJ. Temporal trend in dementia incidence since 2002 and projections for prevalence in England and Wales to 2040: a modelling study. British Medical Journal (BMJ), 2017; 358:j2856.

2. Guzman Castillo M, Ahmadi-Abhari S, Bandosz P, Shipley MJ, Capewell S, Singh-Manoux A, Kivimäki M, Steptoe A, Brunner EJ, O'Flaherty M. Trends in disability and life expectancies in England and Wales, 2015-2025: a modelling study. Lancet Public Health, 2017; 2(7):e307-e313.

Figure 1: IMPACT-Better Ageing Model (IMPACT-BAM) structure



Abbreviations: CVD: Cardiovascular disease; CI: Cognitive impairment; FI: Functional impairment (disability)
Arrows represent transition probabilities

Figure 2: IMPACT-BAM predicted cardiovascular disease prevalence compared with observed estimates from the Health Survey for England in 2011. Error bars represent 95% uncertainty intervals.

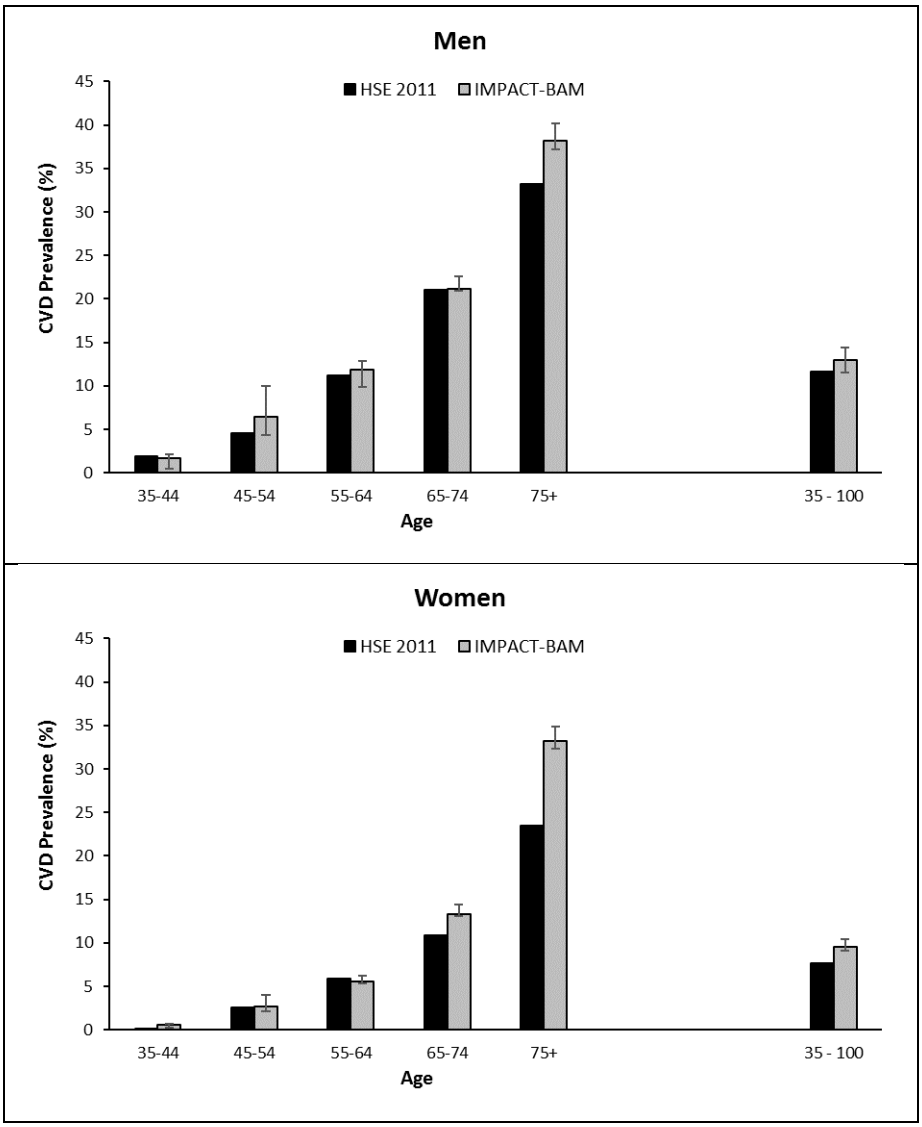
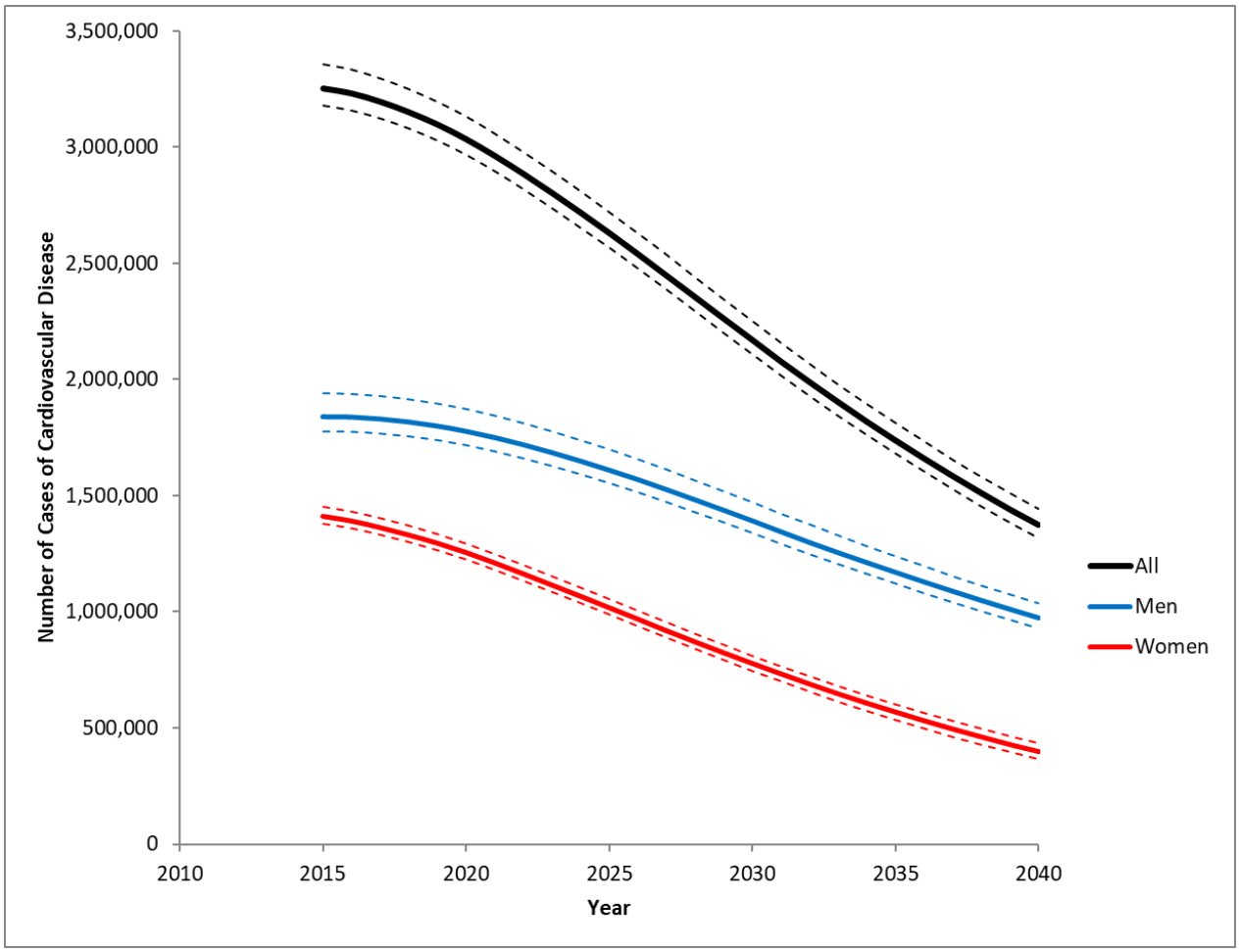


Figure 3: Projected number of people living with cardiovascular disease in England and Wales using IMPACT-BAM, 2015-2040. Dashed lines represent 95% uncertainty intervals.



Forecasting in qPCR procedure by means of hyperbolastic stochastic model

Antonio Barrera[1], Patricia Román-Román[2], and Francisco Torres-Ruiz[2]

[1] Departamento de Matemática Aplicada, E.T.S.I. Informática, Bulevar Louis Pasteur, 35, Campus de Teatinos, Universidad de Málaga, 29071, Málaga, Spain
antonio.barrera@uma.es

[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, s/n, Campus de Fuentenueva, Universidad de Granada, 18071, Granada, Spain
{proman, fdeasis}@ugr.es

Abstract. Quantitative Polymerase Chain Reaction (qPCR, real-time PCR) is a method derived from the classical PCR and used in molecular biology to amplify nucleic acids (commonly DNA or RNA) by taking advantage of the polymerase enzyme (see [1]).

Amplification is obtained by three steps: denaturation, annealing and elongation, repeated along a number of cycles. Unlike traditional techniques, qPCR monitors amplification at the same time it is produced, resulting in a collection of data indexed by time (cycles).

The procedure of monitoring amplification is carried out by the detection and measurement of fluorescence, induced by dyes interacting with molecules or labeled oligonucleotids. In particular, the cycle at which certain threshold of amplicon DNA is reached, is quantified by qPCR technique.

In [2], criteria for the mathematical development of qPCR technique, focusing on the threshold of log-fluorescence, is established. Absolute quantification is achieved using a standard curve constructed by amplification of known amounts of target DNA and plotting the values obtained for the threshold cycle against target DNA concentration. This and other recent studies (see, for example, [3]) show the interest in a mathematical approach to qPCR.

However, this dynamical phenomenon, developed at molecular level, is clearly subject to random influence produced by internal and external conditions, such us the failure of any step of the procedure, as well as the accuracy of measurement instruments. In addition, particular laboratory equipment is necessary in order to obtain data.

These reasons and the mathematical background underlying this technique, allow us to consider the develop and application of a stochastic model describing this phenomenon. Indeed, stochastic diffusion processes have become an appropriate tool to explain dynamical phenomena, taking into account random influence produced by internal and external conditions.

In particular, models based on logistic curve could deal well with the dynamics of fluorescence in qPCR. The type I hyperbolastic growth curve (see [4]) have been successfully applied to stem cell growth, epidemiological dynamics or tumor growth ([4, 5]), due to their flexible condition and the ability to deal with NP-hard problems.

In [6], a stochastic model based on the type I hyperbolastic curve is developed. Firefly algorithm (presented in [7], with successful applications and variants in [8, 9]) is applied to deal with the problem of maximum likelihood estimation for parameters.
In this work, we use this model in order to analyze the evolution of fluorescence and forecasting the production of replicated acids. The same model can also be used to determine first-passage time variable and applying it to calculate first cycle at which certain threshold is achieve.

References

1. Higuchi, R., Fockler, C., Dollinger, G., Watson, R.: Kinetic PCR: Real time monitoring of DNA amplification reactions. Biotechnology **11**, 1026-1030 (1993)
2. Rutledge RG, Côté C.: Mathematics of quantitative kinetic PCR and the application of standard curves. Nucleic Acids Research **31**(16) e93 (2003)
3. Spiess, A.-N., Rdiger, S., Burdukiewicz, M., Volksdorf, T., Tellinghuisen, J.: System-specific periodicity in quantitative real-time polymerase chain reaction data questions threshold-based quantitation. Scientific Reports **6** (2016)
4. Tabatabai, M., Williams, D., Bursac, Z.: Hyperbolastic growth models: theoryand application. Theoretical Biology and Medical Modelling **2**:14 (2005)
5. Eby, W. M., Tabatabai, M. A., Bursac, Z.: Hyperbolastic modeling of tumor growth with a combined treatment of iodoacetate and dimethylsulphoxide. BMC Cancer **10**:509 (2010)
6. Barrera, A., Román-Román, P., Torres-Ruiz, F.: A hyperbolastic type-I diffusion process: Parameter estimation by means of the firefly algorithm. Biosystems **163**, 11-22 (2018)
7. Yang, X.S.: Nature-Inspired Metaheuristic Algorithms. Luniver Press (2008)
8. Horng, M.-H.: Vector quantization using the firefly algorithm for image compression. Expert Systems with Applications **39**, 1078-1091 (2012)
9. Wang, H., Wang, W., Zhou, X., Sun, H., Zhao, J., Yu, X., Cui, Z.: Firefly algorithm with neighborhood attraction. Information Sciences **382-383**, 374-387 (2017)

Effects of Electrical Stimulation on Cortical Phase Synchronization as a Measure of Excitability

Farrokh Manzouri[1, 2], Matthias Dümpelmann[1, 2], Christian Meisel[3] and Andreas Schulze-Bonhage[2]

[1]Biomedical Microtechnology Lab, Faculty of Engineering, University of Freiburg, Freiburg, Germany
[2]Epilepsy center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[3]Department of Neurology, University Clinic Carl Gustav Carus, Dresden, Germany

Introduction:

Epilepsy is one of the most common neurological diseases with about 1% of population suffering from it. Epileptic seizures are episodes of excessive or abnormal synchronous neuronal activity of the brain. About one third of these patients are resistant to pharmaceutical therapy. Because of the ongoing recurrence of seizures, these patients suffer from severe implications on the longevity and quality of life. One emerging new treatment modality for drug-resistant epilepsy patients is direct electrical stimulation of the epileptogenic zone. Studies in hippocampal brain slices in vitro and in human patients with temporal lobe epilepsy have shown that cortical electrical stimulation can eliminate epileptiform discharges, including epileptic seizures. So far, few preliminary studies tested the ability of electrical stimulation to prematurely terminate seizures in epilepsy patients (see for example Fountas et al. 2005; Kossoff et al. 2004; Osorio et al.2005). In addition, Lesser et al. (1997) have reported that neocortical electrical stimulation eliminated evoked afterdischarges in patients with intractable extra temporal neocortical epilepsy. In spite of growing evidence regarding an antiepileptic effect of electrical stimulation, the use of electrical stimulation for treating intractable epilepsy is in its early stages of development.

Animal models have also been used to study the effects of either low or high frequency stimulation on epileptiform activity under a wide variety of conditions in hippocampal brain slices and in other targets. In addition, low-frequency electric stimulation was shown to have an inhibitory effect on epileptic foci in kindled rats. For example, Albensi et al., (2004) reported that prolonged electrical stimulation in the hippocampus may be effective in reducing seizure frequency in patients with epilepsy and that suppression by low frequency stimulation may be mediated by long-term depression (LTD).

To assess the level of excitability under in vivo experimental conditions, usually the size of the evoked cortical activity to external perturbations is measured, which in case of epilepsy patients can be done by electrical stimulation. Recent insights into

ongoing cortical activity have identified global levels of phase synchronization as a measure that characterizes normal levels of excitability and quantify any deviation therefrom.

Meisel et al., (2015) proposed the mean of phase synchronization dynamics as a potential intrinsic excitability measures (IEM) in cortex in humans. Specifically, Meisel et al. (2015) hypothesized that the mean levels of phase synchronization should positively correlate with cortical excitability. To study the effect of low-frequency electric cortical stimulation on the phase synchronization dynamics of epileptic focus in humans, the following methodology was applied.

Materials and Methods

Data analyzed in this study was obtained from three patients who were undergoing presurgical evaluation and were implanted with depth electrodes. The stimulation was administered to the epileptic focus by means of stereotactically implanted depth electrodes. We used biphasic electric current of 250 μs/phase duration that was delivered to two adjacent electrodes with a stimulation frequency of 1 Hz. Each stimulation block consisted of 600 stimulations (10 minutes) with 10 minutes pause intervals. The stimulation was repeated 3 times with stimulus intensities of 1.5 (patient 1), 2 mA (patient 2) and 3 mA (patient 3). For offline analysis, both global and local effects of stimulation were studied. For global analysis, the two stimulation electrodes and other electrodes showing signs of stimulation artifacts were excluded. For the local analysis, two neighboring channels to the stimulation channels which were less affected by the artifacts were selected. To remove slow drifts and high frequency noise, phase neutral filter was performed by applying a second-order Butterworth filter in both directions. Then the electrode montage was converted to common average. The power line noise at 50 Hz was removed by applying a notch-filter on the filtered data to exclude this frequency. The channel selection for global analysis was performed by comparing the standard deviation (SD) in the time window during which evoked responses were analyzed (10-110ms) and the remaining inter-stimulus interval (110-990ms) as baseline. Due to possible artifacts of stimulation, 10 ms before and after each stimulation was removed from analysis. For each patient, channels with an evoked potential standard deviation beyond 3 times the standard deviation of the baseline period were selected for an analysis of excitability and synchronization.

Excitability Measurement

Stimulation-evoked potentials in each selected channel were derived by averaging the responses in each stimulation block (600 stimulations per block) time-locked to the onset of stimulation. To this end, first a band-pass filter (sixth-order Butterworth filter; 1-100 Hz due to sometimes-appearing slow current transients) in both time directions was applied to avoid potential stimulation artifacts or ringing. Then, for each stimulation block, epochs were averaged to get the mean amplitude. For excitability measurement, two different methods were applied. First the peak to peak (P-P)

amplitude in each channel was taken as a measure of excitability. Second, variation of the standard deviation (between 10 and 110 ms after the stimulation onset) between the last stimulation and the first stimulation block was used as a more robust estimator for the size of the response.

Synchrony Measurement

In the selected data segments, the seconds contaminated with artifact were removed from the analysis. In the next step, by using a sixth-order Butterworth filter, IEM in the frequency bands 1-50 Hz, 50-100 Hz and 100-200 Hz were extracted. Then we derived estimates of mean phase synchronization for band-pass-filtered data in time windows of one second. To this end, first a phase trace from each channel by applying Hilbert transform was obtained. Having the phase trace the Kuramoto order parameter over the selected channels was calculated. Finally, by averaging the kuramoto order over the length of the data segment we can quantify the mean synchrony.

Results

A 10 min data segment (started ~20 min prior to the first stimulation) was selected as the baseline. To analyze the effects of the stimulation on the synchrony levels, intracranial Electroencephalography (iEEG) data in a segment of 10-min duration after the last stimulation block was compared to the baseline. To analyze the effects of stimulation on the excitability, the excitability values were measured for each stimulation block and the excitability levels of the last stimulation block was compared to the first stimulation blocks. The results of global analysis and local analysis are presented in tables1 and 2; respectively.

Table 1. Percentage of excitability and mean phase synchrony changes after low frequency electrical stimulation on three patients – global changes

#	Excitability changes (%)		Mean phase synchrony changes (%) in 1-50 Hz	Mean phase synchrony changes (%) in 50-100 Hz	Mean phase synchrony changes (%) in 100-200 Hz
	SD	P-P			
P1	-11.31	-8.50	6.53	0	11.67
P2	4.04	3.10	-33.16	-9.82	-1.94
P3	34.17	33.22	7.00	3.21	0.39

Table 2. Percentage of mean phase synchrony changes after low frequency electrical stimulation on three patients – local changes

#	channel	Excitability changes (%)		Mean phase synchrony changes (%) in 1-50 Hz	Mean phase synchrony changes (%) in 50-100 Hz	Mean phase synchrony changes (%) in 100-200 Hz
		SD	P-P			
P1	HAR4	-7.46	-7.99	10.10	0.95	7.75
	HAR5					
P2	EL4	8.40	7.66	6.03	3.52	-0.19
	EL5					
P3	FIR2	25.85	31.83	-22.39	-2.48	4.98
	FIR3					

Discussion

The results of this study have several implications. First, synchrony changes are higher in the 1-50 Hz frequency range in local analysis. Second, the synchrony changes do not necessarily follow excitability variations. Third, the effects of the low frequency stimulation on excitability can vary between patients. It is important to notice that, excitability changes in the local and global level, covariate with each other; however, synchrony changes in the local level do not covariate with global changes. These results suggest that the cortical excitability can be modulated due to low frequency electrical stimulation. Additional patients are required to systematically assess the seen effects, taking also anatomical regions of stimulation, the underlying etiology and disease progress into account. Comprehensive results may lead to improved treatment options for patients with medically refractory patients.

Using time series analysis for challenging breast lesion detection and classification in DCE-MRI

I.A. Illán[1], A. Bagnall[2], J. Ramirez[1], J.M. Gorriz[1], K. Pinker[3,4], M.A. Marino[3], D. Avendaño[3], A. Meyer-Baese[5]

[1]Department of Signal Theory, Networking and Communications, University of Granada, Spain
[2]Computer Science Sector, School of Information Systems, University of East Anglia, Norwich NR4 7TJ, UK
[3]Department of Radiology, Breast Imaging Service, Memorial Sloan Kettering Cancer Center, New York, USA
[4]Department of Biomedical Imaging and Image-Guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna, Vienna, Austria
[5]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA

Abstract

Non-mass enhancing (NME) lesion detection in magnetic resonance imaging (MRI) of the breast represents a significant challenge. NME lesions exhibit a heterogeneous appearance with high variations in kinetic characteristics and typical morphological parameters, and have a specificity and sensitivity much lower than mass-enhancing lesions. With the study of tissue kinetics with computer aided diagnosis (CAD) systems, new insight can be gained into lesion characteristics at the voxel level. However, the processing of MRI images for lesion detection and segmentation is a big data problem, and data-mining techniques can usefully be applied. We propose to analyze the kinetic curves of DCE-MRI images using an ensemble approach: the Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE), which has been proven to represent a new state of the art for time series classification (TSC). HIVE-COTE is a modular hierarchical structure with a probabilistic voting scheme that includes data transforms in the time, shape and frequency domains. This means inter-patient differences can be accounted for, as well as intra-patient differences in time-to-peak enhancement, wash-in wash-out, or permanent enhancing differences. Different algorithms in the ensemble are evaluated and ranked, and most discriminative features are extracted. We test the performance of the proposed algorithms for NME lesion classification in a multiparametric DCE-MRI database with high temporal resolution.

Relationships between stock exchanges of Shanghai, Shenzhen and Hong Kong considering the split-share reform

Yang MESTRE-ZHOU[1] Roman MESTRE[2]

Abstract.

China wants to improve her financial power and attracts foreign investor, so the Chinese authorities establish the Non-tradable Share (NTS) Reform at the end of 2005. However, the Chinese financial system is composed by three major stock markets, Shanghai (SSE), Shenzhen (SZSE) and Hong Kong (HKSE). In this paper, we analyse the relationships between these markets considering the NTS reform from 1993 to 2017(daily data). By using the causality test and co-integration analysis, we find that the NTS reform modifies the linkages between three markets. SSE becomes the leader of the Chinese markets.

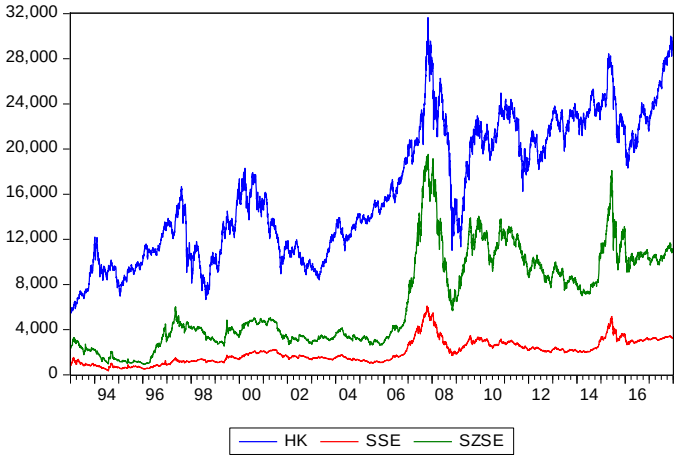
Keywords: Non-tradable Share (NTS) Reform, Chinese markets, Granger causality, co-integration

[1] Corresponding Author: yangzhou26@yahoo.com, University of Montpellier-MRE (Montpellier research in economics)
[2] University of Montpellier-MRE (Montpellier research in economics)

1 Introduction

In an environment of economic globalization, Chinese stock markets become more and more open and the Shanghai Stock-Exchange is one of the biggest financial places in the world. But there are three different Chinese markets: Shenzhen (SZSE), Shanghai (SSE) and the Honk Kong Stock Exchange (HKSE). The relationships between these three markets are influenced by the Chinese policies following the different development phases. The two continental markets were created during the beginning of 90s when Hong Kong was an English possession; it came back to China in 1997 with a specific economic status. HKSE is totally open to the foreign investors whereas SSE and SZSE are still relatively closed. As the correlation between them are getting closer day by day, analyzing the linkages between them can help investors make better investment plans, or seize the opportunity to get high returns with relatively low risks.

Figure 1: Chinese markets indices: 01/01/1993-12/29/2017



By looking at the stock price movement among Chinese markets in Figure 1, a great volatility deserves our attention in early 2006. During the past decade, the mainland markets have gradually opened up to adapt the financial liberalization environment. Since 2002, China has started to implement a program "Qualified Foreign Institutional Investors"(QFII), which allow some foreign institutional investors to buy "A-share" denominated in RMB in the two China's mainland stock exchanges. Furthermore, in order to make the Chinese mainland markets healthier with stable development and attract more foreign investors, the Chinese government has established the Split-Share Reform or Non-tradable Share (NTS) Reform at the end of 2005. The Split-Share refers to that the shares of listed companies in the A-share market are divided into tradable shares and non-tradable shares. This situation is specific to the mainland China stock market. In order to eliminate the differences in the circulation system, the

China Securities Regulatory Commission (CSRC) launch the reform for NTS to achieve an objective: for a same-share there are the same right and the same profit. "NTS reform allowed shareholder to profit immediately."(Yeung H.W.H.,2009). It's a milestone in the process of the modernization of Chinese market, it can better understand the overall institutional mechanisms that promote China's economic growth.(Yeh Y.H.,et al 2009). We suggest this changing can affect the links between its stock markets.

The remainder of this study is organized as follow. Section 2 provides a brief review literature. Section 3 presents the statistical data and methodology used in the paper. Section 4 reports the empirical finding on the overall period. Section 5 presents the results of NTS reform effect on markets linkage, and Section 6 concludes.

2 Literature Review

In the past decade, Chinese markets attract more and more investors thanks the Chinese economic performances. Chinese markets become interesting subject of studies in many fields as volatility analysis, corporate finance and world integration. Many papers focus on China market links with other international financial markets. Wong, Penm, Terrell and Lim (2004) analyze the Co-Movements of Emerging and Developed countries, Gupta and Guidi (2012) study cointegration analysis between Asian countries, Chiang and Chen (2016) study the financial power of China in Asia with DCC-GARCH model. These papers consider Hong-Kong market to represent China. But there are other financial places in China, Shanghai and Shenzhen, with different rules, organizations, structures and opportunities.

Considering this case there are few papers focused on interactions and relationships between the three Chinese markets. Lu and Zhu (2001) find that HKSE is not linked with mainland markets. Wang and Xu (2003) confirm partially this result, the volatility of HKSE is endogenous and the influence of Chinese markets is not significant (but they suppose that it could increase in the future). Hu and Lv (2008) indicate that HKSE and SSE "are isolated and separated". However, other studies highlight strong linkages between Chinese stock markets. Shi (2002) and Li (2009) show that SSE and SZSE are co-integrated and conclude that there is a long-run relationship. Wang and Yu (2004) analyze the impact of Asian Crisis in 1997 and they find that after crisis linkages are tighter with mutual influences and before SSE and SZSE influence HKSE. Guo and Wang (2009) find similar results. Wong and Chen (2011) find that there is no relationship at long-run but at short-run linkages are stronger.

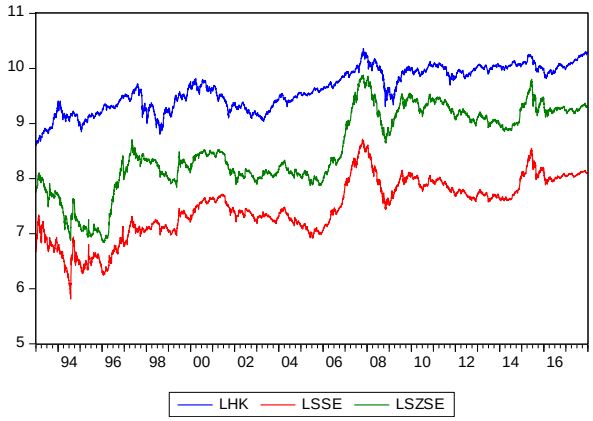
The previous studies are interested but they are focused on the Hong-Kong reunification to China in 1997 and the Asian Crisis of 90's. They don't consider the important split-share reform in the end of 2005. Few papers analyze the effect of the reform on

Chinese markets. Yeung (2009) conclude that the reform will improve the equity liquidity and the market's tability in the long term,. Guo, Lien and Dai (2016) find the reform ameliorate the corporate finance structure. Beltratti, Bortolotti and Caccavaio (2016) find that there is an announcement effect generating positive abnormal returns. The previous studies are based on stocks performances before and after the reform but not on the effect on the markets linkages. It is that why we want to merge the two problematics: the impact of the reform on the market linkages in short and long term.

3 Data and Methodology

In this paper, we use the three Chinese markets indices, the SSE Composite Index (called SSE), the SZSE Component Index (called SZSE) and the Hang Seng Index (called HKSE) for the daily period from January 1, 1993 to December 29, 2017. The data are obtained from Yahoo Finance database. All the indices are analyzed by the logarithm forms in order to smooth and reduce the amplitude effect, note LSSE, LSZSE and LHK.

Figure 2: Indices in log-form



To examine the short-run lead-lag interactions and the long-run co-movements between the three Chinese markets, we use the Granger Causality Test (1969) and Co-integration analysis (1987-1995).

Firstly, before realizing the two analyses we should test the stationarity by Unit Root Test for all the previous logged variables. The Unit Root test is useful to find the integration order of each time series and characterize the trend, and then it is a decisive step of to the Co-Integration Test because it influences all results. In our studies, we

use the Augmented Dickey-Fuller Test (1979-1981) and the Phillips-Perron Test (1988). The choice of these tests is justified by their robustness to the autocorrelation (for ADF test) and heteroskedasticity (for PP Test).

The ADF Test is based on three AR (p) models to determinate the presence of a unit root and characterize the nature of the trend (Stochastic or Determinist Trend with or without constant). The AR order p is find by minimizing the Akaike Information Criterion (AIC) or the Schwarz Bayesian Criterion (SIC or SBC). The AR model treats the autocorrelation issue in errors. However, if we find p=0, we realize the simple Dickey-Fuller (DF) test.

$$\boldsymbol{M1}: \boldsymbol{No\ Trend - No\ Constant}$$
$$\Delta x_t = \rho x_{t-1} - \sum_{j=2}^{p} \phi_j \Delta x_{t-1-j} + \varepsilon_t$$
$$\boldsymbol{M2}: \boldsymbol{No\ Trend - Constant}$$
$$\Delta x_t = \rho x_{t-1} + c - \sum_{j=2}^{p} \phi_j \Delta x_{t-1-j} + \varepsilon_t \qquad (1)$$
$$\boldsymbol{M3}: \boldsymbol{Trend - Constant}$$
$$\Delta x_t = \rho x_{t-1} + bt + c - \sum_{j=2}^{p} \phi_j \Delta x_{t-1-j} + \varepsilon_t$$

Under the null hypothesis the series has a unit root so it is not stationary. But the ADF test doesn't consider the heteroskedasticity problems in errors terms. It is that why, we also the Phillips-Perron Test. PP Test is based on the equation of ADF test without the AR(p) terms but it deals differently with errors issues.

In the previous two tests we use the Cem Ertur strategy[3]. We begin the strategy by testing the unit root in the model 3 framework and also the nullity of c and b parameter (constant and trend) with joint tests. According to the results, we test the unit root in model 2 or we follow the step by Student test on c and b.

According to the Unit Root Test we realize the causality or Co-integration analysis. In our paper, we consider that non-stationary I(1) series can be stationarized thanks the first differences filter. In this case, we have the returns of the indices.

To analyze the short-run relationships, we apply firstly the causality test.

[3] More details on this procedure are available in Terraza and Bourbonnais book (see references).

Supposing two time series x_t and y_t, according to the Granger definition (1969), $y_t$cause x_t if we can predict x_t with the values of y_{t-p} (the past). Granger consider also the reverse relationship, x_t cause y_t, in this case we have bi causality effect or feedback relationship. The Granger Causality test is based on VAR(p) model on stationary variables (or stationarized).

$$\begin{aligned} \Delta x_t &= \alpha_1 + \sum_{i=1}^{p} \beta_{1,i} \Delta x_{t-i} + \sum_{i=1}^{p} \gamma_{1,i} \Delta y_{t-i} + \varepsilon_{1,t} \\ \Delta y_t &= \alpha_2 + \sum_{i=1}^{p} \beta_{2,i} \Delta x_{t-i} + \sum_{i=1}^{p} \gamma_{2,i} \Delta y_{t-i} + \varepsilon_{2,t} \end{aligned} \quad (2)$$

We test the significance of the parameter by nullity test and the order p by minimizing an Information Criterion (AIC or SBIC). By calculation the F-statistic for the two variables, we can determinate the direction of the causality.

Then, to analyze the long-run links we perform a Co-Integration analysis. Engel and Granger (1987) indicate that a long-run relation between two or more non-stationary time series exist if a combination of these chronic is stationary. Engel and Granger method is based on ADF test and on I(1) variables. In this case we estimate a static OLS equation between variables and if the residuals are stationary we can conclude that there is a co-integration relationship.

But in this paper, we use the Johansen Test (with the Trace test and maximum eigenvalue Test, 1988-1995) to determinate the number of co-integration relation between series. If we call X the vector of k non-stationary variables, the method is based on VAR(p) model on VECM form:

$$\mathbf{VAR(p)}$$
$$X_t = A_1 X_{t-1} + \cdots + A_p X_{t-p} + \varepsilon_t$$

$$\boldsymbol{VECM} \quad (3)$$
$$\Delta X_t = \Pi . X_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta X_{t-i} + \varepsilon_t$$

$$Where\ \Pi = \left(A_1 + \cdots + A_p\right) - I\ and\ \phi_i = -(A_{i+1} + \cdots + A_p)$$

Under these equations, if the rank of Π equal to zero, there is no Co-integration relationships and if $rank(\Pi) = k$ so there are k Co-integration relations. The Trace Test and Max Eigenvalues Test determine the number r by testing condition on Π.

4 Empirical Finding on the overall period:

4.1 Unit root test

In this section, to test the stationarity of the three variables, we realize the Augmented Dickey-Fuller test (robust to the autocorrelation) and Philips-Perron (robust to the heteroskedasticity). According to the test statistics in Table 1, we can conclude that SSE and SZSE have a unit root at the 1%, 5% and 10% risk level and HKSE has a unit root at the 1% risk level. So we consider that the three series are DS processes (Difference Stationarity) without derivate. They are stationary after filtering by the fist differences, noted ΔLHK, ΔLSSE and ΔLSZSE. We have then the returns of the index, noted DLHK, DLSSE and DLSZSE. Consequently, the three series are order 1 integrated, noted I(1).

Table 1: Unit Root Test Results

1993-2017	ADF INTERCEPT AND TREND	ADF INTERCEPT	PP INTERCEPT AND TREND	PP INTERCEPT
LHK	-3.769(0)*	-2.256(0)***	-3.787(4)*	-2.264(3)***
ΔLHK	-80.47(0)	-80.473(0)	-80.470(4)	-80.473(4)
LSSE	-2.794(6)***	-1.556(6)***	-2.782(7)***	-1.735(5)***
ΔLSSE	-32.874(5)	-32.877(5)	-81.425(3)	-81.431(3)
LSZSE	-2.008(0)***	-1.124(0)***	-2.205(13)***	-1.223(12)***
ΔLSZSE	-36.788(3)	-36.791(3)	-78.813(10)	-78.818(10)

Notes: Values in the parentheses are the Lag Length (ADF) or the Bandwidth (PP) used in the estimation of the unit root test. *denote accept null hypothesis at the 1% risk level, ** 5% risk level, ***10% risk level.

4.2 Granger Causality

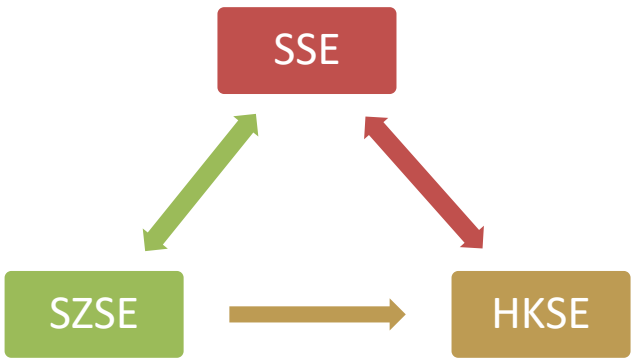
In the section, we present the results for the Granger causality test with the three stationary series. Table 2 shows that there are a bi-causality or feedback effect between SSE and SZSE and between SSE and HKSE. The relationship between SZSE and HKSE is not clearly explicit, SZSE influences HKSE but the feedback effect (of HKSE on SZSE) is not totally significant. Figure 3 summarizes these results.

Table 2: Granger Causality Test Results

1993-2017	Null Hypothesis	
Lags	DLSSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSSE
1	16.6617***	4.91134**
2	8.42871***	3.80355**
3	5.65007***	3.49105**
4	4.15061***	3.92196***
5	3.48354***	3.95562***
6	3.85098***	4.31938***
Lags	DLSZSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSZSE
1	16.3241***	2.61425
2	8.25906***	1.68973
3	5.51263***	1.69372
4	4.07722***	2.17590*
5	3.26633***	2.67300**
6	3.21570***	2.22889**
Lags	DLSZSE does not Granger Cause DLSSE	DLSSE does not Granger Cause DLSZSE
1	28.8855***	8.84715***
2	14.5196***	4.56582**
3	9.86937***	3.12900**
4	7.59595***	3.58632***
5	6.18879***	3.96737***
6	5.93504***	3.42903***

Note: Null hypothesis is not accepted at the ***1%, **5% and *10% risk level.

Figure 3: Causality results (Pairwise Granger Test):



4.3 Co-integration

In this section, we present the results of the Co-Integration Analysis by the Johansen Co-integration Test. Previously, we find that all the series are I(1) (without constant), so we build the test on this structure and on a VAR(1) model selected by minimized Schwarz Criterion (SC).

Table 3: Co-integration Analysis Results: 1993-2017

Hypothesized No. of CE(s	None	At most 1	At most 2
LSSE, LSZSE and LHK			
Trace Statistic	31.81	11.5	1.61
Max-Eigen Statistic	20.31	9.89	1.61
Critical Value(0.05)-Trace	29.8	15.49	3.84
Critical Value(0.05)-Max	21.13	14.26	3.84

The Trace test (Table 3) shows that there is a cointegration relationship between the three markets with a risk level of 5%. This confirms what we can observe graphically on Figure 1 and 2 where we notice a common trend between SSE, SZSE and HKSE. In addition, from a graphical point of view, the trend was interrupted around the end of 2005 - early 2006. We clearly identify two volatility regimes after and before this break especially for Shanghai and Shenzhen: Low Volatility from 1993 to 2005 and High Volatility from 2006 to 2017. This break can affect the robustness of the co-integration test, then, we need to analyze more precisely it.

5 Non-Tradable Share Reform effects on markets linkages

The rupture observed previously, coincide with the Non-Tradable Share reform established at the end of 2005. The previous results don‘t consider this important change in the mainland markets organizations, so, we can suppose that the relationships are not totally appreciate.

To improve the analysis and include the effect of the reform, we divide the sample in two periods:

- First period from January 1, 1993 to December 30, 2005 corresponding to the ante-NTS Reform phase.
- Second period from January 2, 2006 to December 29, 2017 corresponding to the post-NTS Reform phase.

In order to confirm our hypothesis on this break, we carry out the Chow test (1960) of comparison of the regression coefficients (Table 4). The Chow test confirms the presence of rupture (statistics are superior to Fisher tabulated).

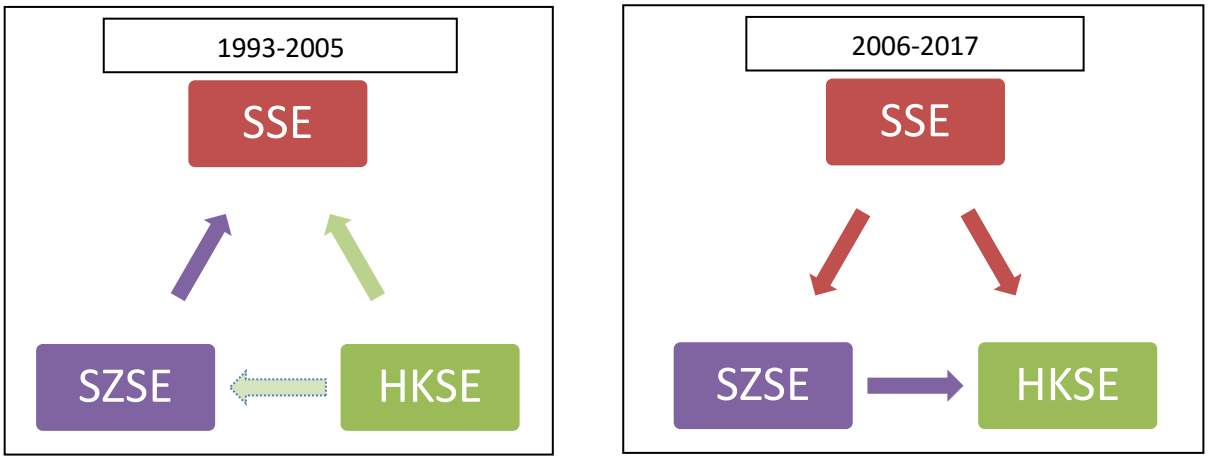
Table 4: Chow Test Results

Variable explained	LHK	LSSE	LSZSE
Chow Test stat	561.25	112.68	322.21

We reiterate the stationary tests on these two sub-periods (Appendix 1) and we conclude that all series are DS processes without constant for the first period. For the second period HKSE and SSE have the same structure but SZSE is a DS process with constant. At the end, all series are I(1) for the two periods.

We follow the same methodology as previous. Figures 4 summarize results of Granger Causality Tests for the two sub periods (Appendix 2):

Figures 4: Causality results (Pairwise Granger Test):



We notice that SSE linkages with other markets are reversed after the reform. SSE was both influenced by SZSE and HKSE during the first period whereas it influences SZSE and HKSE after the reform. It is similar for HKSE, during the first period it influences SSE and SZSE, but after, it is influenced by others. Similar conclusions can be made for SZSE. Compare to Figure 3, the bi-causality relationship, see for the global period, is in fact decomposed across time. This result highlights the effect of the reform on the linkages between Chinese markets.

Moreover, the co-integration tests indicate that there are no co-integration relations for the first period, but there are two cointegrations relationships for the second period (Table 5). This indicates that the cointegration overall period is time-varying: there is a long-term relationship between the two mainland stock markets and HKSE since the NTS reform.

Table 5: Co-integration Analysis Results by Periods

5.1: First Period: 1993-2005

Hypothesized No. of CE(s)	None	At most 1	At most 2
LSSE, LSZSE and LHK			
Trace Statistic	27.15	12.38	3.45
Max-Eigen Statistic	14.77	8.93	3.45
Critical Value(0.05)-Trace	29.8	15.49	3.84
Critical Value(0.05)-Max	21.13	14.26	3.84

5.2: Second Period: 2006-2017

Hypothesized No. of CE(s)	None	At most 1	At most 2
LSSE, LSZSE and LHK			
Trace Statistic	33.62	15.5	2.52
Max-Eigen Statistic	18.13	12.98	2.52
Critical Value(0.05)-Trace	29.8	15.49	3.84
Critical Value(0.05)-Max	21.13	14.26	3.84

6 Conclusion

We analyze the relationships between three Chinese markets from 1993 to 2017; we find bi-causality linkages between SSE-SZSE, and SSE-HKSE, but not long-run co-movements. By diving in two our sample, we consider the NTS reform break at the end of 2005. Results indicate that the feed-back relationships are time varying: on the first period, HKSE and SZSE mutually Granger cause SSE, but on the second period, SSE influences both SZSE and HKSE. Furthermore, after the reform, the two mainland stock markets establish a long term relationship with HKSE.

So we can conclude as follow. First, before the reform, as a free financial market, HKSE influences the two mainland markets in the short term, but after the reform, the results were reversed. The reform ameliorates the Chinese market structure and changes the rank and financial power of each market, especially SSE becoming the most important stock market in China. Second, before the reform, SSE and SZSE have relatively closed cause of the non-tradable share structure preventing the development of the mainland markets. After the reform, the mainland stock markets SSE and SZSE have achieved internationalized because we find a long-run equilibrium relationship between them and HKSE.

To conclude, we show that the NTS reform modify/reveres the linkages (short-run and long-run) between the three Chinese markets. If we globally analyze the relationships, we can't appreciate this changing.

7 Appendixes

Appendix 1: Unit Root Test Results by Sub-Periods

1993-2005	ADF INTERCEPT AND TREND	ADF INTERCEPT	PP INTERCEPT AND TREND	PP-INTERCEPT
LHK	-3.129(3)**	-3.009(3)*	-3.043***	-2.965*
ΔLHK	-31.289(2)	-31.283(2)	-56.508	-56.507
LSSE	-2.097(0)***	-1.959(0)***	-2.075***	-1.954***
ΔLSSE	-59.302(0)	-59.307(0)	-59.309	-59.313
LSZSE	-1.457(0)***	-1.409(0)***	-1.585***	-1.507***
ΔLSZSE	-59.318(0)	-57.326(0)	-57.357	-57.364
2006-2017	ADF INTERCEPT AND TREND	ADF INTERCEPT	PP INTERCEPT AND TREND	PP INTERCEPT
LHK	-3.03(0)***	-2.633(0)**	-2.911***	-2.518***
ΔLHK	-57.615(0)	-57.625(0)	-57.682	-57.692
LSSE	-2.653(0)***	-2.841(0)**	-2.695***	-2.868*
ΔLSSE	-55.218(0)	-55.202(0)	-55.314	-55.312
LSZSE	-3.201(0)**	-3.565(0)	-3.21**	-3.533
ΔLSZSE	-53.606(0)	-53.562(0)	-53.77	-53.749

Notes: Values in the parentheses are the Lag Length (ADF) or the Bandwidth (PP) used in the estimation of the unit root test. *denote accept null hypothesis at the 1% risk level, ** 5% risk level, ***10% risk level.

Appendix 2: Granger Causality Test Results by Sub-Periods

A2.1: First Period from 1993 to 2005:

1993-2005	Null Hypothesis	
Lags	DLSSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSSE
1	2.71066*	1.78423
2	1.57557	1.94506
3	1.14622	2.8999**
4	0.9076	2.6359**
5	0.78047	3.17236***
6	1.21385	2.77156**
Lags	DLSZSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSZSE
1	2.99101*	1.20482
2	1.83088	0.66565
3	1.31418	1.89888
4	1.19704	1.95909*
5	0.96634	3.09744***
6	1.07353	2.55412**
Lags	DLSZSE does not Granger Cause DLSSE	DLSSE does not Granger Cause DLSZSE
1	20.4229***	4.50260**
2	10.2155***	2.4244*
3	6.6766***	1.6992
4	5.12789***	1.61107
5	4.72686***	1.62736
6	4.32814***	1.36996

Note: Null hypothesis is not accepted at the ***1%, **5% and *10% critical value.

A2.2: Second Period from 2006 to 2017

2006-2017	Null Hypothesis	
Lags	DLSSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSSE
1	20.6503***	2.47728
2	10.4601***	1.08884
3	8.01568***	0.72851
4	6.007***	1.13196
5	5.11573***	0.95634
6	4.54068***	1.84219*
Lags	DLSZSE does not Granger Cause DLHK	DLHK does not Granger Cause DLSZSE
1	14.5989***	0.83947
2	7.42745***	0.54646
3	6.12369***	0.33832
4	4.79332***	0.59645
5	3.89163***	0.44992
Lags	DLSZSE does not Granger Cause DLSSE	DLSSE does not Granger Cause DLSZSE
1	3.18683*	5.93907**
2	1.72101	3.52179**
3	2.46539*	3.06367**
4	2.22016*	4.69871***
5	2.30647**	4.84741***
6	2.17796**	4.45648***

Note: Null hypothesis is not accepted at the ***1%, **5% and *10% critical value.

8 References

Beltratti A., Bortolotti B. and Caccavaio “M., Stock market efficiency in China: Evidence from the split-share reform”, The Quarterly Review of Economics and Finance, vol.60, 2016, p.125-137

Bourbonnais R and Terraza M., Analyse des Séries Temporelles Application à l’Economie et à la Gestion , ed. DUNOD

Chiang T.C. and Chen X.Y., “Empirical Analysis of Dynamic Linkages between China and International Stock Markets”, Journal of Mathematical Finance, vol.6,2016,p.189-212

Chow G C., “Tests of Equality Between Sets of Coefficients in Two Linear Regressions”, Econometrica, vol. 28(3), 1960, p. 591-605

Dickey D.A and Fuller W.A., “Distribution of the estimators for autoregressive time series with unit-root”, Journal of the American Statistical Association, vol.74, 1979, p.427-431.

Dickey D.A and Fuller W.A., “Likelihood ratio statistics for autoregressive time series with a unit root”. Econometrica, 1981, vol.49, p.1057-1072

Engel R.F. and Granger C.W.J., “Co-Integration and error correction: Representation, Estimation and Testing”, Econometrica, vol. 55(2), 1987, p. 251-276.

Granger C.W.J., “Investigating causal relations by econometrics models and cross-spectral methods”, Econometrica, vol. 37(3), 1969, p. 424-438.

Gregory C., Chow: “Tests of Equality Between Sets of Coefficients in Two Linear Regression”, Econometrica, vol.28 (3), 1960, P.591-605

Gupta R. and Guidi F. “,Cointegration relationship and time varying co-movements among Indian and Asian developed stock markets”, International Review of Financial Analysis,vol.21,2012,p.10-22.

Guo L .and Lien, D.D. and Dai Y., “The Effects of China's Split-Share Reform on Firms’ Capital Structure Choice” , Applied Economics, Vol. 48, 2016 - Issue 27.

Guo X. and Wang X., “An empirical study on China inland and Hong-Kong security markets’ integration”, Modern Business Trade Industry, Vol.5 , 2009, p.118-120.

Hu J. and Lv P., “Analysis on the reaction between Shanghai stock market and Hong-Kong stock market”, Journal of Shanxi Finance and Economics University, vol.30 (12), 2008, p.93-98.

Johansen, S., “Statistical Analysis of Cointegration Vectors”, Journal of Economic Dynamics and Control, vol.12, 1988, p.231-254.

Johansen, S. and Juselius, K.., “Maximum Likelihood Estimation and Inference on Cointegration - with Applications to the Demand for Money", Oxford Bulletin of Economics and Statistics vol.52, 1990, p.169-210.

Johansen, S. “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models”, Econometrica,vol.59,1991,p. 1551-1580

Johansen, S., “Likelihood-Based Inference in Cointegrated Vector Autoregressive Models”. Oxford University Press, New York, 1995

Li S., ”Granger Causality between Shanghai and Shenzhen stock indices”, Security Times, vol.25 (A15), 2009.

Lu Z. and Zhu H., “Granger Causality between China stock markets”, Journal of Management Sciences in China, vol.4(5), 2001, p.7-12.

Phillips P.C.B, and Perron P.,”Testing for unit-root in time series regressions”, Biometrika vol.75, 1988, p.335-346.

Shi D., “Cointegration research on Shanghai and Shenzhen stock exchanges’ indices’ fluctuation”, Quantitative and Technical Economics, vol.19(9), 2002, p.103-105.

Wang H. and Yu B., “The integration trend: the China mainland stock markets and Hong-Kong stock market 1993-2002”, School of Economics Nankai University Working Paper 2004.

Wang H. and Xu X., “Hong-Kong and inland stock markets’ fluctuating transmition: empirical research on pre and post rejoining”, Journal of Economics Studies (by CCER) ,vol. 15(1), 2003, p.49-56.

Wong H.T and Chen Z., “Integration analysis of the People Republic of China stock markets”, Labuan Bulletin of International Business and Finance, vol. 9, 2011, p.26-43.

Wong W.K.,Penm J.,Terrell R.D. and Lim K.Y.C., “The Relationship between Stock Markets Of Major Developed Countries And Asian Emerging Markets”, Journal of Applied Mathematics and Decision Sciences,vol.8(4),2004, p.201-218.

Yeh Y.H.,Shu P.G.,Lee T.S.,Su Y.H.,"Non-Tradable Share Reform and Corporate Governance in the Chinese Stock Market", Corporate Governance: An International Review,vol.17(4),2009,p.457-475

Yeung.H.W.H., "Non-Tradable Share Reform in China: Marching Towards the Berle and Means Corporation?" . Comparative Research in Law & Political Economy, 2009, Research Paper No. 48.

Economic and Environmental Benefits Based on Scenario Analysis in Transportation Sector: A Case Study of Kuwait

S. Al-Osaimi and K. J. Sreekanth

Energy and Building Research Center, Kuwait Institute for Scientific Research, Kuwait

(sosaimi, sreekanthkj) @kisr.edu.kw

Abstract

Transport represents the largest oil consumer sector in the world and hence one of the main challenges for climate change and energy security of supply policies of any country. Improving energy efficiency in the transport sector is the need of the hour. This study explores a scenario-based analysis of future energy consumption and CO_2 emissions for Kuwait between 2016 and 2027 using the time series forecasting. The impact of different energy scenarios are analyzed for the Kuwait transportation system by considering two major scenarios: The reference (REF) scenario, the low-carbon scenario (70/30). Data relating to transport Energy usage in the current scenario has been collected and employed to obtain the potential Energy demand in the Transport sector. A time series model was developed in this study using MS-Excel platform to model the energy consumption and CO_2 emission trends for the state of Kuwait for 2016–2027. The reference or business as usual (BAU) scenario is based on the energy consumption characteristic of transportation sector using the trends related to socio-economic prospects and the number of vehicles. The alternative scenario took into account energy efficiency and emission reduction, if 30% of the present private vehicles are replaced with public transport systems. The results show that fuel consumption in the transport sector will increase by 40.34% between 2016 and 2027 in the BAU scenario. Maximum reduction in CO_2 emissions in the transport sector of Kuwait was observed by 2027 in alternative scenario of 70/30, with an associated high financial savings.

Key words: Fuel economy, energy consumption, and transportation sector.

1 Introduction

Transport is usually called the ‘engine’ of financial system of the country, albeit ‘lubricant’ would seem a more suitable automotive symbol. Reasonable modes of freight/passenger transport permit other economic sectors to optimize the countless steps in the value chain from raw materials to final products and thus strengthen the economy of the country (CCAP, 2008). Besides the afore-mentioned economic as-

adfa, p. 1, 2011.

pect, the transport sector is extremely dependent on the price and accessibility of fuel. Contemporary years have shown that the oil price can rise to unanticipated heights due to geopolitical instabilities, natural disasters and technical hindrances (Fan and Lei, 2016). Even if worldwide oil resources are still significant, they are much more inadequate than the resources of coal and gas. Hence, energy security is also a vital component that needs enhanced consideration while discussing the importance of transport sector (Khalili et al, 2015).

2 Background

With its liberal market-based economic policy and openness to the world of international investment, the State of Kuwait is striving to become a "hub for foreign and regional investments" by 2030. Most of the people travel by private vehicles (85%). Public transport bus services are provided by a private company (City Bus) and state-owned Kuwait Public Transportation Corporation. Given Kuwait's rapid economic growth, low price of old vehicles, and easy access to loans to finance new vehicle purchases, the vehicle fleet is expected to continue to grow in the coming years, and this will not be matched by similar growth in the capacity of road networks and transport facilities. There is an increasing pressure therefore to find energy efficient feasible and optimal solutions for the escalating problems of the transport sector. Transport policy planning is lacking through which roads, ports and airports, land use, and urban features can be integrated with community needs. The transport sector's management practices are characterized by unavailability of an integrated strategy (AboulNaga et al.,1997). As mentioned above, the objective of this study is evaluate the effectiveness of fuel efficiency enhancement and emission reduction scenarios, both economically and environmentally, for future system development pertaining to road transport sector in Kuwait using time series forecasting.

3 Approach and Methodology

Energy demand and emission reduction analysis and evaluation for road transport sector and corresponding scenario analysis to analyze the impacts of identified energy saving measures, are performed using time series forecasting method, to calculate the

prediction of energy demand in both the cases. The energy demand in all the cases is calculated with the bottom-up method. The road transport sector energy demand is calculated based on this basic equation (1):

$$ED_{i,j} = NV_{i,j} \text{ X } FE_{i,j} \text{ X } VKT_i \quad (1)$$

Where, NV is number of vehicles,

FE is the fuel economy,

VKT kilometer travelled by the vehicles,

i is the vehicle type and

j is the type of fuel used.

The FE and VKT are estimated by the survey conducted, while the NV models are obtained from national records and ministries such as ministry of information (MOI), considering the logistics and socio-economic factors. The number of private vehicles (NV) data were collected and validated with historical record of vehicle registration from the MOI and the quantity of fuel used for transportation too was collected. In this study, the fuel consumption in terms of demand is projected from current situation to 2027. It is expected that averaged fuel economy (FE) will be improved considerably for private car segment if the energy efficient transportation measures are followed. As a result, the measured scenario is defined as follows:

1. Business as usual or baseline scenario: The potential contribution of electric vehicle and other automotive technology will not change from current situation in all vehicles.
2. Partial efforts scenario (Scenario II): Here minimal efforts need to be undertaken by the authorities to achieve a modal Split of 70:30 between public and private mode in road transport.

4 Data Collection and Analysis

The survey questionnaire used was prepared by the team members and circulated among the attendees through email. The questionnaire for the survey has 17 questions, which are of objective type. The consent to attend the survey as well as to participate in the data collection by giving the vehicles was collected from all the participants. The vehicle population in the study area between the years 2006 to 2017 was collected from the MOI, which gives the data regarding the total number of commer-

cial and non-commercial vehicles in the city and the data considered for this study have been given in Table 1.

Table 1. Vehicle Population in the Study Area from 2006 to 2015

Type of Vehicle	Private	Taxis	Cruising Taxis	On Demand Taxis	Diplomatic Corp	Governmental
2006	975,315	597	2,753	5,949	855	3,360
2007	1,038,327	545	2,577	6,500	839	3,280
2008	1,088,801	498	2,482	6,470	850	3,056
2009	1,141,191	475	2,331	6,843	911	2,716
2010	1,193,085	472	2,212	6,935	965	2,662
2011	1,256,655	466	2,113	7,272	991	2,549
2012	1,331,703	456	2,247	7,911	1,023	2,663
2013	1,411,372	475	4,455	10,323	1,080	2,640
2014	1,485,926	472	4,757	10,323	1,104	2,699
2015	1,552,738	525	6,109	10,824	1,124	2,699
2016	975,315	597	2,753	5,949	855	3,360

The data on private vehicle population from the year 2006 to 2017 has been collected, which has been projected to find out the expected private vehicle population for 2027 (Table 2). These projected values have been considered for various scenario analyses.

Table 2. Private Vehicle Population and Projection for 2027

Private vehicles			
year	**Vehicle population**	**Growth rate**	**forecasted population**
2006	975,315		
2007	1,038,327	6.461	1,038,327
2008	1,088,801	4.861	1,088,801
2009	1,141,191	4.812	1,141,191
2010	1,193,085	4.547	1,193,085
2011	1,256,655	5.328	1,256,655

2012	1,331,703	5.972	1,331,703
2013	1,411,372	5.982	1,411,372
2014	1,485,926	5.282	1,485,926
2015	1,552,738	4.496	1,552,738
2016		5.305	1,635,106
2017		5.176	1,719,743
2018		5.211	1,809,364
2019		5.256	1,904,458
2020		5.334	2,006,049
2021		5.335	2,113,073
2022		5.264	2,224,311
2023		5.184	2,339,630
2024		5.174	2,460,673
2025		5.249	2,589,830
2026		5.243	2,725,605
2027		5.250	2,868,700

Energy Sector

With respect to the energy sector the fuels that were of prime concern were petrol and diesel. Petrol of three different types in terms of octane number is used by the private vehicles. Thus data relating to the fuel consumption per day by different petrol

driven vehicles have been collected and the individual fuel efficiency values have been used in the model analysis.

Emissions Sector

An inventory regarding the GHGs and other urban pollutants from transport sector were collected from the vehicles on a monthly basis. Emission factors for trace gas emissions viz. CO, CO_2, NOx, hydrocarbon (HC), particulate matter (PM) from various types of vehicles have been identified.

5 Development of Model and Scenario Analysis

The time series model applicable for the selected scenarios of road transport sector was developed based on the collected data pertaining to vehicular emissions and fuel consumption.

The baseline situation of the transport sector is established in the BAU scenario using the assessment model of this study. The BAU scenario identifies the CO_2 emissions assessment under the assumption that the trend of present social and economic development will go on in the future after a basic year, and that technological efficiency will also persist to improve based on patterns seen from the past to the present.

5.1 Base case or Business as Usual Scenario

Assessment results of the BAU scenario are shown in Table 3. Total fuel consumption in the transport sector for private vehicles in 2017 is reported at 3, 218, 484, 201.20 liter and the corresponding fuel cost is USD 1,258,893,734.68. At the same time, for the year 2027 the total fuel consumption in the transport sector for private vehicles is reported at 7,977,836,543.71 liter and the corresponding fuel cost is USD 3,120,490,210.08. This means, there is an increase of 40.34% in fuel consumption while following the BAU scenario.

Table 3. Total fuel consumption from 2016 to 2027

Year	Total fuel consumption (liter/year)	Fuel Price/year
2016	3,218,484,201.20	380,853,963.81
2017	4,801,258,201.39	568,148,887.16
2018	5,032,082,257.15	595,463,067.10
2019	5,296,279,083.75	626,726,358.24

2020	5,578,808,257.55	660,158,977.14
2021	5,876,440,990.41	695,378,850.53
2022	6,185,793,220.86	731,985,531.13
2023	6,506,494,273.45	769,935,155.69
2024	6,843,114,235.55	809,768,517.87
2025	7,202,299,229.07	852,272,075.44
2026	7,579,890,413.91	896,953,698.98
2027	7,977,836,543.71	944,043,991.01

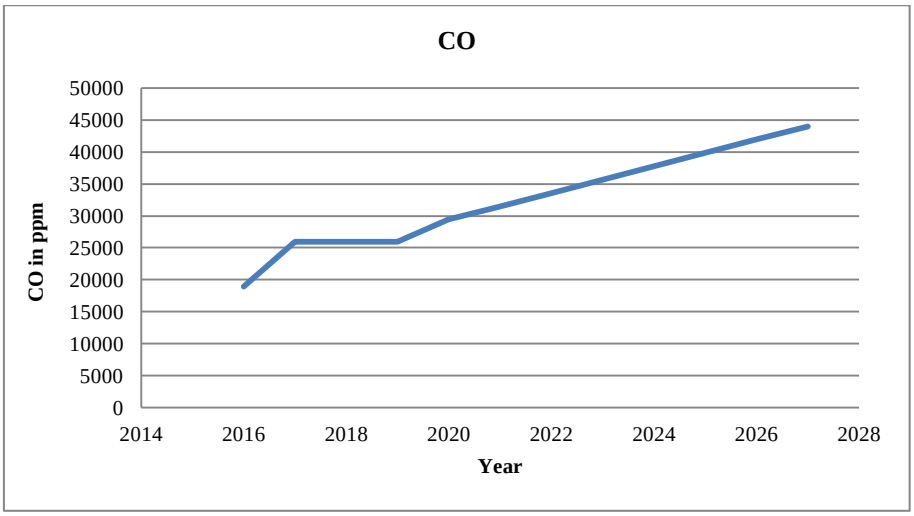


Fig. 1. CO emission in BAU scenario from 2014 to 2027.

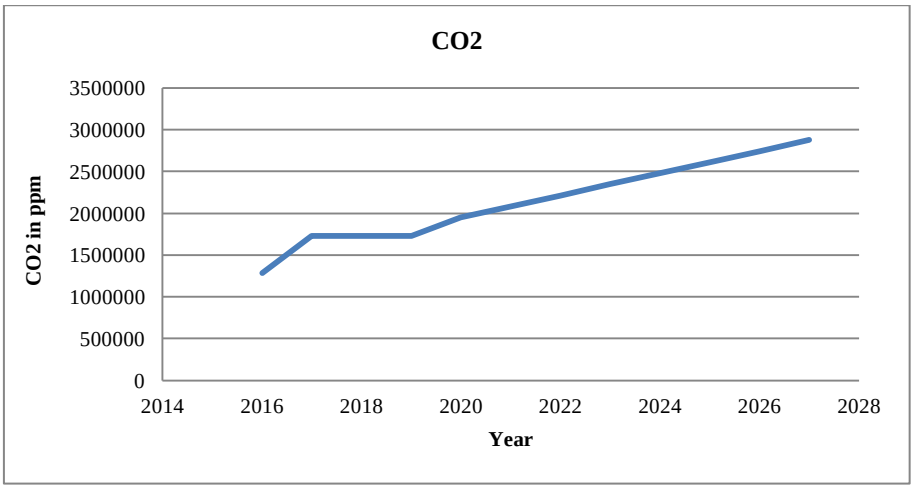


Fig. 2. CO_2 emission in BAU scenario from 2014 to 2027.

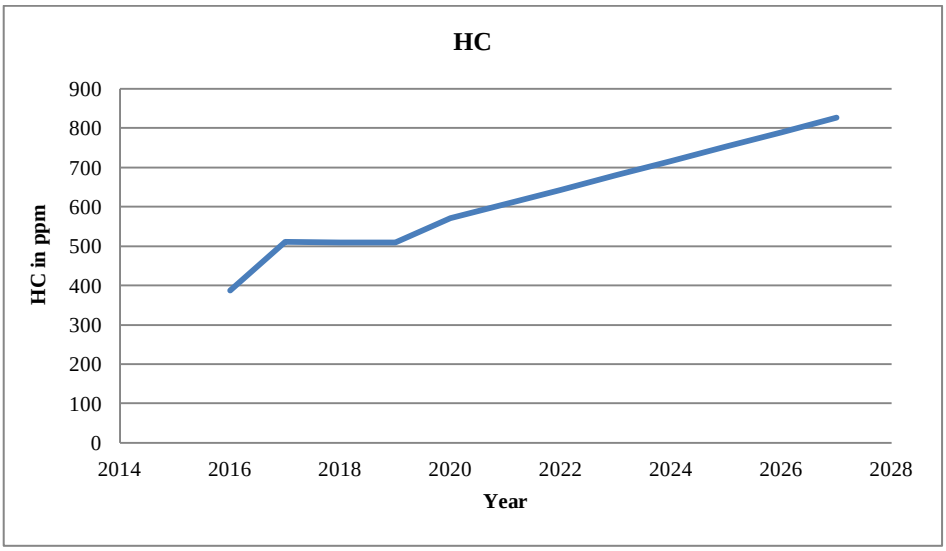


Fig. 3. HC emission in BAU scenario from 2014 to 2027.

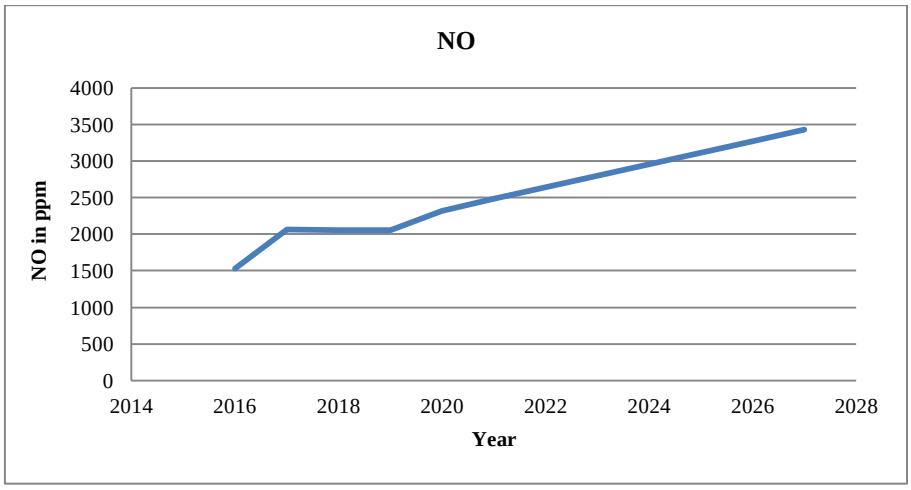


Fig. 4. NO emission in BAU scenario from 2014 to 2027.

Fig. 1 to Fig 4 explains the emissions in ppm from 2016 to 2027 in terms of CO, CO2, NO and HC while following the BAU scenario. There can be gradual increase in emissions if we follow the business as usual or baseline scenario.

5.2 Alternative Scenario Analysis and Results

There is one alternative scenario identified and tested for its feasibility; which is, 70/30 scenario. This means, partial efforts scenario such that minimal efforts are undertaken by the authorities to achieve a modal Split of 70:30 between public and private mode in road transport sector. Public transport has been augmented in a phase wise manner and simultaneously the growths of cars have been restricted for the year 2027. The hypothesis considered is that one bus will replace 40 cars. This rise can be attributed to the launch of metro rail in the country and a maximum utilization of its services. The tendency of change in share of trips between public and personalized modes of travel to the proportion of 70:30 can be observed in the results given in Table 4.

Table 4. 70/30 Scenario: Fuel Consumption and Economic Savings

Year	Diesel		petrol		Economic Savings in Kuwaiti Dinars
	Total fuel consumption liter/year	Price in Kuwaiti Dinars	70% of fuel consumption liter/year	Price in Kuwaiti Dinars	
2016	113,210,832.41	10,755,029.08	2,252,938,940.84	266,597,774.67	172,501,933.44
2017	168,885,227.83	16,044,096.64	3,360,880,740.97	397,704,221.01	257,334,282.51

2018	177,004,510.65	16,815,428.51	3,522,457,580.00	416,824,146.97	269,705,819.36
2019	186,297,687.44	17,698,280.31	3,707,395,358.62	438,708,450.77	283,866,045.28
2020	196,235,708.24	18,642,392.28	3,905,165,780.28	462,111,284.00	299,008,834.77
2021	206,704,999.79	19,636,974.98	4,113,508,693.29	486,765,195.37	314,961,133.63
2022	217,586,527.03	20,670,720.07	4,330,055,254.60	512,389,871.79	331,541,565.45
2023	228,867,251.39	21,742,388.88	4,554,545,991.42	538,954,608.98	348,730,263.04
2024	240,707,926.60	22,867,253.03	4,790,179,964.88	566,837,962.51	366,772,170.56
2025	253,342,331.36	24,067,521.48	5,041,609,460.35	596,590,452.81	386,023,501.92
2026	266,624,177.62	25,329,296.87	5,305,923,289.73	627,867,589.29	406,261,354.70
2027	280,622,013.18	26,659,091.25	5,584,485,580.60	660,830,793.70	427,590,176.75

From the table 4, it is evident that there is huge saving potential by incorporating this scenario. At the same time, this suggested scenario needs facilities such as, excellent public transport system, connecting different segments of the country and the launch of metro rail in the country and a maximum utilization of its services. Table 5 show the emission savings in scenario 2 - 70/30 scenarios- in the transport sector. The quantity of CO_2 and CO emissions that can be reduced by introducing the proposed scenario is enormous and hence this will surely fetch carbon emission reduction programs an upper hand for the state of Kuwait.

Table 5. Emission Savings in Scenario 2: 70/30 Scenarios in the Transport Sector

Year	Saving emissions (ppm)				
	O_2	CO	CO_2	HC	NO
2016	45,860,121,197.33	15,498,195,362.85	1,052,301,115,408.81	316,439,142.63	1,249,447,338.08
2017	64,053,082,075.26	22,332,578,418.61	1,486,455,830,226.78	438,608,620.61	1,770,950,103.98
2018	67,403,552,336.95	23,453,540,843.29	1,563,618,750,318.45	460,657,552.63	1,861,080,990.66
2019	70,948,766,067.87	24,686,113,910.28	1,645,780,229,558.46	484,883,048.72	1,959,020,643.41
2020	74,733,490,981.99	26,002,984,422.10	1,733,571,235,778.50	510,747,901.77	2,063,525,392.36

2021	78,720,564,052.83	27,390,258,020.44	1,826,058,219,856.58	537,996,584.77	2,173,615,613.15
2022	82,864,633,926.34	28,832,157,534.95	1,922,187,013,485.35	566,318,224.30	2,288,040,796.55
2023	87,160,732,158.85	30,326,954,231.16	2,021,842,365,110.13	595,678,864.75	2,406,663,754.94
2024	91,670,079,454.43	31,895,949,415.78	2,126,444,393,746.07	626,496,904.15	2,531,174,901.48
2025	96,481,706,991.48	33,570,120,851.47	2,238,058,329,960.07	659,380,804.48	2,664,032,546.30
2026	101,539,903,117.38	35,330,084,066.58	2,355,391,846,614.42	693,949,817.97	2,803,698,391.00
2027	106,870,773,256.67	37,184,922,257.16	2,479,050,503,713.49	730,382,355.81	2,950,893,253.06

6 Conclusions and Recommendations

A major goal of improved transportation technologies and associated energy policies is the reduction of CO_2 emissions, which is one of the prime goals of energy efficiency and carbon policies. The results obtained from the baseline scenario in estimating energy consumption and CO_2 emissions in the transport sector were based on the energy consumption characteristics of the present transport sector and socio-economic development prospective prevailing in the country. According to the CO_2 emissions results in the BAU scenario, CO_2 emissions in the transport sector are projected to continually increase until 2027. Energy consumption in the transport sector is expected to increase by about 42.4% from 2016 to 2027, and CO_2 emissions are expected to increase considerably too.

Based on the findings, utilization tendency, and historical data, the potential energy consumption of the transport sector is projected to expand. Energy and economic facets have the same trend, along with similar emissions of GHG, which are increasing due to the high dominance of fossil fuel consumption. Car ownership is growing in parallel to population growth; however this trend must be dejected. The findings of this study point to the need for a set of recommendations for nationwide implementation, to prevent negative ecological impacts and damage to the environment in terms of GHG emissions, and moreover, to provide equal access to mobility by making sure that individual should be able to move freely. Yet, there remains technological, economic and institutional ambiguity relating to the introduction of new technologies in future transportation associated markets. To surmount such restrictions, technological

supply systems or policies required to be endorsed at the national level. It is also essential to institute a long-term comprehensive plan and prepare an ample system for executing such plans to activate the establishment of such measures by linking the plan to a long-term national CO_2 reduction goal.

7 References

1. AboulNaga, M., K. Al-Sallal and R. El Diasty. 1997. The impact of city urban patterns on building energy consumption in hot climates: Al-Ain city as a case study. Proceedings of the International Solar Energy Society (ISES) Solar World Congress, S. Korea. 147-158.
2. Center for Clean Air Policy (CCAP). 2008. Transportation Emissions Guidebook, "Part One: Land Use, Transit & Travel Demand Management." Available at http://www.ccap.org/guidebookAccess/login.php. Accessed on 4/15/08
3. Khalili, N., Duecker, S., Ashton, W., Chavez, F., 2015. From cleaner production to sustainable development: the role of academia. J. Clean. Prod. 96, 30-43.

Tourism – the factor of employment sustainability in Croatian economy

Justin Pupavac[1] and Drago Pupavac[2][0000-0001-2201-0598]

[1] Faculty of Tourism and Hospitality Management, Primorska 42, 51414 Ika, Croatia
[2] Polytechnic of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
justin.pupavac@gmail.com
drago.pupavac@veleri.com

Abstract. The number of people employed within the Croatian economic system has been decreasing. In 2015 this number is by 12.73% smaller than in 2008. Adding to this negative trend is the fact that young experts and scientists are leaving the country which altogether leads us to question sustainability of employment in Croatian economy. If new jobs are not created and employment is not increased, there will be no sustainable development. In relation to this the main goal of this paper is to determine the role of tourism in job preservation. This scientific paper is based on econometric analysis, thus anticipating the needs for human resources in Croatian tourism sector until 2030. The methods of correlation and regression analysis have been used to investigate the rules and regulations which govern relations among stochastic phenomena in tourism. The statistical analysis refers to the period from 1986 to 2015. Data analysis and numerical calculations have been performed using *SPSS* software.

The main finding of this paper points to the fact that if tourism is regarded as an intensive labor activity, it will remain an important source of employment in Croatian economy. In other words, it will remain an important factor contributing to sustainability in employment. Considerable difficulties have occurred due to different approaches in measuring the number of people employed in tourism, or due to the fact that tourism is characterized by great seasonal fluctuations of workforce. The scientific methodology applied lends objectivity in the assessment of possibilities of tourism in creating new jobs. These findings should be helpful with a scientific approach in direct and indirect contribution of tourism in overall employment.

Key words: tourism, sustainability, employment

1 Introduction

Tourism is one of the most important sectors of Croatian economy, creating new jobs for the locals and opportunities for earning. Lately, tourism has been expanding globally and it is one of the fastest growing industries. Only in 2015, its annual growth rate was 2,8%, while the global economy had a growth of 2,3%. According to the WTTC estimates, the share of tourism in the global GDP has been growing steadily since 2010. In 2015, it was estimated at 9,8% (7,2 trillion USD). Currently, there are about 264 million tourism employees in the world, which makes it one of the 11 most important professions. Estimates for tourism growth in the course of the next decade are very encouraging – it is expected that in 2026 there would be 370 million workplaces or 100 million more than in 2015, which is an increase of more than 40%. The contribution of tourism in GDP and employment is apparent in national or regional economies as well. For instance, the direct share of tourism in Croatian GDP was 10,2% in 2016. In particular regions of Spain, as the Canary Islands, tourism makes for 31,4% of GDP and 35,9% of total employment, while in the Balearic Islands it is 45,5% of GDP and 34,8% of total employment (Perles-Ribes, et al., 2016). Accordingly, the main goal of this study is to explore the role of tourism in creating new jobs and sustaining employment in Croatian tourism by the year 2030 on the macro level. Also, the study will try to provide the determining variables of the total tourism employment and movements of tourism employment. The underlining hypothesis of this project is: tourism as the work intensive sector is the basic factor of employment sustainability in Croatian economy. In order to prove this hypothesis and maintain the goals of the research, secondary data sources will be used. The results are based on statistic methods of descriptive statistics and of correlation and regression analysis.

2 Theoretical background and research problem

Numerous authors have already been exploring the role of tourism in creating new workplaces and increase in total employment. Thus the analysis of the fundamentals of sustainable tourism management (Blažević, Maškarin Ribarić, Smolčić Jurdana, 2013), highlights not only the direct and total share of tourism in GDP and employment, but workforce as well, as an important indicator of sustainable tourism, and this is done through analysis of the number of employees in the HORECA sector, the number of permanent employees, the number of seasonal employees and the educational structure of employees. Obadić and Pehar (2016) have researched the impact of tourism on the main macroeconomic variables (GDP, employment and capital investments) in several Mediterranean countries (France, Greece, Spain, Italy) with special attention given to Croatia. According to this research, one of the main restrictions of tourism employment is its seasonal nature, while the importance of tourism in creating new workplaces in post-crisis period is particularly regarded. Rešetar and Čeh Časni (2014) tackle the importance of employment in tourist facilities for local residents, whose life quality and demographic structure depend on it, as tourism is the main source of employment in these areas. Here, they used the Holt-Winters additive and multiplicative model of exponential flattening to predict the level of tourism employment across different age groups. Pavlić, Tolić and Svilokos (2014) have studied the direct and indirect contribution of tourism to employment increase in Croatia. They based the research on quartal data for the period between 2000 and 2012. Interrelation of tourism and employment was explored by applying the Granger test of causality and Johansen's cointegrational approach. Dedu (2012) has done a research on the importance of human resources and the influence of tourism on employment. He focused in particular on the connection between tourist arrivals and employment within the tourist sector of Romania. Beneki, et al. (2016) have used the linear logarithm model to explore the possibilities of employment growth in the hotel industry of Greece. Their empirical results point to capacities of the hotel industry as the main factor for employment growth within the tourist sector. Baum, et al (2016) have focused on sustainability and workforce, stating that sustainable tourism and the contribution of tourism in achieving sustainable growth and full productive employment while maintaining dignity of work is among the 17 goals of the UN's sustainability Agenda 2030. Witt, Song and Anhill (2004) focus on prognostics of tourist demand as the main factor in planning and decision-making in tourism. Their predictions on employment generated by the Danish tourist industry are based on estimates of tourist spendings, using the vector autoregressive model, and then direct, indirect and induced effects of tourism in total

employment are estimated using the input-output model. The main effect of how much foreign tourist spend on employment is estimated within the retail industry, tourism and hospitality. A scientific discourse by Petenlar, Rasekhi and Ebrahimzadeh (2016) where the impact of direct foreign investments on employment is discussed is rather interesting. Besides direct foreign investments as an independent variable, GDP, tax rate and index of globalization are also treated as independent variables. Their econometric analysis is based on panel data for 48 countries, including Iran. The main result of their work is the fact that direct foreign investment and index of globalization have a negative impact on employment in tourism, while the GDP and tax rate have a positive impact. Chao, Hazari, Laffargue and Yu (2009) have researched short-term and long-term effects of tourism on wealth and employment in small open economies using the dynamic simulation model. It is no secret that tourism can make unprofitable goods into profitable ones. Increased demand for unprofitable goods inflates their relative pricing, which results with expansion of the unprofitable goods sector while the profitable goods sector is shrinking. This movement of total output within a short period of time leads to employment increase. However, in the long term, higher relative price leads to higher pays which would eventually have a negative impact on employment. The results show that when the effect of total output is dominant, tourism expansion, employment growth and wealth are inevitable. However, in reality, tourism can have a negative effect on employment and wealth if the costs are increased. Darabos and Konyves (2015) highlight tourism as an important factor in the long-term stabilization of market and increase in total employment on the global level and in the EU. Bohlin, Brandt and Elbe (2016) have studied the role of tourism in regional development of peripheries, focusing on such Swedish regions, i.e. the exact contribution of tourism in eradicating development disparities between particular regions and in increasing employment in the periphery. The EU has started several programs and projects that should stimulate growth and employment of the periphery, and these are often based on tourist programs of development. Their statistical analysis is based on data about tourist overnight stays in certain regions. Figueroa and Rotarou (2016) have taken the Eastern Island of Chile as an example to show that locals, fully aware of negative effects of uncontrolled tourism development, are fully dependent on tourism as the main source of income and employment. Nengroo, Bhat and Khaki (2016) have analyzed the economic impact of tourism on development of Jamma and Kashmir in India. Their research is based on secondary data. Estimates on movements in tourism employment in these Indian regions by the year 2020 are based on tourist transport

and the WTTC's estimate of one million rupees (about 15 000 USD) generating about 90 workplaces in tourism, which is much more than in agriculture (40 workplaces) or manufacture (13 workplaces). According to Okun's Law, each 2% fall of GDP (in relation to potential GDP), means an increase of 1 index point of unemployment rate (Samuelson, 1992:574). As the registered unemployment rate in Croatia in 2008 was 13,2%, and the fall of GDP according to Okun's Law was 12% in 2014 (in relation to 2008), unemployment rate would be 19,2%, which means a 6% increase. The official statistics for 2014 show a registered unemployment rate to be 17,3%. This slight discrepancy between the calculation and reality is actually negligable, and the calculated prediction is rather accurate, since economy is not an exact science. The number of employees in Croatian economy has been constantly decreasing since 2008. In 2015 there were 12,7% less employees than in 2008. There was a loss of about 200 000 workplaces during the recession. However, Croatian economy has suffered from another worrying trend over the past years – highly educated young people and scientists have been moving out of the country in great numbers. This raises a question about sustainability of employment in Croatia. Without creating new workplaces and increasing employment, there is no sustained development. Recession has caused the loss of workplaces (cf. Fig. 1) in the secondary sector (construction, processing industry), but also in the tertiary sector (retail, transport), which makes the sustainability questionable in of all of its components (economic, social and ecological).

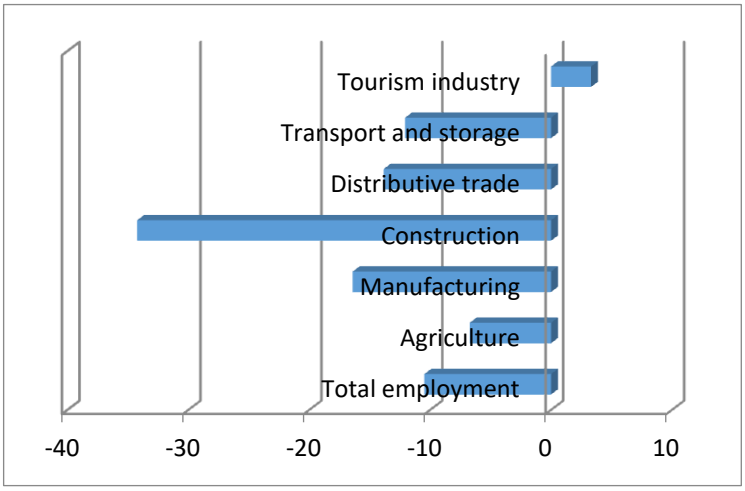


Fig. 1. Labor market dynamics in Croatia 2009-2014

Negative effects of recession have affected tourism as well and diminished its contribution to new employments. After the absolute and relative fall of this contribution of tourism in 2009 and 2010, the sector has somewhat recovered and got back to its former numbers, despite the ongoing crisis and negative trends in Croatian economy.

3 Methodology

To make a scientifically based estimate of the number of employees in tourism by 2030, the model of multiple regression was applied. The fundamental hypothesis of the model is that the number of employees in Croatian tourism (facilities and food services) depends on: 1) tourist arrivals – TA, 2) the number of available beds – B, 3) overnight stays – TN, 4) gross domestic product – GDP and 5) time – T. Accordingly, the model for such estimation can be written as follows:

$$ET = f(TA, B, TN, GDP, T) \quad (1)$$

The variable of the number of employees is dependent, while the arrivals, beds, overnight stays, investments in tourist economy and GDP are independent. With the assumption that the number of employees depends on the afore mentioned independent variables, the linear model is the following:

$$ET = b_0 + b_1TA + b_2B + b_3TN + b_4GDP + b_5T \quad (2)$$

After the estimate on the number of employees in tourism by 2030 is calculated, the estimate on the share of tourism in seasonal and total employment can be made. Total share of tourism in the seasonal and total employment by 2030 will be predicted based on the established relations between the average number of employees in tourism, the number of seasonal workplaces and the total contribution of tourism employees to total employment (cf. Table 1).

Table 1. The relationship between the average number of tourism employees, the number of seasonal workplaces and the total contribution of tourism to employment

	2010.	2011.	2012.	2013.	2014.
AE	89,6	81	82,6	84	85
AE+SW	122	130,5	121,6	127,8	129,7
TCE	294,2	309,6	286,2	299,1	300,6
(AE+SW)/AE	1,36	1,61	1,47	1,52	1,52
TCE/AE	3,28	3,82	3,46	3,56	3,53

Legend: AE – average number of employees in tourism; SW – number of seasonal workers; TCE – total contribution to employment

Data in Table 1 show that there is one seasonal worker per two employees in tourism, that is, on average, there are 3,5 workers connected to tourism per one tourism employee.

4 Research results

Based on data on the number of arrivals, overnight stays, available beds, GDP and the number of employees in tourism from 1986 to 2015, a correlation analysis was made to confirm the correaltion between these variables (cf. Table 2).

Table 2. Correlation analysis

Correlations (tourism.sta) Marked correlations are significant at p < ,05000 N=30 (Casewise deletion of missing data)							
	Means	**Std.Dev.**	**TA**	**TN**	**B**	**GDP**	**ET**
TA	8409,1	3497,77	1,00	0,96	0,98	0,92	0,84
TN	45861,3	19379,05	0,96	1,00	0,94	0,92	0,91
B	834,5	127,51	0,98	0,94	1,00	0,94	0,83
GDP	265883,6	44442,83	0,92	0,92	0,94	1,00	0,91
ET	78332,0	15261,21	0,84	0,91	0,83	0,91	1,00

Correlation analysis has confirmed the existence of a statistically strong and positive connection between the movements of the number of employees in tourism and tourist arrivals (r=0,85), overnight stays (r=0,91), beds (r=0,84) and GDP (r=0,91).

Since there was a strong and positive interdependence between the number of employees in tourism, tourist arrivals, overnight stays, available beds and GDP, a regression analysis was conducted using the same data. While evaluating the parameters of the function (2), we could not find a conclusive regression model. After many trials and errors procedures, a simpler model has been developed for the estimate of the number of employees in Croatian tourism industry:

$$ET = f(TA, GDP, t) \quad (3)$$

Statistical and econometric attributes of the linear relationship between the number of the employed in tourism and tourist arrivals, GDP and time as the examined variables can be seen in Table 3.

Table 3. Regresssion analysis for the number of employees in the tourism industry

Regression Summary for Dependent Variable: ET (tourism.sta) R= ,97742666 R2= ,95536287 Adjusted R2= ,95021243 F(3,26)=185,49 p						
	Beta	**Std.Err. - of Beta**	**B**	**Std.Err. - of B**	**t(26)**	**p-level**
Inter-cept			17454,50	7801,194	2,23	0,03
TA	0,67	0,133236	3,24	0,641	5,05	0,00
GDP	0,51	0,117903	0,19	0,045	4,33	0,00
t	-0,55	0,056807	-1063,60	108,649	-9,78	0,00

Regression analysis between the number of employees in tourism (ET), tourist arivals (TA), GDP and time (t) has resulted with the following model of multiple linear regression:

$$ET = 17454{,}50 + 3{,}24\,TA + 0{,}19\,GDP - 1063{,}60\,T \quad (4)$$

According to regression analysis (cf. Table 3), it can be concluded that there is a statistically significant correlation between the number of employed in the tourism industry, tourist arrivals, GDP and time ($R=0{,}97$; $F(3,26)=185{,}49$; $p<0{,}01$).

Based on the given model (4), an estimate of the number of employees in Croatian tourism by 2030 was made. It seems appropriate to assume that the considered variables - tourist arrivals and GDP - will increase in the coming period, so if we anticipate that the average growth rate of tourist arrivals will grow at an annual rate of 5% and GDP will grow at an annual rate of 3%, the number of employees in tourism by 2030 will also be rise (as shown in Table 4).

Table 4. Estimate of the number of employees in tourism by 2030

Year	Number of employees in the tourism industry
2020.	105926
2025.	127566
2030.	155417

Data in Table 4 show a significant increase of an average number of employees in tourism and this by 67 586 workplaces. This significant growth will affect the grow of seasonal employment and an increase in the total contribution of tourism to employment (cf. Table 5).

Table 5. Estimate of the average number of employees in tourism, the number of seasonal workplaces and the total contribution of tourism to employment by 2030.

	2020.	2025.	2030.
AE	105926	127566	155417
AE+SW	158889	191350	233126
TCE	370741	446482	543961

According to obtained data, the total contribution of tourism to employment will increase by 243 361 workplaces.

5 Conclusion

Current economic environment is marked by a slight economic recovery and shift from the global recession, which has made a devastating and ongoing impact on Croatian economy, visible in decreased GDP in relation to past values and decrease in total employment. In Croatia the recession has caused a loss of about 200 000 workplaces, making sustainability questionable in all of its components (economic, social and ecological). Negative effects of recession have affected tourism as well by diminishing its contribution to total employment. After the absolute and relative decrease of this contribution in 2009 and 2010, the tourist industry has somewhat recovered and got back to its prerecession values in spite of ongoing negative trends in Croatian economy. Thus, the tourism industry has proven its economic importance by sustaining the existing workplaces and creating new ones. To make an estimate of the number of employees in tourism by 2030, the correlation between the number of employees in tourism and tourist arrivals, beds, overnight stays and GDP has been made for the period between 1986 and 2015. The correlation analysis has proven the existence of a statistically strong and positive correlation between the number of employees in tourism and the mentioned variables. Afterwards, a regression model was created to include the variables of tourist arrivals, GDP and time. Then an estimate of the number of employees in tourism, the number of seasonal workers and the total contribution of tourism to employment by the year 2030 was made based on this regression model. With an assumed 5% annual growth rate of tourist arrivals and 3% annual growth rate of GDP, the number of employees in tourism is expected to grow by 77%, the number of seasonal workers by 94,3% and the growth of total contribution of tourism to employment by 81%. The estimates are encouraging, so in 2030, the tourism industry might directly or indirectly employ 543 961 people.

References

Baum, Tom; Cheung, Catherine; Kong, Haiyan; et al.: Sustainability and the Tourism and Hospitality Workforce: A Thematic Analysis, Sustainability, Volume: 8 Issue: 8 Article Number: 809 Published: AUG 2016.

Beneki, Christina C.; Rerres, Kyriakos; Chionis, Dionysios P.; et al.: How to stimulate employment growth in the Greek hotel industry, Tourism Economics, Volume: 22 Issue: 5 Pages: 865-883, 2016.

Blažević, B., Maškarin Ribarić, H., Smolčić Jurdana, D.: The Analytical Base for Management of Sustainable Tourism, Ekonomski pregled, 64 (2) 143-158, 2013.
Bohlin, Magnus; Brandt, Daniel; Elbe, Jorgen.: Tourism as a vehicle for regional development in peripheral areas - myth or reality? A longitudinal case study of Swedish regions, EUROPEAN PLANNING STUDIES Volume: 24 Issue: 10 Pages: 1788-1805 Published: OCT 2016.
Chao, C., et al.: A Dynamic Model of Tourism, Employment and Welfare: The Case of Hong Kong, Pacific Economic Review, 14:2., pp. 232 – 245, 2009.
Darabos, E., Konyves, E.: Significance of Tourism in Economy of the EU's Member States, Particulary Concerning the Employment, Annals of the University of Oradea, pp. 689 – 695, 2015.
Dedu, E.: The Relationship Between Tourism and Employment in the Tourist Sector, Managerial Challenges of the Contemporary Society, Risoprint Publishin Houese, pp. 121 – 124, 2012.
Eugenio, Figueroa B.; Rotarou, Elena S.: Sustainable Development or Eco-Collapse: Lessons for Tourism and Development from Easter Island SUSTAINABILITY Volume: 8 Issue: 11 Article Number: 1093 Published: NOV 2016.
Nengroo, A., Bhat, G., Khaki, A.: Economic Impact of Tourism in Jammu and Kashmir, Indian Journal of Economics and Development, Vol 4 (3), pp. 1 – 5, 2016.
Obadić, A., Pehar, L.: Employment, Capital ana Seasonality in Selected Mediterranean Countries, Zagreb International Review of Economics & Business, Vol 19, Special Conference Issue, pp, 43-58, 2016.
Pavlić, I., Šuman Tolić, M., Svilokos, T.: Impact of Tourism on the Employment in Croatia, Recent Advances in Business Management and Marketing, (aviable: http://www.wseas.us/e-library/conferences/2013/Dubrovnik/MATREFC/MATREFC-34.pdf), 2014.
Perles-Ribes, et al.: Unemployment effects of economic crises on hotel and residential tourism destinations: The case of Spain, Tourism Management, 54, pp. 356 – 368, 2016.
Petenlar, S., Rasekhi, S., Ebrahimzadeh, S.: The Effect of Foreign Direct Investment on Employment in Tourism Industry: A Case Study of Selected Countries, International Journal of Management, Accounting and Economics, Vol 3, No 9., pp. 459-472, 2016.
Rešetar, M., Čeh Časni, A.: Forecasting level of Employment in Tourist Accomodation Sector for Selected EU Countries Using Smoothing Models, Zbornik Ekonomskog fakulteta u Zagrebu, Vol 12, No 2, 2014.
Samuelson, P., Nordhaus, W.: Economics, 14th edition, McGraw-Hill, New York, (1992).
Witt, S., Song, H., Anhill, S.: Forecasting tourism-generated employment: the case of Denmark, Tourism Economics, 2004, 10 (2), pp. 167 – 176, 2004.

Enhancement of time series analysis by including label variables

José Carlos García-García, Ricardo García-Ródenas, and Francisco P. Romero

Escuela Superior de Informática, University of Castilla-La Mancha,
Paseo de la Universidad 4, 13071 Ciudad Real, Spain
{josecarlos.garcia,ricardo.garcia,franciscop.romero}@uclm.es

Abstract. The area of time series modeling is in his full development to accommodate a growing volume of time series data from heterogeneous data sources. In this context, there is a need for particular attention to process additional information. In this paper, we propose a methodology to include "label variables" (geographical locations, demographic data) to the analysis of time series. These variables allow to partition the dataset using tree clustering and perform a tuned time series analysis on each of the partition components. Some experiments illustrate the feasibility and efficiency of the approach to the prediction of energy consumption and the establishment of patterns of electrical consumers.

Keywords: Decision Tree, Forecasting Electricity Consumption

1 Introduction

Nowadays, there is an increasing use of time series data that can be obtained from multiple areas such as economics, social sciences or energy consumption. New sensor technologies are currently generating large volumes of data. This situation drives new approaches to the processing of time series.

Classic time series approaches are categorized into representation and indexing, similarity measure, segmentation, visualization and mining [1]. In this work, a generic time series task $\mathcal{M}$ is carried out on an specific dataset $\{\mathbf{x}_i\}_{i\in I}$. The meaning of $\mathbf{x}_i$ depends on each specific problem, from a single temporary observation in the instant i to a complete time series $\mathbf{x}_i$. In this context, we assume that it is possible to apply $\mathcal{M}$ in any data subset $J \subseteq I$, and the existence of a measure F to evaluate the performance in the task execution in any data subset $J \subseteq I$. Also, for each observation i additional information $\mathbf{z}_i$ is available, these *label variables* represent other characteristics of each element i such as the geographical location where the information was captured, demographic aspects, etc.

The main purpose of this work is to include the information represented by $\{\mathbf{z}_i\}_{i\in I}$ in order to obtain a data partition $\{I_j\}$ so that the $\mathcal{M}$ task can be performed separately on each of the components of the partition, I_j, in the most effective possible way.

Examples of these problems are the piecewise regression models [3] in which $\mathcal{M}$ are regression models and the *label variables* correspond to the moments where the observations are captured. Another example is represented by the the so-called prediction-after-classification approaches to predict traffic flows [2]. In these problems, $\mathbf{x}_i$ represent the traffic flow time series for each day i, $\mathcal{M}$ is a traffic flow prediction model and the labels $\mathbf{z}_i = \mathbf{x}_i$. In order to solve this problem, it is necessary to carry out a time series clustering process, (for example using $k-$means [4]). The predictive model is executed on each subset of time series.

The rest of the paper is organized as follows. In Section 2 we illustrate our methodology and the mathematical principles of the proposal and a brief enumeration of the key aspects of the resolution of the proposed model. Section 3 presents the experiments and the results obtained. We terminate with some conclusions and possible future research lines.

2 Methodology

2.1 Optimization model

The proposed method consists of iterative perform a partitioning of the *label data* $\{\mathbf{z}_i\}_{i\in I}$ using a decision tree. The main difference with the classic decision tree algorithms consists of the use of a global quality measure F of the $\mathcal{M}$ model in each of the clusters of the partition instead of the common quality indices such as the Gained Information or Gini index.

We will explain the procedure to partition the data, so, in Fig. 1 we show each of the steps to understand the method. This example illustrates a division in the case of $\mathbf{z}_i \in \mathbb{R}^2$. Figure 1.(a) considers the region `R1` consisting of the space $\mathbb{R}^2$. On the other hand, Fig. 1.(b) shows a graph in which the region `R1` is divided into `R2` and `R3`. This initial step uses a hyperplane in order to divide the space $\mathbb{R}^2$ into two new regions, specifically, we consider:

$$\texttt{R2} = \texttt{R1}^+ = \left\{\mathbf{z} \in \texttt{R1} : \mathbf{w}_1^T \cdot \mathbf{z} + b_1 \geq 0\right\}$$
$$\texttt{R3} = \texttt{R1}^- = \left\{\mathbf{z} \in \texttt{R1} : \mathbf{w}_1^T \cdot \mathbf{z} + b_1 < 0\right\}$$

Figure 1.(c) indicates that the region `R2`, in turn, is divided into two new partitions, `R4` and `R5`. So, we specify other hyperplane to carry out the following step:

$$\texttt{R4} = \texttt{R2}^+ = \left\{\mathbf{z} \in \texttt{R2} : \mathbf{w}_2^T \cdot \mathbf{z} + b_2 \geq 0\right\}$$
$$\texttt{R5} = \texttt{R2}^- = \left\{\mathbf{z} \in \texttt{R2} : \mathbf{w}_2^T \cdot \mathbf{z} + b_2 < 0\right\}$$

Finally, in Fig. 1.(d) we can see how the region `R5` is partitioned into two new ones, `R6` and `R7` through the third hyperplane:

$$\texttt{R6} = \texttt{R5}^+ = \left\{\mathbf{z} \in \texttt{R5} : \mathbf{w}_3^T \cdot \mathbf{z} + b_3 \geq 0\right\}$$
$$\texttt{R7} = \texttt{R5}^- = \left\{\mathbf{z} \in \texttt{R5} : \mathbf{w}_3^T \cdot \mathbf{z} + b_3 < 0\right\}$$

If we take into account the regions associated with the 1-degree nodes (leaves) of the considered tree, we obtain a partition of the space $\mathbb{R}^2$. Let N_1 be the set of these nodes. Note that the cardinality of the set N_1 is $m+1$, where m is the number of hyperplanes considered in the procedure. In turn, this partition of the space induces a partition $\{I_j\}_{j=1}^{m+1}$ of the data $\{\mathbf{z}_i\}_{i\in I}$ through the following relationship:

$$i \in I_j \Leftrightarrow \mathbf{z}_i \in \mathtt{Rj} \tag{1}$$

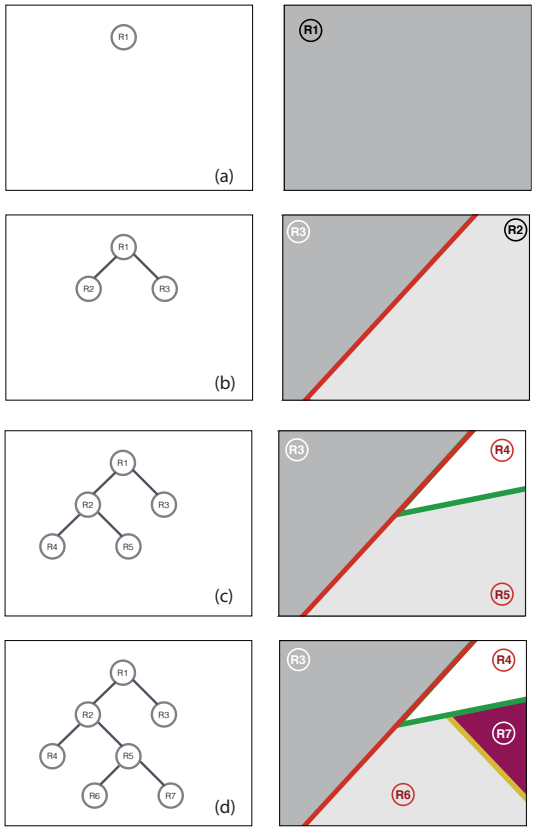


Fig. 1: Illustration of the relationship between decision tree and space partition

Let $\mathbf{W}$ be the matrix whose rows are the normal vector of each hyperplane $\mathbf{w}_j$ and let $\mathbf{b} = (\cdots, b_j, \cdots)$ be the vector of independent terms. Let $I_j(\mathbf{W}, \mathbf{b})$ be the procedure described previously in order to establish what data are contained by the component j according to the set of considered hyperplanes, which are

defined by the matrix $\mathbf{W}$ and the vector $\mathbf{b}$. Note that if a hyperplane is defined by the vector $(\mathbf{w}_j, b_j)$, then the vector $\lambda(\mathbf{w}_j, b_j)$ in which $\lambda > 0$ defines the same hyperplane. This fact lets us consider only the vectors $(\mathbf{w}_j, b_j)$ whose norm is less than 1, that is $\|(\mathbf{w}_j, b_j)\| \leq 1$. To simplify the problem structure we consider $\|\cdot\|_\infty$.

Suppose we have a time series task $\mathcal{M}$ which is applicable to any set of data $\{\mathbf{x}_i\}_{i\in J}$ where $J \subseteq I$. Let $\mathcal{F}(J)$ be an index related to the goodness of adjustment of the task $\mathcal{M}$ applied to the subset of data defined by the set J. In order to establish the partition of the data, we solve the following optimization problem:

$$\begin{array}{c} \text{Minimize } \sum_{j\in N_1} \mathcal{F}(I_j(\mathbf{W}, \mathbf{b})) \\ \text{subject to: } \|(\mathbf{w}_j, b_j)\|_\infty \leq 1; j \in N_1 \end{array} \tag{2}$$

The problem (2) has infinite solutions because if there is a hyperplane that separates two sets of points, therefore, there are infinite hyperplanes that divide them. For this reason, we introduce a procedure of *tie breaker* based on the choice of those hyperplanes that obtain the widest margins between points located in different components of the partition. The concept of margin is defined by only one hyperplane. We extend this definition to tree structure as follows:

Let $j \in \{1, \cdots, m\}$ and for all the points $\mathbf{z}_i \in \mathtt{R}_j^+ = \mathtt{R}_{j'}$, with $j' \in N_1$, hold:

$$\mathbf{w}_j \cdot \mathbf{z}_i + b_j \geq 0 \tag{3}$$

Let $b_j^i = -\mathbf{w}_j \cdot \mathbf{z}_i$, so it is satisfied

$$\mathbf{w}_j \cdot \mathbf{z}_i + b_j^i = 0 \tag{4}$$

We denote the point projection $\mathbf{z}_i$ in the hyperplane $(\mathbf{w}_j, b_j)$ by $\mathbf{y}_i$ and the distance by $\|\mathbf{z}_i - \mathbf{y}_i\|$. Because each point belongs to the previous hyperplanes, we have:

$$\mathbf{w}_j \mathbf{y}_i + b_j = 0 \tag{5}$$

$$\mathbf{w}_j \mathbf{z}_i + b_j^i = 0 \tag{6}$$

If we subtract both equations, we obtain:

$$\mathbf{w}_j \cdot (\mathbf{y}_i - \mathbf{z}_i) = b_j^i - b_j \tag{7}$$

On the other hand, the following relationship of scalar product is met:

$$\mathbf{w}_j \cdot (\mathbf{y}_i - \mathbf{z}_i) = \|\mathbf{w}_j\| \|\mathbf{y}_i - \mathbf{z}_i\| \cos(\alpha) \tag{8}$$

where α is the angle which the vectors $\mathbf{w}_j$ and $\mathbf{y}_i - \mathbf{z}_i$ form. Due to the selection method of points $\mathbf{z}_i$ and $\mathbf{y}_i$ and the normal vector $\mathbf{w}_j$ of the hyperplane,then $|\cos(\alpha)| = 1$. If we join both previous relationships, we obtain:

$$\|\mathbf{w}_j\| \|\mathbf{y}_i - \mathbf{z}_i\| = |b_j^i - b_j| \Rightarrow \|\mathbf{y}_i - \mathbf{z}_i\| = \frac{|b_j^i - b_j|}{\|\mathbf{w}_j\|} \tag{9}$$

So, we define:

$$d_j^+ = \min_{i \in \mathtt{R}_j^+} \|\mathbf{y}_i - \mathbf{z}_i\| = \min_{i \in \mathtt{R}_j^+} \frac{|b_j^i - b_j|}{\|\mathbf{w}_j\|} \tag{10}$$

In a similar way, we get d_j^- and the penalization term is:

$$\mathcal{P}(\mathbf{W}, \mathbf{b}) = \sum_{j=1}^{m} \min\{d_j^+, d_j^-\} \tag{11}$$

Finally, we consider the following problem:

$$\begin{array}{l} \text{Minimize } \sum_{j \in N_1} \mathcal{F}(I_j(\mathbf{W}, \mathbf{b})) - \delta \mathcal{P}(\mathbf{W}, \mathbf{b}) \\ \text{subject to: } \|(\mathbf{w}_j, b_j)\|_\infty \leq 1; j \in N_1 \end{array} \tag{12}$$

where $\delta \to 0^+$. This problem seeks the data partitioning that best fits the time series tasks and obtains the widest margins.

2.2 Resolution algorithms

The resolution of the model (12) requires methods based only on functional evaluations that can be applied to general computing models $\mathcal{M}$. The fact that the problem (12) has a bi-level optimization structure, that is, given a partition defined by $(\mathbf{W}, \mathbf{b})$, a second optimization procedure must be applied to each component in order to fit the time series approach $\mathcal{M}$. So, the type of method to be used is restricted. We propose the use of surrogate-assisted optimization or also named *Response Surface Methods* (RSMs) such as *Efficient Global Optimization* (EGO) and *Metric Stochastic Response Surface* (MSRS). These methods don't use derivates and are applied to expensive optimization problems such as the problem (12) in many domains of application. If the computational burden is small for the problem at hands, the metaheuristics can be used.

3 Experiments

We have carried out the following experiments:

- *Experiment* 1. We compare resolution strategies based on meta-heuristic algorithms with surrogate-assisted optimization methods.
- *Experiment* 2 *and* 3. We have applied the methodology to time series clustering in Experiment 2 and times series prediction in Experiment 3.

3.1 Experiment 1: Resolution methods

The first key about the methodology introduced in this paper to be analyzed is the appropriate optimization method to solve the problem (12). In Experiment 1, we take into account a basic clustering analysis. We assume $\mathbf{z}_i = \mathbf{x}_i$ and the model $\mathcal{M}$ to be adjusted is the mean of data, i.e $\mathcal{M}(\{x_i\}_{i\in J}) = \overline{\mathbf{x}}_J = \frac{1}{|J|}\sum_{i\in J}\mathbf{x}_i$. In this experiment, the quality measure $\mathcal{F}$ of adjustment of the model $\mathcal{M}$ is the *sum of squared error* (SSE) of instances within each of the regions with respect to its cluster center. This is $\mathcal{F}(I_j) := \sum_{i\in I_j}\|\mathbf{x}_i - \overline{\mathbf{x}}_{I_j}\|_2^2$. We have generated four random samples of 200 points each one distributed as a normal bivariate $\mathbf{x} \sim N(\mu, \Sigma)$ whose means are: $\mu_1 = (0, 1)$, $\mu_2 = (-1, 0)$, $\mu_3 = (1, -1)$ and $\mu_4 = (-1, 1)$ and variance-covariance matrix $\Sigma = \sigma^2 I_2$ where I_2 is the identity matrix and standard deviation $\sigma^2 = 0.25$.

The optimization methods proposed for the comparison are EGO and MSRS as sampling strategies based on response surfaces; and *Genetic Algorithm* (GA) and *Particle Swarm Optimization* (PSO) as heuristic methods. We set MSRS with *cubic radial basis function* (RBFcub), meanwhile EGO uses Kriging as surrogate method. Both strategies perform 200 main iterations and use *Symmetric Latin Hypercube Design* (SLHD) in order to carry out the design of initial points. For GA, we set 50 individuals of population and 200 generations that result in 10.000 functional evaluations. In the case of PSO, the above amounts remain unchanged for the number of particles and iterations, respectively. Table 1 presents a comparison of the resolution methods mentioned above concerning quality and CPU time. From these results, we can conclude that heuristic methods (GA and PSO) obtain the best results according to the quality measure and CPU time in a not too computationally expensive problem.

Table 1: Comparison of algorithms

Method	$\mathcal{F}$- Value	CPU (s)
PSO	236.47	18.43
GA	239.39	21.84
EGO	240.23	4263.67
MSRS	241.34	1564.25

In Fig. 2 we show a set of four groups of points that are partitioned through hyperplanes. The solution obtained by PSO and EGO are worse than MSRS and GA (because a point is included in an incorrect partition whose cluster center is closer to this point than the cluster center of the correct partition). On the other hand, the objective value $\mathcal{F}$ in GA is higher than in PSO due to the penalization since in PSO there is more separation between the yellow hyperplane and the points which are the closest to both regions. The same phenomenon is applied to methods based on response surfaces such as EGO and MSRS.

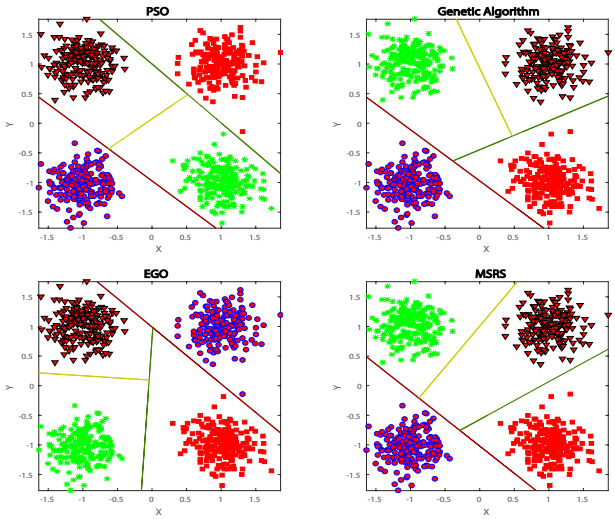


Fig. 2: Data partition with different methods and algorithms

The following test lets us know the evolution of algorithms regarding functional evaluations. We can observe the results in Fig. 3. For this test, we take into account the same set of data but modify some parameters of GA and PSO (Number of individuals/particles = 10, number of generations/iterations = 20).

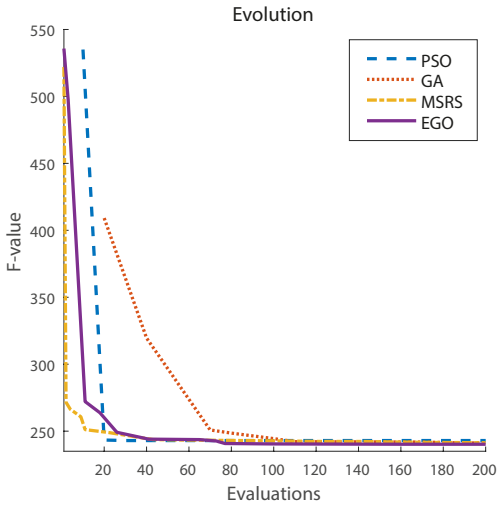


Fig. 3: Comparison of algorithms performance

In Fig. 3, we can observe which heuristic methods converge to an optimum solution slower than strategies based on response surfaces. As a conclusion, metaheuristic algorithms are adequate for piecewise machine learning problems in which the adjustment of models $\mathcal{M}_j$ is not computationally expensive, however, if the computational cost is high, the use of strategies based on surrogate models is more appropriate because they reached a correct solution with fewer iterations than heuristic methods.

3.2 Experiment 2: Time series clustering with spatial constraints

The second experiment consists of a real world problem. The objective is to solve a time series clustering problem. The variables $\mathbf{z}_i$ correspond to geographical coordinates related to the residence of clients while the variable $\mathbf{x}_i$ is associated with their temporal energy consumption. The data collected in $\mathbf{x}$ correspond to mean consumption of energy of a client per hour during a day. The clients are distributed geographically throughout the country according to $\mathbf{z}$, that is, the longitude and latitude corresponding to their cities to show all clients in a map. Specifically, the objective of this problem is to partition Spain into regions and, inside each of them, determine consumption patterns. In each of these partitions, we adjust a model $\mathcal{M}$ through a clustering analysis by using $k-$means algorithm. The parameter k is determined by considering the value that minimizes *Bayesian Information Criterion* (BIC) for all customers from Spain. We obtain $k = 6$ and this value is applied to all regions. In addition, we carry out 5 $k-$means replicas in each region to get a good adjustment of the model.

The evaluation of the objective function (12) is expensive because it performs the $k-$means algorithm $3 * 5$ times over a large dataset of clients. Due to the conclusions from the first experiment, we have decided to use an algorithm based on response surfaces. In addition, as we can see in that experiment, the MSRS method spends less time than EGO. For these reasons, we have chosen MSRS. This method is set to 250 functional evaluations and the surrogate model is RBFcub. The results are shown in Fig. 4.

We can see how hyperplanes divide customers into three regions. Catalonia and Balearic Islands (red circles) are separated from the rest of Spain because they contain quite a few factories that consume a large amount of energy. On the other hand, Canary Islands, in other partition (red squares), is characterized by a lower consumption influenced by weather and a small number of companies. This conclusion is shown by Fig. 5, in which we can observe 6 groups of customers and, in each chart, the mean consumption of the three regions divided by hyperplanes.

3.3 Experiment 3: Forecasting of time series with spatial constraints. Application to electrical consumption in Spain

In many problems, the variable $\mathbf{z}$ is used as data label because it adds information when or where the data $\mathbf{x}$ was obtained. The objective of this experiment is to show that the methodology based on piecewise time series lets us improve the base model and achieve more precise results. Specifically, we deal with the

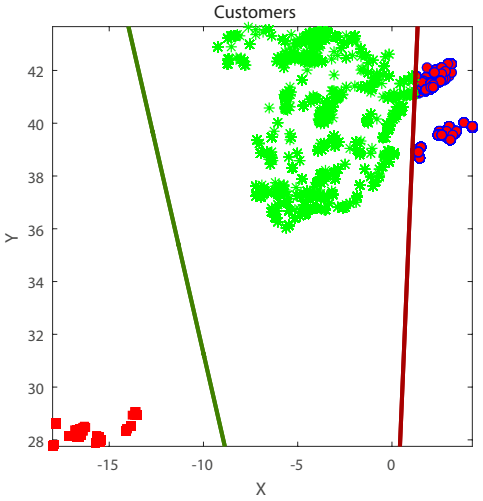


Fig. 4: Partition of Spain with respect to the electrical customers

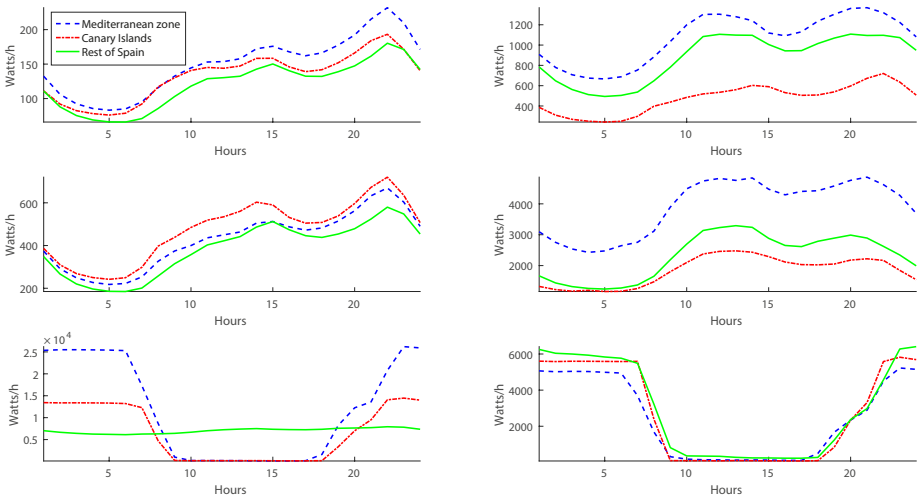


Fig. 5: Groups characterization

prediction of energy consumption of homes and companies in Spain. To carry out this computational experience, we suppose two situations: i) the information supplied by the location label **z** isn't taken into account and a regression model is adjusted, ii) we include location information of variable **z** to estimate a piecewise regression model. Formally, we consider:

– *Regression model.* The prediction is carried out by way of all records of consumption gathered throughout Spain. We consider this expression

$$y_{k+1} = \mathcal{M}(\overbrace{y_k, y_{k-1}, y_{k-2}, y_{k-3}, y_{k-4}}^{\mathbf{x}}) \tag{13}$$

where y_k is the total energy consumption at hour k.

– *Piecewise regression model* The country is divided into regions $\mathbf{R}_j$ to predict the user consumption in each of them.

$$y^j_{k+1} = \mathcal{M}_j(\overbrace{y^j_k, y^j_{k-1}, y^j_{k-2}, y^j_{k-3}, y^j_{k-4}}^{\mathbf{x}}) \tag{14}$$

where y^j_k is the total energy consumption in the area $\mathbf{R}_j$ at hour k. The national consumption is $y_k = \sum_j y^j_k$.

Firstly, we aggregate consumption per days during 2015. As it is necessary two set of data for the prediction, a random selection is made to decide which days are considered for the training of the model and which ones for the test. In this case, the sets are composed of 65 % and 35 % of total days registered in 2015[1], respectively.

With the aggregate consumption for each of the training days, a linear regression model is obtained to predict the consumption for the following hours. We use a *time window* of 5 hours for the training and the next hour is for the test. We don't estimate between 0 and 5 hours because this is the first period for the training.

In the prediction step, we obtain all estimated hours related to consumption at a national level. The indicator of quality of the adjusted regression model is the sum of squared errors. It is calculated by comparing the real consumption with the estimated one for each of days and, in turn, for each of hours.

Next, we fit a piecewise regression model to the previous problem. The measure related to the adjustment quality of the regression model in each region is obtained by way of the sum of squared error related to real and predicted energy consumption together with the penalization because of adding a new hyperplane. This function is optimized by performing the algorithm based on response surfaces, MSRS, with the surrogate model RBFcub through 270 main iterations. The regions obtained by the piecewise procedure to predict the energy consumed at the next hour are shown in Fig. 6.

Finally, we compare both models by considering three cases with predictions to one, two and three next hours of energy demand. The results described previously correspond to the next hour; however, we repeat the same procedure to estimate the two and three next hours. The table 2 shows the mean relative error (throughout the day) for both models. We can see the piecewise regression model achieves lower relative error than regression model in the three first cases.

[1] There are 277 days registered in the data-set: 180 days for the training and 97 for the test

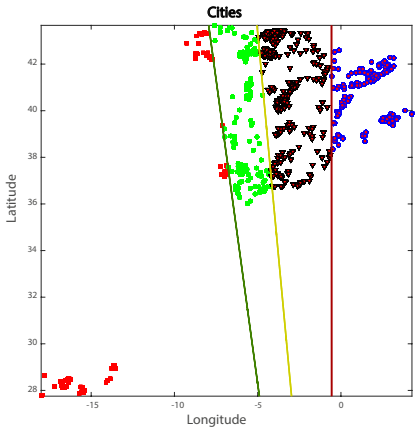


Fig. 6: Regions of Spain by using the piecewise model

Table 2: Mean relative error

Problem	Spain (%)	Piecewise (%)
1 H	0.1446	0.1193
2 H	0.2529	0.2037
3 H	0.3032	0.2346

4 Conclusions

In this paper, we present a methodology to include additional label variables in the analysis of time series. This methodology has been illustrated with a problem of clustering of time series for the establishment of patterns of electrical customers in different geographical locations and for the improvement of short-time forecasting of electrical consumption in which the geographical location of consumption is considered.

A future line of research will be to extend this scheme with hard partitions to fuzzy partitions. The level of membership obtained for each time series allows to establish its weight for analysis.

Acknowledgements

This work has been supported by Ministry of Economy and Competitiveness - FEDER EU with grant number TRA2016-76914-C3-2-P and the Predoctoral FPU fellowship from Ministry of Education, Culture and Sports with number 16/00792.

References

1. Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence 24*(1), 164–181.
2. García-Ródenas, R., M. López-García, and T. Sánchez-Rico (2017). An approach to dynamical classification of daily traffic patterns. *Computer-Aided Civil and Infrastructure Engineering 32*(3), 191–212.
3. López-García, M., R. García-Ródenas, and A. G. Gómez (2014). Hybrid metaheuristic optimization algorithms for time-domain-constrained data clustering. *Applied Soft Computing 23*, 319 – 332.
4. López-García, M. L., R. García-Ródenas, and A. G. Gómez (2015). K-means algorithms for functional data. *Neurocomputing 151*, 231–245.

Direct and Recursive Strategies for Multi-Step Ahead Wind Speed Forecasting

Sameer Al-Dahidi[1] and Hisham ElMoaqet[2]

[1]Department of Mechanical and Maintenance Engineering, School of Applied Technical Sciences, German Jordanian University, Madaba, Jordan

Sameer.Aldahidi@gju.edu.jo

[2]Department of Mechatronics Engineering, School of Applied Technical Sciences, German Jordanian University, Madaba, Jordan

Hisham.Elmoaqet@gju.edu.jo

Abstract. In this work, we investigate the capability of two different prediction strategies, i.e., direct and recursive, in accurately multi-step ahead forecasting the wind speed. For illustration purposes, the prediction strategies are implemented within the Artificial Neural Networks (*ANNs*) technique. To demonstrate our results, we use real wind speed dataset collected over a three-year period from the West Texas Mesonet, with 5-minute observation period from near Lake Alan, Garza. Three standard performance metrics, namely Mean Absolute Error (*MAE*), Root Mean Square Error (*RMSE*) and Symmetric Mean Absolute Percentage Error (*SMAPE*) are used for comparing the performances of the two strategies. Results show the effectiveness of the strategies for multi-step ahead wind speed forecasting compared to the classical persistence prediction strategy widely used in literature, in particular for longer time horizons.

Keywords: Multi-Step ahead; Wind speed; Forecasting; Direct; Recursive; Persistence; Artificial Neural Networks.

1 Introduction

With a diminishing pool of the world's fossil fuel supply (e.g., natural gas, oil, coal, etc.), energy security or access to cheap energy has become extremely important to the functioning of modern economies. Utilization of clean and renewable energy resources can help assuage this dependence on fossil fuels and can result in direct economic benefits, if implemented in an economic and scientific pragmatic manner. Among these, wind energy is one of the most widely tapped resources, and is a low cost and environment friendly resource, yet the stochastic and intermittent nature of wind constitutes a challenge in electricity generation (Ahmed & Khalid, 2017; Al-Dahidi, Baraldi, Zio, &

adfa, p. 1, 2011.

Legnani, 2017; Bokde, Troncoso, Asencio-Cortés, Kulat, & Martínez-Álvarez, 2017; The European Wind Energy Association (EWEA), 2017).

Electricity generation prediction from wind plants depends on weather data, such as wind speed. In practice, being able to effectively predict the electrical power output from wind plants with intermittent energy sources at time in the future is one of the milestones in ensuring the reliability of electric power distribution (Al-Dahidi et al., 2017; Bokde et al., 2017). For instance, inaccurate forecast of wind speed will indeed lead to a large electrical power output deviation, therefore accurate wind speed forecasting is of paramount importance for wind plants owners to ensure a sustainable contribution of their plants to the electricity grid, consequently, an affordable and reliable energy source to the consumers (Ahmed & Khalid, 2017; Al-Dahidi et al., 2017).

The wind speed forecasting approaches can be typically classified into model-based and data-driven approaches (Zhang, Su, Zhang, Zhao, & Zhao, 2014). Model-based approaches, such as numeric weather prediction (*NWP*) methods, require building mathematical/physics-based models of the atmosphere and oceans for weather/wind speed forecast (Liu, Shi, & Erdem, 2010). However, their practical deployment is in fact limited by the difficulty of building such models without making simplifications and assumptions, which reduce their predictability performances (Al-Dahidi et al., 2017). Contrary, data-driven approaches, such as, statistical, probabilistic and/or hybrid methods, based upon machine learning techniques (Zhang et al., 2014), do not use any explicit physics-based models, but completely depend on the availability of weather data (Al-Dahidi et al., 2017).

Over the last few decades, several data-driven methods such as Artificial Neural Networks (*ANNs*) (Ahmed & Khalid, 2017), Hidden Markov Models (*HMMs*) (Hocaoglu, Gerek, & Kurban, 2010), Recurrent Neural Networks (*RNNs*) (More & Deo, 2003), Support Vector Machines (*SVMs*) (M. A. Mohandes, Halawani, Rehman, & Hussain, 2004) and hybrid systems, have been developed and applied with success to wind speed forecasting. For example, (Niu, Wang, Zhang, & Du, 2018) have proposed a hybrid approach for multi-step-ahead wind speed forecasting by combining optimal feature selection technique and an *ANN* optimized by a modified bat algorithm with a cognition strategy.

In this regard, the objective of the present work is to employ the *ANN*, given their ability of successfully solving highly non-linear problems (Al-Dahidi et al., 2017; Han, Liu, & Yan, 2011), for multi-step ahead wind speed forecast. In particular, we aim to investigate the capability of two different prediction strategies, namely direct and recursive (Ben Taieb, Bontempi, Atiya, & Sorjamaa, 2012) in accurately forecasting the wind speed values and to compare their forecasting accuracy with respect to the well-known persistence strategy of literature, as a benchmark.

The comparison is performed considering wind speed time series data collected over a three years period from the West Texas Mesonet station with an observation period of

5 minutes. Three standard metrics, namely Mean Absolute Error (MAE), Root Mean Square Error ($RMSE$) and Symmetric Mean Absolute Percentage Error ($SMAPE$) (Ahmed & Khalid, 2017; Zhang et al., 2014) are used as performance metrics for comparing the different prediction strategies.

The remaining of this paper is organized as follows. Section 2 states the problem. In Section 3, the adopted prediction strategies for multi-step ahead forecasting of wind speed are briefly introduced. In Section 4, the results of the application of the two strategies to a real case study are presented and compared with those obtained by the persistence strategy of literature. Finally, some conclusions are drawn in Section 5.

2 Problem Statement

Let us assume that we have available N wind speed values, $\vec{v}$, observed for a period of time, Y, in a wind energy plant with a constant observation period, $\Delta t = t_i - t_{i-1}$, at predefined times $t_1, \dots, t_i, t_N$, $i = 1, \dots, N$, and recorded as a time series data of $\vec{v} = [v_1, \dots, v_i, v_N]$.

The objective is, then, to develop a prediction model, that utilizes historical wind speed values, $\vec{v}$, to provide H-step ahead predictions of the wind speed in the plant, $\hat{\vec{v}} = [v_{N+1}, \dots, v_{N+h}, v_{N+H}], h = 1, \dots, H$, where H is the forecast horizon, i.e., at a given time t_N, the objective is to provide the wind speed forecast for the following H time steps.

3 Methods

This Section briefly illustrates the basics of the two prediction strategies employed in this work for the multi-step ahead forecasting of wind speed. In particular, we consider the Recursive strategy (Subsection 3.1) and the Direct strategy (Subsection 3.2) (Ahmed & Khalid, 2017).

It is worth mentioning that the basic idea behind the different prediction strategies lies in the way they relate the wind speed forecasts, $\hat{\vec{v}}$, to the historical wind speed values, $\vec{v}$.

3.1 Recursive strategy

The idea underpinning this strategy is, first, to forecast one-step ahead at time step t_N, i.e., t_{N+1} that is at $h = 1$, by building a prediction model trained on the previous d wind speed values:

$$\hat{v}_{N+1} = f_1(v_N, \dots, v_{N-d+1}) \quad (1)$$

where d is the time series model order (embedding dimension). Afterwards, the forecasted wind speed values are fed as inputs to the input time series prediction model for forecasting successive time steps in the horizon, $h = 2, \dots, H$, using:

$$\hat{v}_{N+h} = f_h(\hat{v}_{N+1}, v_N, \dots, v_{N-d+h}) (2)$$

In this regard, the recursive strategy may lead to inaccurate forecasts for longer time horizons due to the forecast error accumulation that arises from each forecasted wind speed value (Ahmed & Khalid, 2017).

3.2 Direct strategy

The idea underpinning this strategy is to forecast each time step in the horizon, $h \in [1, H]$, independently from each other. Consequently, H prediction models, f_h, are in need to be built for each time horizon, h, $h = 1, \dots, H$. In this regard, the h-th forecast is performed by considering only the previous d wind speed values regadless the previous wind speed forecasts:

$$\hat{v}_{N+h} = f_h(v_N, \dots, v_{N-d+1}) (3)$$

Despite its simplicity in building the H prediction models and the fact that it avoids the forecast error accumulation as in the recursive strategy, this strategy requires more computational efforts as compared to the recursive strategy and may yield to uncorrelated results for multi-step ahead forecasting in some cases (Ahmed & Khalid, 2017).

For illustration purposes, the *ANNs* prediction technique is employed as the base prediction model, due to its simplicity and its potential in providing accurate wind speed forecasts compared to other techniques of literature, like for instance, Auto-regressive time series model (Mohamed A. Mohandes, Rehman, & Halawani, 1998).

4 Real Case Study: Results and Discussions

In this Section, we tackle the problem of multi-step ahead wind speed forecasting: Subsection 4.1 describes the wind speed data and presents the pre-processing steps carried out on the data for their proper utilization for the forecasting task, Subsection 4.2 illustrates the optimization of the *ANN* model structure for accurate one-step ahead wind speed forecasting and Subsection 4.3 discusses the results obtained by the application of the optimum *ANN* model using the different prediction strategies for multi-step ahead wind speed forecasting.

4.1 Data description and pre-processing

The wind speed data was collected over a $Y = 3$ years period from the West Texas Mesonet station (a network of meteorological monitoring instruments dispersed across West Texas), sampled at $\Delta t = 5$ minutes interval from near Lake Alan, Garza (Institute National Wind, 2018). The wind speed observations have been collected at an altitude of 10 m. The data consists of the average wind speed and direction with a peak instantaneous 3 seconds gust (during the 5 minutes observation period) (Institute National Wind, 2018). It is worth mentioning that only the average wind speed values are considered for building the prediction model.

Figure 1 shows an example of the observed wind speed values of one year. One can easily recognize the large variability of the wind speeds (i.e., highly nonstationary) at different times/seasons of the year.

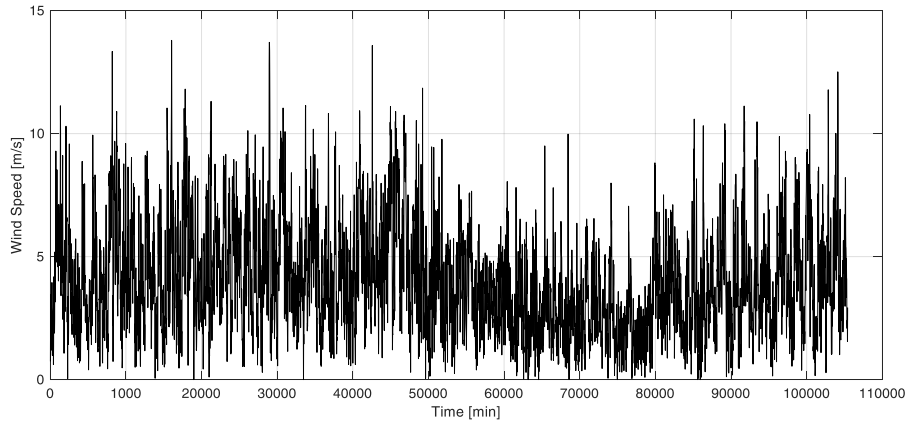


Fig. 1. An example of wind speed time series for one year.

For this reason, the data is corrected, i.e., it is convex smoothed, by using a filter of the form:

$$\bar{v}_i = \lambda v_i + (1 - \lambda)\bar{v}_{i-1} (4)$$

where λ is a smoothing filter taken equals to $\lambda = 0.4$ and $\bar{v}_i$ is the corrected wind speed value of the v_i measured at time instant t_i, $i = 1, \dots, N$. Figure 2 shows an example of the pre-processing performed on one particular day.

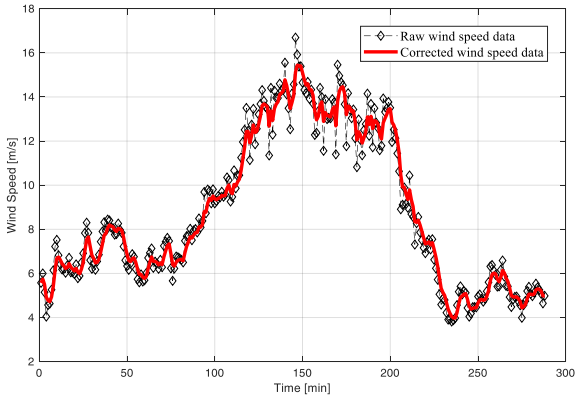


Fig. 2. An example of the corrected wind speed time series of one day.

The $Y = 3$ years available time series data are divided into:

1. **Training time series data** that contains a portion of 60% selected randomly from the first two years data and it is formed by $N_{train} = 126142$ time steps. It is used for building (training) the *ANN* base prediction model;
2. **Validation time series data** that contains a portion of 40% selected randomly from the first two years data and it is formed by $N_{valid} = 84094$ time steps. It is used to optimize the structure of the *ANN* prediction model in terms of number of hidden neurons, N_h, and the number of previous time steps to be used for the multi-step ahead forecast (embedding dimension), d;
3. **Test set time series data** that contains the third year and it is formed by $N_{test} = 105150$ time steps. It is used to evaluate the performances of the two different prediction strategies with respect to the persistence strategy.

4.2 Optimum *ANN* model structure

It is worth mentioning that *ANN* is assumed to be characterized by three layers: input, hidden and output layers. The *ANN* structure is optimized in term of *i)* number of hidden neurons, N_h, and *ii)* number of the previous time steps (embedding dimension), d. To this aim, we follow an exhaustive search procedure by considering different candidates for N_h, i.e., $N_h = 10,20, \dots ,100$, and d, i.e., $d = 1,2, \dots ,20$. The maximum candidate values of N_h(i.e., 100) and d (i.e., 20) are arbitrary; taking into account the computational efforts required for larger values of N_h and d. The performance of each *ANN* candidate structure, that comprises the combination of different values of N_h and d, is evaluated by computing the Root Mean Square Error ($RMSE$), the Mean Absolute Error (MAE) and the Symmetric Mean Absolute Percentage Error ($SMAPE$) as standard performance metrics widely used in literature (Ahmed & Khalid, 2017; Zhang et al., 2014):

$$RMSE = \sqrt{\frac{\sum_{k=1}^{M}(v_k - \hat{v}_k)^2}{M}} (5)$$

$$MAE = \frac{\sum_{k=1}^{M}|v_k - \widehat{v_k}|}{M} (6)$$

$$SMAPE = \frac{\sum_{k=1}^{M}\frac{|v_k - \hat{v}_k|}{(|v_k| + |\hat{v}_k|)}}{M} \%(7)$$

where, M is the overall number of the forecasts made, v_k and $\hat{v}_k$ are the k-th true and forecasted wind speed values, respectively. It is worth mentioning that smaller the values of these metrics, more accurate the forecast will be.

4.3 Application results

To ensure robust performance metrics, 5-fold Cross Validation (CV) is implemented in which the training and validation time steps are randomly selected from the first two years data with portions of 60% and 40%, respectively, for 5 times (i.e., 5 folds), and the average values of the performance metrics are then calculated.

The minimum *RMSE* value is found at the number of hidden neurons, $N_h^{opt} = 60$ with embedding dimension, $d^{opt} = 18$. Figure 3 shows the average values of the three performance metrics of the optimum number of hidden neurons, $N_h^{opt} = 60$ for different numbers of d that span the interval [1,20]: the star marker indicates the optimum number of the previous time steps, d. This is a good compromise between large number of hidden neurons and large embedding dimensions. In fact, one can recognize that for $d \geq 2$ (i.e., 10 minutes) the embedding dimension has a small influence on the accuracy of the forecasts. However, it is expected that for larger d values (i.e., in hours), the accuracy of the forecasts would be largely reduced, due to the high variability of the wind speed data measured during a long historical horizon.

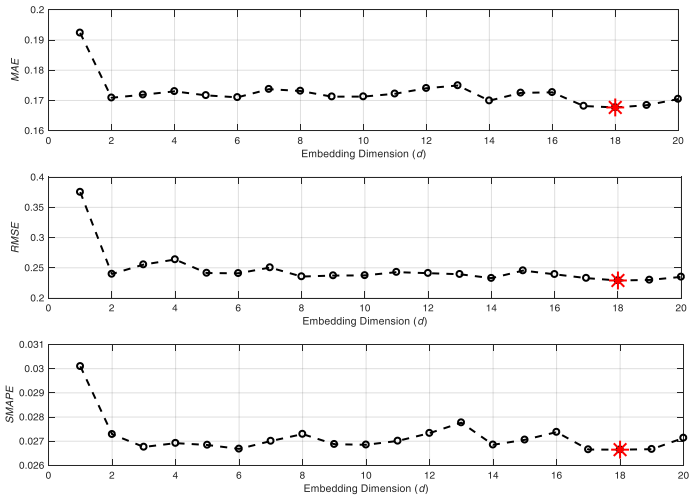


Fig. 3. MAE, $RMSE$ and $SMAPE$ values obtained by one-step ahead forecasting, considering the optimum number of hidden neurons, $N_h^{opt} = 60$, at different d values.

To effectively evaluate the performance of the adopted prediction strategies, their results are compared to those obtained by the application of a well-known prediction strategy (the persistence strategy) of literature to the same test data. In practice, the persistence strategy assumes that the wind speed forecast through the whole horizon will be the same as the v_N, $\hat{v}_{N+h} = v_N$, $h = 1, \dots, H$.

Figure 4 shows the average values of the three performance metrics of the recursive (circles) and the direct (squares) strategies obtained by the 5-fold cross validation procedure, with respect to those obtained by the benchmark persistence (diamonds) strategy, for a horizon of $H = 18$ steps (time in minutes) ahead. One can easily recognize the following:

- As expected, the forecasting errors increase significantly with the time horizon for the three prediction strategies;
- The recursive and direct strategies outperform the persistence, even for longer time horizons. This even is more clear when looking at the *RMSE* performance metric rather than the other metrics;
- The direct and recursive strategies have almost similar performance at the shorter time horizons, whereas the direct strategy starts to outperform the recursive strategy at the longer time horizons. This is due to the error accumulation in the recursive strategy as long as the time horizon increases, i.e., the recursive is used to rely on the previously obtained wind speed forecasts to provide the forecast of the following time step.

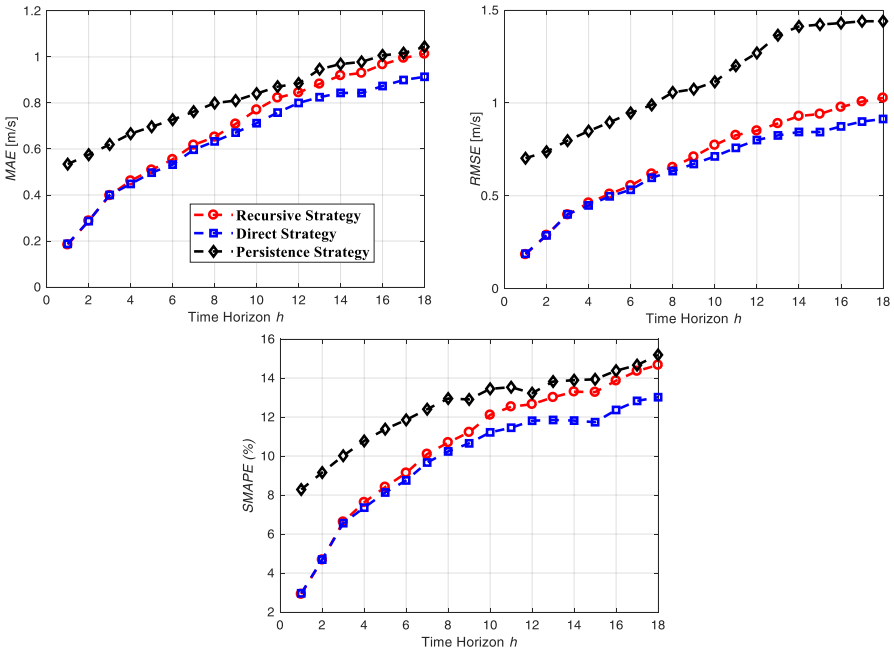


Fig. 4. MAE, $RMSE$ and $SMAPE$ values obtained by the recursive (circles), direct (squares) and persistence (diamonds) prediction strategies for 18-step ahead.

This is even more clear when looking at the standard deviation values calculated over the 5 CV iterations of the two strategies along the $H = 18$ horizon (Figure 5). One can recognize the error accumulation of the recursive strategy with respect to that of the direct strategy.

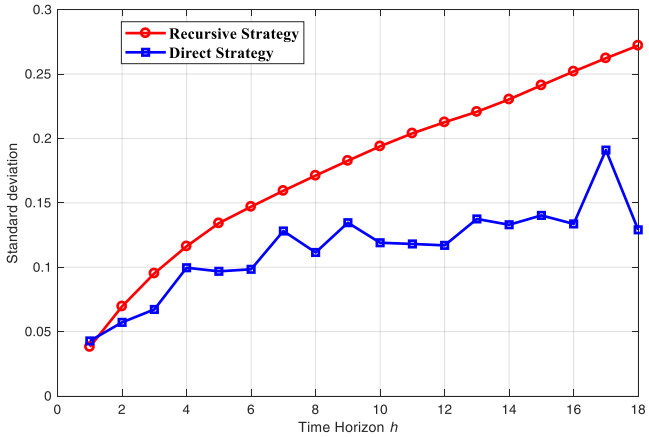


Fig. 5. Standard deviation values calculated over the 5-fold CV obtained by the recursive (circles) and direct (squares) prediction strategies for 18-step ahead.

5 Conclusions

The present work aims at investigating the capability of two different prediction strategies, i.e., recursive and direct, implemented on the basis of the Artificial Neural Networks (*ANNs*), in providing multi-step ahead wind speed forecasts. Results show the effectiveness of the two strategies in providing accurate forecasts, especially for longer time horizons, when compared to a benchmark model (persistence) by resorting to well-known performance metrics, i.e., *MAE*, *RMSE* and *SMAPE* calculated and interpreted statistically using a cross validation procedure. In this work, the obtained results are not necessary to be generalized to other types of wind speed datasets, whereas the investigated prediction strategies provide insights into the prediction performance that can significantly vary with different prediction strategies which need to be investigated on a case by case basis in multi-step ahead time series prediction problems.

Despite of the simplicity of the work carried out, the main objective is to establish a baseline for our future contributions in wind speed time series prediction. The development and the application of advanced computational methods for quantifying the uncertainties in forecasting processes, forecasting by resorting to ensemble of base prediction models, pre-processing the raw data for enhancing the forecast are few to mention for our future works for enhancing the accuracy of predicting near future wind speeds.

References

Ahmed, A., & Khalid, M. (2017). Multi-step Ahead Wind Forecasting Using Nonlinear Autoregressive Neural Networks. In *Energy Procedia* (Vol. 134, pp. 192–204). http://doi.org/10.1016/j.egypro.2017.09.609

Al-Dahidi, S., Baraldi, P., Zio, E., & Legnani, E. (2017). A Dynamic Weighting Ensemble Approach for Wind Energy Production Prediction. In *2017 2nd International Conference on System Reliability and Safety* (pp. 296–302). Milano. http://doi.org/10.1109/ICSRS.2017.8272838

Ben Taieb, S., Bontempi, G., Atiya, A. F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*. http://doi.org/10.1016/j.eswa.2012.01.039

Bokde, N., Troncoso, A., Asencio-Cortés, G., Kulat, K., & Martínez-Álvarez, F. (2017). Pattern sequence similarity based techniques for wind speed forecasting. In *Proceedings ITISE 2017. Granada, 18-20, September, 2017.ISBN: 978-84-17293-01-7* (pp. 786–794). Granada.

Han, S., Liu, Y., & Yan, J. (2011). Neural network ensemble method study for wind power prediction. In *2011 AsiaPacific Power and Energy Engineering Conference APPEEC 2011* (Vol. Wuhan). Retrieved from

http://www.scopus.com/inward/record.url?eid=2-s2.0-79955830495&partnerID=40&md5=c3ddc3d74feb8b4050bc5a4daf067614

Hocaoglu, F. O., Gerek, Ö. N., & Kurban, M. (2010). A novel wind speed modeling approach using atmospheric pressure observations and hidden Markov models. *Journal of Wind Engineering and Industrial Aerodynamics*, *98*(8–9), 472–481. http://doi.org/10.1016/j.jweia.2010.02.003

Institute National Wind. (2018). West Texas Mesonet. Retrieved from http://www.depts.ttu.edu/

Liu, H., Shi, J., & Erdem, E. (2010). Prediction of wind speed time series using modified Taylor Kriging method. *Energy*, *35*(12), 4870–4879. http://doi.org/10.1016/j.energy.2010.09.001

Mohandes, M. A., Halawani, T. O., Rehman, S., & Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, *29*(6), 939–947. http://doi.org/10.1016/j.renene.2003.11.009

Mohandes, M. A., Rehman, S., & Halawani, T. O. (1998). A neural networks approach for wind speed prediction. *Renewable Energy*, *13*(3), 345–354. http://doi.org/10.1016/S0960-1481(98)00001-9

More, A., & Deo, M. C. (2003). Forecasting wind with neural networks. *Marine Structures*, *16*(1), 35–49. http://doi.org/10.1016/S0951-8339(02)00053-9

Niu, T., Wang, J., Zhang, K., & Du, P. (2018). Multi-step-ahead wind speed forecasting based on optimal feature selection and a modified bat algorithm with the cognition strategy. *Renewable Energy*, *118*, 213–229. http://doi.org/10.1016/j.renene.2017.10.075

The European Wind Energy Association (EWEA). (2017). *Wind in power - 2016 European statistics*. Brussels. Retrieved from https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2016.pdf

Zhang, W., Su, Z., Zhang, H., Zhao, Y., & Zhao, Z. (2014). Hybrid wind speed forecasting model study based on SSA and intelligent optimized algorithm. *Abstract and Applied Analysis*, *2014*. http://doi.org/10.1155/2014/693205

Identification of multiregime periodic autoregressive models by genetic algorithms

Domenico Cucina, Manuel Rizzo, and Eugen Ursu

Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy
dcucina@unisa.it
Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00100 Rome, Italy
manuel.rizzo@uniroma1.it
GREThA UMR-CNRS 5113, Université de Bordeaux, Avenue Léon Duguit, 33608 Pessac cedex, France
eugen.ursu@u-bordeaux.fr

Abstract. This paper develops a procedure for identifying multiregime Periodic AutoRegressive (PAR) models. In each regime a possibly different PAR model is built, for which changes can be due to the seasonal means, the autocorrelation structure or the variances. Number and locations of changepoints which subdivide the time span are detected by means of Genetic Algorithms (GAs), that optimize an identification criterion. The method is evaluated by means of simulation studies, and is then employed to analyze shrimp fishery data.

Keywords: Seasonality, Structural changes, Genetic algorithm

1 Introduction

This paper is concerned with seasonal time series which may display many discontinuities, that can be specified by changepoints in time (or structural changes). As defined in [16] a changepoint is "a time where the structural pattern of a time series first shifts". In many cases, the changepoints are located at known times and it is easy to take into account their effects. When changepoints are located at unknown times and their features are ignored, the time series estimation can be misleading [20]. In fact, an undetected changepoint can lead to: misinterpretation of the model, biased estimates and less accurate forecasting [9]. Taking all these into account, changepoint detection becomes a demanding job especially if its identification is required soon after occurrence. In the past four decades several techniques have employed for changepoint detection [4, 6, 27]. For a recent review of changepoint analysis in time series see [1].

Periodic time series models have been introduced because standard seasonal autoregressive integrated moving average (SARIMA; [3]) cannot be filtered to achieve second-order stationarity, and this is because the correlation structure of these time series depends on the season [26]. [22] also showed that seasonal differencing maintains the seasonal correlation structure, whereas the periodic term is completely removed by seasonal standardization or by spectral analysis. General overviews of periodic models and their applications are presented in [10, 7].

We shall focus on time series recorded monthly which display periodic dynamics and possible structural changes, which may imply the existence of several regimes in time. Changepoint detection procedures for periodic series have been studied in [17], which focused on mean shifts. In our proposal we allow also the whole model structure to switch at each changepoint time, as far as periodic data usually display both seasonal effects and various kind of discontinuities.

We propose a procedure based on Genetic Algorithms to detect changepoints and estimate resulting PAR models. These kind of methods are well suited for complex global optimization, as they have been widely applied to hardly tractable identification and estimation problems [25, 6]. The procedure is based on an identification criterion, such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) or MDL (Minimum Description Length). The resulting method will also allow to perform subset selection, as we allow intermediate parameters to be constrained to zero. This modification leads to gain in parsimony, and could also contribute to improve forecasting ability of resulting models.

The article is organized as follows: Section 2 describes proposed methodology for model building; in order to illustrate the efficiency of the proposed procedure some simulations are presented in Section 3; an application to French Guiana shrimp fishery is included in Section 4; comments close the paper in Section 5.

2 Methodology

2.1 Model description

We consider a periodic time series of period s, observed for N years and possibly subdivided into M regimes in time. The multiregime PAR model is specified as follows:

$$X_{(n-1)s+k} = a^j + b^j[(n-1)s + k] + Y_{(n-1)s+k}, \tag{1}$$

where $j = 1, ..., M$ is the index of regimes, $k = 1, ..., s$ is the index of periods, a^j and b^j are trend parameters, which may vary with the regime, and μ_k^j are the seasonal means. $X_{(n-1)s+k}$ is referred to the observation in season k of year n ($n = 1, ..., N$), while $Y_{(n-1)s+k}$ follows a PAR given by:

$$Y_{(n-1)s+k} = \sum_{i=1}^{p} \phi_i^j(k) Y_{(n-1)s+k-i} + \epsilon_{(n-1)s+k}, \tag{2}$$

where $i = 1, \ldots, p$ denotes the lag, $\phi_i^j(k)$, $i = 1, ..., p$ indicate the autoregressive parameters of regime j and season k. As far as our identification procedure will allow to identify subset models, the autoregressive parameters can be constrained to zero, in order to get more parsimonious models. The error process is a periodic white noise with $E(\epsilon_{(n-1)s+k}) = 0$ and $\text{var}(\epsilon_{(n-1)s+k}) = \sigma_j^2(k)$, so that also the residual variances are allowed to change in each regime and season. We shall assume that each regime is periodic stationary with period s [18].

The regimes are specified by $M - 1$ changepoint years $\tau_1, ..., \tau_{M-1}$, defined in such a way that τ_{j-1} and $\tau_j - 1$ denote, respectively, the first and the last year of regime j ($j = 1, \ldots, M$). We also assume that $\tau_0 = 1$ and $\tau_M = N + 1$. In order to ensure reasonable estimates, we require that each regime contains at least a minimum number ω of years, therefore $\tau_j \geq \tau_{j-1} + \omega$ for any regime j. For the sake of simplicity we assume that the total number of observations T is a multiple of s ($T = N \times s$).

2.2 Model building

The identification of our multiregime PAR model consists in the choice of changepoints $M-1$, the changepoint times $\tau_1, \ldots, \tau_{M-1}$ and the specification of subset models (we shall assume the same maximum autoregressive order p for all models). The discrete search space is prohibitively large, so we shall base the procedure on GAs [11].

They are a nature-inspired optimization method, often employed when it is required to find an optimal solution from a prohibitively large discrete set. In GAs metaphor, the search strategy is based on the evolution of a population of individuals, coded in binary vectors named chromosomes to suitably represent the problem solutions, towards populations which are better able to adapt to the environment. The goodness of individuals in such populations is called fitness, and it is related to the objective function of the problem at hand. At each iteration (named generation in the GA terminology) the evolution takes place by means of three main operators: the selection, which chooses the individuals that will generate

the offspring; the crossover, that allows pairs of individuals to combine, producing possibly better solutions; mutation, which simulates the rare random changes happening in nature, and facilitates the exploration of the search space. Lastly, the elitist strategy ensures that the best solution is always retained in each generation of the algorithm (for an account on GA operators and strategies see [8]). The flow of generations generally stops when a prefixed criterion is met, for example the reaching of a fixed number of generations.

We shall optimize a fitness function based on an identification criterion, such as AIC, BIC, Hannan-Quinn, that combine a measure of goodness of fit and a penalization on the number of parameters. In particular, we will consider a criterion inspired by the Normalized Akaike's Information Criterion (NAIC), introduced in [24] for threshold models:

$$g = [\sum_{j=1}^{M}\sum_{k=1}^{s} n_{j,k}\log(\hat{\sigma}_j^2(k)) + IC\sum_{j=1}^{M}\sum_{k=1}^{s} P_{j,k}]/T, \tag{3}$$

where $\hat{\sigma}_j^2(k)$ is the model residual variance of series in regime j and season k, $n_{j,k}$ and $P_{j,k}$ are, respectively, sample size and number of parameters of regime j and season k, IC is the penalization term. The choice of IC specifies the magnitude of penalization on number of parameters: for example a value equal to 2 resembles the structure of an AIC, while $IC = ln(N)$ leads to the analogous to BIC criterion. The final fitness f will be a scaled exponential transformation of g: $f = \exp(-g/\beta)$, where β is a problem dependent constant. This is a quite common procedure in GAs [8, 13] as it allows to control the shape of fitness function without changing the solutions ranking.

The fitness evaluation step is carried out conditioning on the model structure of a generic solution, and the model parameters are estimated consequently. These latter are the trend intercepts and slopes a^j and b^j, the seasonal means μ_k^j, the autoregressive parameters $\phi_i^j(k)$ and the residual variances $\sigma_j^2(k)$, for all $j = 1, \ldots, M$; $k = 1, \ldots, s$; $i = 1, \ldots, p$. These parameters are estimated by Ordinary Least Squares, except for the autoregressive ones, which must account also for the subset selection constraints.

With respect to the model structure, we shall adopt the following strategy in our GA: the generic chromosome will binary encode only the regime structure $[M-1, \tau_1, ..., \tau_{M-1}]$. Conditioning on such structure, all the possible 2^p subset autoregressive models will be enumerated and evaluated in the fitness evaluation step, and only the best will be retained.

This is an exact strategy with respect to the subset selection, and it is computationally feasible only if the maximum autoregressive order p is small.

The chromosome encoding works as follows: the first two or three bits (depending on the maximum number of regimes allowed) give the number of changepoints $M-1$; subsequent bit intervals, whose length is custom fixed, produce changepoint times $\tau_1, ..., \tau_{M-1}$. This part of encoding must ensure the constraints on minimum number of years per regime ω. We shall adopt a procedure introduced in [2] for multiregime nonlinear models identification, which allows each chromosome to be legal (the constraints on minimum regime length ω always hold), so there is no computational time wasted on evaluating infeasible solutions.

3 Simulation studies

To illustrate the efficiency of the proposed procedure, we use a set of one thousand simulated series and apply our method to each series. The series will contain a century ($N = 100$) of monthly data ($s = 12$). We consider the following options for the objective function, in order to study the sensitivity of penalization: values of IC equal to 2 and $ln(N)$, which resemble a generalization of AIC and BIC criteria, and also $IC = 3$, successfully adopted in [2] for the identification of nonstationary nonlinear models by GAs. Concerning the choices on GA implementation, we employed a population of 50 solutions and used operators of tournament selection, bit-flip mutation (rate 0.1) and parameterized uniform crossover (rate 0.7). As far as GA as stochastic methods, namely each GA run may lead to a different results, we shall report the correct number of changepoints identification rate, and also the mean and standard deviation of changepoint locations.

Model 1: 1 changepoint with no trend Our first simulation model consists of 1 changepoint located at the end of the 60-th year ($\tau_1 = 61$). For the second regime, we consider that the variances are four times smaller than for the first regime. We consider the same autoregressive structure in both regimes and an order equal to one for the PAR model.

We apply our method to the realization in Figure 1. Table 1 reports empirical frequency distribution of the estimated changepoints. We see that BIC ($IC = \ln N$) has the best percentage rate (100%) of correct identification. The $IC = 3$ has a correct rate of identification of 73%; 19.5% of runs estimate 2 changepoints and 6.2% of runs estimate 3 changepoints.

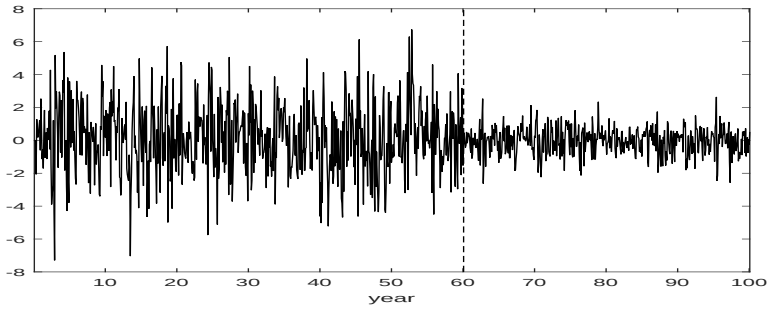


Fig. 1: A realization from the process in *Model 1*. The vertical dash line indicates the true changepoint location at the end of the 60-th year.

The AIC criterion seems to overestimate the number of changepoints as 84.4% of runs estimates 4 changepoints or more.

Number of	IC=ln N			IC=3			IC=2		
changepoints	%	mean	se	%	mean	se	%	mean	se
0									
1	100	60.97	0.68	73.0	60.96	0.59	0.1	24	0
2				19.5	50.40 70.34	15.29 9.05	4.3	48.76 69.86	16.43 9.62
3				6.2	50.90 67.11 81.73	14.50 9.88 8.18	11.3	48.27 65.18 79.92	17.45 12.12 9.73
≥ 4				1.3			84.4		

Table 1: Summary of the estimated changepoints for the *Model 1*. The true number of changepoints is 1.

As for where the changepoints are estimated, the mean and the standard-error of the estimated locations are also reported in Table 1. The proposed procedure performs very well in locating the changepoints for our method combined with BIC.

Model 2: 2 changepoints with different trends

In the second simulation experiment we consider a model with two changes in trend parameters at times $\tau_1 = 31$ and $\tau_2 = 61$. We use a PAR model of order 1 for each regime with same parameters from one regime

to other. For illustrative purposes, Figure 2 shows a typical realization of the model defined above.

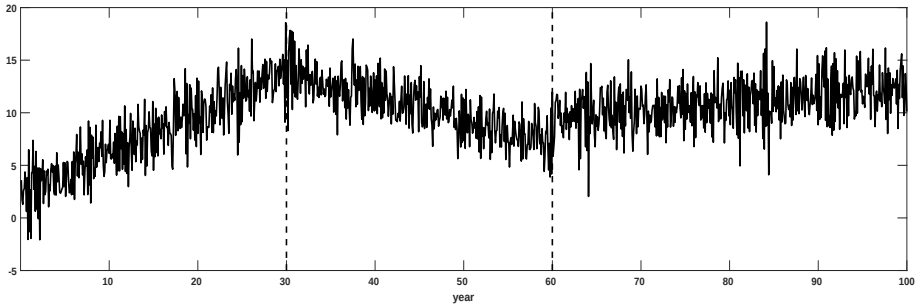


Fig. 2: A realization from the process of *Model 2*. The vertical dash line indicates the true changepoint locations at the end of 30-th and 60-th year.

For our method combined with BIC criterion, the selected number of changepoints is generally equal to 2, while only for 80 of the 1000 analyzed series 3 changepoints were selected and in 15 cases a the number of changepoints greater or equal to 4 was chosen. The $IC = 2$ and $IC = 3$ criteria seem to overestimate the number of changepoints. Lastly, in order to evaluate the changepoints location, we report the mean and the standard-error for estimated locations in Table 2.

Number of	IC=ln N			IC=3			IC=2		
changepoints	%	mean	se	%	mean	se	%	mean	se
0									
1	0.01	32.00	0.00						
2	90.4	30.94 60.97	1.48 0.45	38.1	30.86 60.99	1.43 0.14	1.4	31.00 61.00	1.17 0
3	8.0	27.51 51.05 71.40	6.33 12.58 10.42	22.7	28.07 50.19 70.01	6.23 11.95 9.96	4.1	25.58 44.51 65.68	7.64 12.07 8.45
≥ 4	1.5			26.6			94.5		

Table 2: Summary of the estimated changepoints for the *Model 2*. The true number of changepoints is 2.

4 Data analysis

We illustrate the main findings by analyzing the shrimp French Guiana fishery, a case study that has been accounted in [23]. Two shrimp species are mainly exploited in this fishery, the brown and the pink shrimps (respectively, Farfantepenaeus subtilis and Farfantepenaeus brasiliensis). The F. subtilis represents more than 85% of shrimp landings. We denoted by C the total catch of this shrimp in tons for the whole French Guiana fleet. This catch C is the product of the catchability coefficient q, the fishing effort measured by the number of days at sea E and the abundance of the fish population B. Based on the Schaeffer relation $C = qEB$, the catch-per-unit-effort (CPUE) is equal to the ratio C/E. CPUE is the catch extracted from one unit of fishing effort. We use the data collected by IFREMER (French institute of research for the exploitation of the sea) on C and E between January 1989 to December 2014 to get the CPUE. The data are represented in Figure 3.

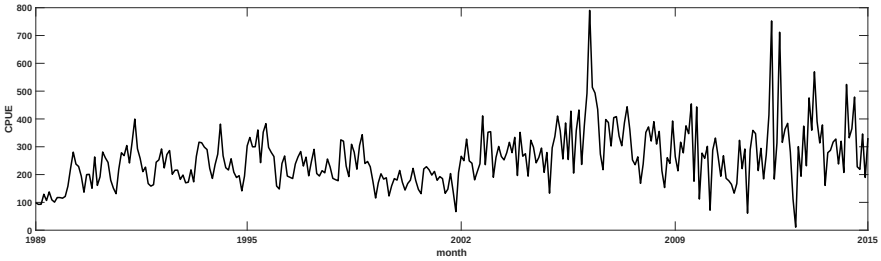


Fig. 3: Plot of monthly average of catch per unit effort between 1989 and 2014.

We will build various kind of PAR models using the BIC criterion, and also evaluate the forecasting accuracy, a standard one-step-ahead procedure, of resulting models by means of the measures Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) [12]. The first model considered in our analysis is a subset PAR without changepoints (denoted by *Model 1* in Table 3). We then estimated a PAR model with at least one changepoint: we impose an upper bound for the order of the PAR models on each regime equal to one. To avoid having too few observations in any regime we set a minimum span of $\omega = 10$ (*Model 2* in Table 3).

	Years of changepoint	$RMSE$	MAE	$MAPE$	$Fitness$
Model 1	/	108.19	85.58	29.05	0.413
Model 2	2002	89.29	67.98	22.53	0.426
Model 3	1996,2002	89.29	67.98	22.53	0.429

Table 3: Results on evaluation criteria of the forecast errors for CPUE

Using our model we found one changepoint corresponding to 2002 (Figure 4).

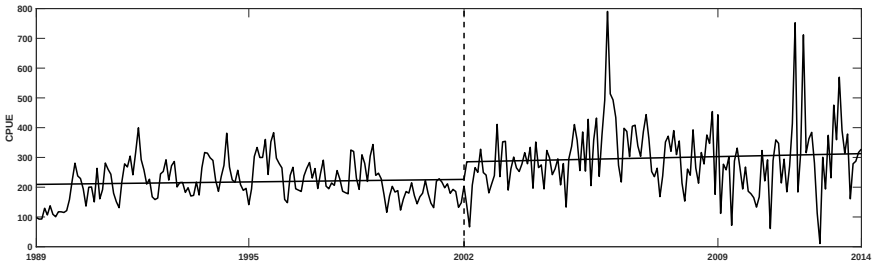


Fig. 4: Changepoint detected on years 2002 for CPUE

To ascertain the type of changes in the time series data due to changepoints, we calculate the 12 seasonal means and the 12 seasonal variation for all years up until the first changepoint. From the changepoint onwards we calculate the seasonal means or the seasonal standard deviations for each period until a detection of a new changepoint. To note that the model used by [17], designed to detect mean shifts, fails to identificate the changepoint. This could be explained by important changes in variance and not so important changes in mean (Table 4).

The changepoint corresponding to 2002 could be linked to the comments made in [15]: it reports a strong correlation between the Southern Oscillation Index (SOI) and the fish recruitment between $2002 - 2009$ ($R^2 = 0.81$), and a lack of correlation between $1990 - 2001$ ($R^2 = 0.001$). The El Niño and La Niña variables are captured through the SOI. The El Niño yields a disruption of temperature in the tropical Pacific Ocean that has important weather and climate consequences around the globe and are associated with physical and biological changes in our oceans that affect fish abundance and distribution. El Niño usually currents last for several months, resulting in the reduction of nutrients and a corresponding dissi-

Month	percentage change in mean	percentage change in variance
Jan.	16.69	107.53
Feb.	41.14	-63.80
Mar.	32.43	365.05
Apr.	69.86	1028.67
May	54.39	270.73
June	71.16	96.70
July	69.15	251.81
Aug.	16.63	78.41
Sep.	2.55	180.97
Oct.	26.87	-9.12
Nov.	22.78	140.97
Dec.	54.87	171.18

Table 4: Percentage change in mean and variance before and after 2002 of CPUE

pation of fish stocks. The La Niña is opposite for this other phase of the SOI, when sea surface temperatures in the central and eastern tropical Pacific are unusually low and when the trade winds are very intense. The sea surface temperature (SST) is a good indicator of global warming due to greenhouse gases. A change in temperature could have an impact on the movement of shrimp populations, on their rate of growth and/or mortality ([21]). The recruit abundance as well as the stock biomass and the fishing mortality were monthly performed by virtual population analysis (VPA) calculations.

If we set the minimum span $\omega = 6$ and we use our model combined with BIC we found two changepoints corresponding to 1996 and 2002 (Figure 5). The evolution of CPUE shows an increase from 1990 to 1996, followed by a decrease until 2002. The changepoint corresponding to 1996 could be linked to the evolution of biomass which has been decreasing steadily over time since 1996 followed by an improvement between 2003 and 2005, but without affecting the overall trend of decline ([14]). Fishing effort was concentrated in the shallow waters until 1995, for which the biomass was highest. Moreover, the French Guiana marine fishing area might be affected by changes in the SST since the latter significantly increased between 1970 and 2004, with an accentuation of this phenomenon since 1995 ([5]). These results, however, should be interpreted with caution as the number of observations is very important for good estimation results.

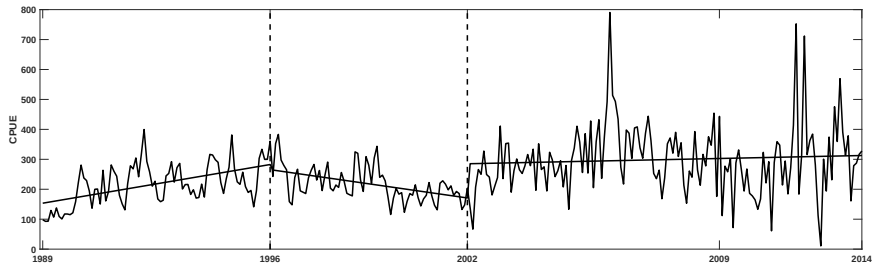


Fig. 5: Changepoint detected on years 1996 and 2002 for CPUE

5 Conclusions

The goal of our research was to develop a computational procedure for building multiregime models in time series with a periodic behaviour. Our procedure has been found effective both in simulation studies and in analysis of time series related to catch per unit effort of shrimps in French Guyana. The reasons for such changepoints are possibly due to both human activities and climatic oscillations. It is hoped that the results presented in this article will be useful in hydrology and finance, where interest lies in detecting changes in the volatility of time series due to changes in instrumentation and institutional changes, respectively.

Acknowledgments. The authors thank professor Francesco Battaglia for his valuable and constructive remarks. Part of this work has been carried out with the financial support of the French National Research Agency in the frame of the Investments for the future Programme, within the Cluster of Excellence COTE (ANR-10-LABX-45).

References

1. Aue A, Horváth L (2013) Structural breaks in time series. Journal of Time Series Analysis 34:1–16
2. Battaglia F, Protopapas MK (2012) Multi-regime models for nonlinear nonstationary time series. Computational Statistics 27:319–341
3. Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden-Day, San Francisco, CA
4. Davis RA, Lee TCM, Rodriguez-Yam GA (2006) Structural break estimation for nonstationary time series models. Journal of the American Statistical Association 473:223–239
5. Diop B, Sanz N, Duplan Y.J.J., Guene E.H.M., Blanchard F, Pereau J-C,Doyen L (2018) MEY fishery management facing climate warming. Ecological Economics:

6. Doerr B, Fischer P, Hilbert A, Witt C (2017) Detecting structural breaks in time series via genetic algorithms. Soft Computing 21(16):4707–4720
7. Franses PH, Paap R (2004) Periodic time series models. Oxford University Press
8. Goldberg DE (1989) Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley
9. Hansen BE (2001) The new econometrics of structural change: dating breaks in U.S. labor productivity. Journal of Economic Perspectives 15:117–128
10. Hipel KW, McLeod AI (1994) Time series modelling of water resources and environmental systems. Elsevier, Amsterdam
11. Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press
12. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22:679–688
13. Kreinovich V, Quintana C, Fuentes O (1993) Genetic algorithms: what fitness scaling is optimal? Cybernetics and Systems 24(1):9–26
14. Lampert, L (2011) Etude de la crise de la pêche de la crevette en Guyane, VOL.1. IFREMER:1–79
15. Lampert, L (2013) Etude de la crise de la pêche de la crevette en Guyane, VOL.2. IFREMER:1–56
16. Li S, Lund R (2012) Multiple changepoint detection via genetic algorithms. Journal of Climate 25:674–686
17. Lu Q, Lund R, Lee TCM (2010) An MDL approach to the climate segmentation problem. The Annals of Applied Statistics 4:299–319
18. Lund RB, Basawa IV (1999) Modeling and inference for periodically corelated time series. In: Gosh S (ed) Asymptotics, Nonparametrics and Time Series, Marcel Dekker, New York, Statistics : textbooks and monographs, vol 158, pp 37–62
19. Lund RB, Basawa IV (2000) Recursive prediction and likelihood evaluation for periodic ARMA models. Journal of Time Series Analysis 21:75–93
20. Lund RB, Wang XL, Lu Q, Reeves J, Gallagher C, Feng Y (2007) Changepoint detection in periodic and autocorrelated time series. Journal of Climate 20:5178–5190
21. Magraoui A, Baulier L, Blanchard F (2014) Effet du changement climatique sur le stock guyanais de crevettes pénéïdes. IFREMER:1–31
22. Moeeni H, Bonakdari H, Fatemi SE (2017) Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. Journal of Hydrology 547:348–364
23. Sanz N, Diop B, Blanchard F, Lampert L (2016) On the influence of environmental factors on harvest: the French Guiana shrimp fishery paradox. Environmental Economics and Policy Studies pp Online First Articles, DOI 10.1007/s10,018–016–0153–6
24. Tong H (1990) Non-linear time series: a dynamical system approach. Oxford University Press
25. Ursu E, Pereau JC (2015) Application of periodic autoregressive process to the modeling of the Garonne river flows. Stochastic Environmental Research and Risk Assessment 30(7):1785–1795
26. Vecchia AV (1985) Periodic autoregressive-moving average (PARMA) modeling with applications to water resources. Water Resources Bulletin 21:721–730
27. Yau CY, Tang CM, Lee TCM (2015) Estimation of multiple-regime threshold autoregressive models with structural breaks. Journal of the American Statistical Association 511:1175–1186

Change Detection for Streaming Data using Wavelet-based Least Squares Density Difference

Nenad Mijatovic*, Rana Haber, Mark Moyou,
Anthony O. Smith, and Adrian M. Peter

Florida Institute of Technology, Melbourne, FL, 32901, USA
nmijatov2005@my.fit.edu

Abstract. In this paper, we present a novel algorithm for detecting change points in a continuous time-series stream, based on the ℓ_2 distance between two distributions. The distributions are non-parametrically modeled using wavelet expansions, inspiring the name of our method: *Wavelet-based Least Squares Density-Difference* (WLSDD). Using the least squares method, we show that the ℓ_2 distance between two wavelet expanded densities results in a closed-form expression in their coefficients. This circumvents the need to evaluate the densities and, instead, allows us to work directly with the differences between their coefficients. The method demonstrates superior change detection performance on both synthetic and real data sets, stationary or non-stationary, in comparison to other competing techniques.

Keywords: Non-parametric Density Difference Estimation, Wavelets, Change Point Detection, Data Stream Processing

1 Introduction

Recently, there has been an increased demand for algorithms capable of processing vast amounts of data during transfer from their origin to the central data lake (e.g., relational database, NoSQL database, or Hadoop HDFS). There are many terms related to these types of connected networks—the most popular one being the Internet of Things (IoT). This data transfer and analysis process is commonly referred to as *stream processing*. In [10], the authors established 8 basic rules for real-time streaming data, some of which motivated this research: rule 3 requires that the system has a built-in mechanism to handle imperfections, and rule 8 requires the system to process and respond to data changes instantaneously.

In this work, we propose the *Wavelet-based Least Squares Density-Difference* (WLSDD) framework that is able to achieve these desired objectives and identify reliable change points in stationary and non-stationary data streams in an efficient manner. The WLSDD method non-parametrically models density-difference between distributions at two time instances through wavelet expansion

* The authors acknowledge partial support from NSF grants Nos. 1263011 and 1560345. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. Also, we would like to thank Antonina "Nina" Mijatovic, Brian S. Smith, and Dr. Song Qin for their kind review of this work.

of their densities. Then, based on the algebraic property that the ℓ_2 distance between two such densities is available in closed-form, we formulate a change detection algorithm that is triggered on significant deviations from this metric. Numerous experimental evaluations illustrate the superior performance of WLSDD when compared to other contemporary methods.

1.1 Related Work

A recent comprehensive survey of the change point methods in [1] serves as a good summary of the current state-of-the-art algorithms. Here we provide a brief taxonomic summary, starting with non-parametric methods of change point detection.

An early work, described in [4], proposed change point identification in stationary time series using Kolmogorov-Smirnov tests. A common framework, capable of detecting both outliers and change points from non-stationary streaming data, was detailed in [19]. As a probabilistic approach used for time-series data, it re-learns the parameters of the auto-regressive (AR) model by forgetting past data (exponential decay). With each new data point, the framework incrementally learns a probability density function and assigns a score (calculated based on the probability) with respect to the baseline. Any rapid excursions from the scores indicate change points. The authors in [14] suggested the generative additive model with two main components: a piecewise linear trend and seasonality. The breaks in the trend component are contributed to the change points. Hence, this model was called the Breaks For Additive Seasonal and Trend (BFAST) and was applied to the sensed image time series data in order to detect phenological changes caused by seasonal climatic variations, see [15]. A similar additive model, which adds a holiday component to accommodate for human behavior in social networks, was suggested in [13]. The model was termed prophet, and it provided forecast-at-scale with the capability to detect changes in the streaming data.

Kifer *et al.* [7] developed a method for detecting and estimating changes in the streaming data. The method utilized a two-window paradigm: the data from the "reference" window were compared to a set from the current window. Aside from the non-parametric assumption, they also incorporated two more requirements: a few false positives and a user-understandable interpretation of the change. To compare and interpret the change, the authors proposed a distance called the relativized discrepancy that takes the relative magnitude of a change into account.

In many real situations, the direct estimation of probability densities may not be an easy task. To help with this, the authors in [6] worked on estimating the logarithm of the likelihood ratio (importance) of probability densities. The improved Kullback-Leibler Importance Estimation Procedure (KLIEP) was implemented for online applications, see [12]. The authors presented a likelihood ratio as the linear model for basis functions. The model parameters were determined by solving the convex optimization problem derived from the Kullback-Leibler divergence criteria. For basis functions, the authors used the Gaussian kernel with a spread parameter that was determined from the cross-validation

procedure. The online KLIEP algorithm updated the model parameters using the aging parameter for older samples. Yet another approach that calculated differences between densities, without explicitly calculating individual densities, was presented in [11]. The authors proposed an ℓ_2 distance, calculated by applying the least-squares fit to a density-difference model. They applied their model to the change point detection by thresholding the density-difference distance between kernel density estimators (KDE). Their work was termed Least-Squares Density-Difference (LSDD) and serves as the progenitor of our work.

1.2 Wavelet-based Density as Orthogonal Series Estimators for Batch Processing

Our wavelet-based density estimator belongs in the category of non-parametric orthogonal series estimators. Chentsov, in his seminal work [2], provided a theory that estimates probability densities of a given data set by using a series of orthogonal bases, $\{\varphi_j\}_{j\in\mathbb{Z}}$. The author applied his theory on an unknown squared integrable density function in $L^2(\mathbb{R})$ that is continuous on the given interval, with limited amplitude, variance, and energy value. Formally, a probability density function is estimated as

$$\hat{p}(x) = \sum_j \hat{b}_j \varphi_j(x), \tag{1}$$

where the coefficient $\hat{b}_j$, related to the jth orthogonal function $\varphi_j(x)$, can be then calculated using

$$\hat{b}_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(x_i), \tag{2}$$

as soon as the batch of data (of size n), $\{x_i\}_{i=1}^n$, is available. Observing the probability density function as functional data and applying Chentsov's eqs. (1) and (2), we can decompose it in wavelet representation, written as

$$\hat{p}(x) = \sum_l \hat{\alpha}_{j_0,l}\phi_{j_0,l}(x) + \sum_{j=j_0}^{J}\sum_l \hat{\beta}_{j,l}\psi_{j,l}(x), \tag{3}$$

where $\phi_{j_0,l}(x) = 2^{j_0/2}\phi\left(2^{j_0}x - l\right)$ is the scaling function (a.k.a. father wavelet), $\psi_{j,l}(x) = 2^{j/2}\psi\left(2^j x - l\right)$ is the wavelet function (a.k.a. mother wavelet), l is the translation parameter, j_0 represents the starting resolution, and J is the ending resolution for the wavelet basis (see [3] for more details). Assuming that we are provided with random data samples, we can follow eq. (2) to approximate the scaling and wavelet coefficients, as

$$\hat{\alpha}_{j_0,l} = \frac{1}{n}\sum_{i=1}^{n}\phi_{j_0,l}(x_i) \text{ and } \hat{\beta}_{j,l} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,l}(x_i), \tag{4}$$

and then calculate the estimated probability density function, using eq. (3). We chose to work with orthogonal, compactly supported wavelet bases (such as the Haar, Daubechies, Symlets, and Coiflets basis families). An excellent overview of algorithms used to estimate probability density functions can be found in [16]. However, a majority of algorithms assume that all data is available, causing them to regard estimation as a batch process.

1.3 Wavelet-based Density as Orthogonal Series Estimators for Stream Processing

In the case of streaming data, eq. (4) is still applicable. However, if the current number of samples, n, becomes too large, the new data will have no effect on the coefficients. To avoid this issue, Wegman and Caudle [18] suggested applying an exponential smoothing technique on the latest data. This approach forces older coefficients, estimated using elderly data samples, to vanish because of the aging or weighting parameter that is related to a window size N.

In [5], the authors recognized the limitations of the exponential aging approach as an inability to adapt to moderate changes in the streaming data. To account for this constraint, they suggested to evaluate only the relevant scaling and wavelet coefficients impacted by the new data. For each new sample, they perform evaluation $2^j + 2P$ times (P is wavelet family order) when calculating the estimated scaling or wavelet coefficient, using eq. (4). Formally, their model for updating coefficients can be expressed as

$$\hat{b}_{j,l}^{n+1} = \hat{b}_{j,l}^{n} + \hat{b}_{j,l}^{n+1,addition} - \hat{b}_{j,l}^{n+1,discounting}, \tag{5}$$

where the estimator coefficient is only updated for the new data within the basis support range, and left unchanged otherwise,

$$\hat{b}_{j,l}^{n+1,addition} = \begin{cases} \frac{\varphi_{j,l}(x_{n+1})}{N} & \text{if } 2^j l \leq x_{n+1} \leq 2^j\,(l+2P-1), \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$\hat{b}_{j,l}^{n+1,discounting} = \begin{cases} \frac{\varphi_{j,l}(x_{n-N+1})}{N} & \text{if } 2^j l \leq x_{n-N+1} \leq 2^j\,(l+2P-1). \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

In eqs. (6) and (7), the basis function $\varphi_{j,l}$ is either the scaling or the wavelet basis for the given resolution j and translation l. Recall that the addition term is applicable to the new data, x_{n+1}, providing it is within the support range of the lth translation. Similarly, aging is applied to the last data sample in the sliding window, x_{n-N+1}, again, providing it belongs to the support of the jth basis for the lth translation. The main challenge of this approach is to select an appropriate sliding window size, N, which depends on the stochastic properties of both the data at hand and the application.

2 Density-Difference using Wavelets

In this section, we provide theoretical and algorithmic approaches for our density-difference estimation method, giving rise to a change point application. We assume that the data sets, $\{x_{1i}\}_{i=1}^{N}$ and $\{x_{2i}\}_{i=1}^{N}$ (N is data set size or corresponding window size), are drawn from two different distributions $p_1(x)$ and $p_2(x)$. We seek a measure that estimates the difference between distributions for the provided data sets.

2.1 Direct Density-Difference Estimator based on Wavelets

Similar to [11], we want to estimate the true density-difference function, $f(x) = p_1(x) - p_2(x)$, between the distributions $p_1(x)$ and $p_2(x)$, by solving

$$\underset{g}{\text{argmin}} = \int \left(g\left(x\right) - f\left(x\right)\right)^2 dx, \tag{8}$$

where $g\left(x\right)$ presents the estimated density-difference function. For $g\left(x\right)$, we assume the wavelet expansion

$$g\left(x\right) = \sum_{l} \alpha_{j_0,l}\phi_{j_0,l} + \sum_{j=j_0}^{J}\sum_{l} \beta_{j,l}\psi_{j,l}, \tag{9}$$

which can be conveniently presented in the matrix form, as

$$g\left(x\right) = \phi\alpha + \sum_{j=j_0}^{J} \psi_j\beta_j,$$

where α and β_j are the vector representations of the scaling coefficients and the wavelet coefficients at j, respectively, are defined as

$$\alpha = \left[\alpha_{j_0,1} \cdots \alpha_{j_0,L}\right]^T \text{ and } \beta_j = \left[\beta_{j,1} \cdots \beta_{j,L_j}\right]^T,$$

where L is the number of scaling translates and L_j is the number of wavelet translates. The matrix representations of the scaling matrix, ϕ, and wavelet matrix, ψ_j, are defined as

$$\phi = \begin{bmatrix} \phi_{j_0,1}\left(x_1\right) & \ldots & \phi_{j_0,L}\left(x_1\right) \\ \vdots & & \\ \phi_{j_0,1}\left(x_N\right) & \ldots & \phi_{j_0,L}\left(x_N\right) \end{bmatrix} \text{ and } \psi_j = \begin{bmatrix} \psi_{j,1}\left(x_1\right) & \ldots & \psi_{j,L_j}\left(x_1\right) \\ \vdots & & \\ \psi_{j,1}\left(x_N\right) & \ldots & \psi_{j,L_j}\left(x_N\right) \end{bmatrix}.$$

Using the above matrix representation and with some derivations, eq. (8) can be rewritten as

$$\alpha^*, \beta_j^* = \underset{\alpha}{\text{argmin}} \left(\alpha^T\alpha - 2\alpha^T\hat{h}_\phi\right) + \underset{\beta_j}{\text{argmin}} \left(\sum_{j=j_0}^{J} \beta_j^T\beta_j - 2\beta_j^T\hat{h}_{\psi_j}\right), \tag{10}$$

where $\hat{h}_\phi$ and $\hat{h}_\phi$ are the evaluated expected values of the scaling and wavelet functions for the distributions at all points $\{x_{1i}\}_{i=1}^N$ and $\{x_{2i}\}_{i=1}^N$ in the domain, which simplifies to

$$\hat{h}_\phi \approx \frac{1}{N}\sum_{i=1}^{N} \phi\left(x_{1i}\right) - \frac{1}{N}\sum_{i=1}^{N} \phi\left(x_{2i}\right) \text{ and } \hat{h}_{\psi_j} \approx \frac{1}{N}\sum_{i=1}^{N} \psi_j\left(x_{1i}\right) - \frac{1}{N}\sum_{i=1}^{N} \psi_j\left(x_{2i}\right).$$

Eq. (10) can be solved in the closed form after applying the least squares optimization method, as

$$\alpha^* = \hat{h}_\phi \text{ and } \beta_j^* = \hat{h}_{\psi_j}.$$

Therefore, the estimated density-difference, $\hat{g}\left(x\right)$, can be computed using

$$\hat{g}\left(x\right) = \phi\hat{h}_\phi + \sum_{j=j_0}^{J} \psi_j\hat{h}_{\psi_j}, \tag{11}$$

which we commonly refer to as the *Wavelet-based Least Squares Density Difference* (WLSDD). Finally, the ℓ_2-based density-difference distance, using eq. (11), is estimated in the following closed form

$$d_{WLSDD} = \hat{h}_\phi^T\hat{h}_\phi + \sum_{j=j_0}^{J} \hat{h}_{\psi_j}^T\hat{h}_{\psi_j}. \tag{12}$$

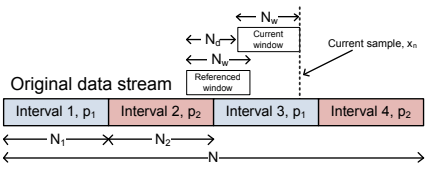


Fig. 1: Change point window approach with data stream samples from two alternating distributions.

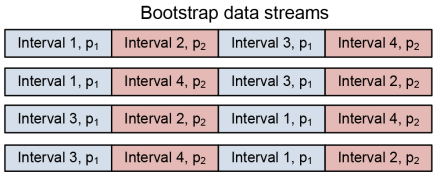


Fig. 2: Bootstrap approach with an optimal threshold for change detection (permutations on interval level).

2.2 Change Point Detection using WLSDD

Our proposed WLSDD algorithm, in order to detect change points in the streaming data, uses a two-window approach. This strategy, as illustrated in Fig. 1, is organized as follows: the reference window contains N_w samples of the stream from the past, and the current window contains the last N_w samples, including the current, last seen sample, x_n. We introduce an offset, N_d, between the current and reference windows (where $N_d \leq N_w$).

For each new sample in each window, we analyze the scaling and wavelet coefficients in an efficient manner. We begin with adding new scaling and wavelet coefficients using eq. (6) and removing old coefficients using eq. (7); then, we update all coefficients by applying eq. (5). After this update, we estimate one density-difference between the reference and current windows using eq. (12).

Finally, we compare the WLSDD value against the threshold value. If WLSDD exceeds the threshold, the algorithm reports a change. At the very beginning (with the first $N_d + N_w$ samples), the algorithm is trained to learn the initial density-differences between the reference and current windows. Thereafter, upon change detection, the algorithm continuously adapts by re-learning the new density-differences without performing any change thresholding until the next $N_d + N_w$ samples pass.

3 Results

In this section, we begin with an introduction of other density-difference measures and then we display the results obtained from synthetic and real data.

3.1 Some Other Distance Measures

In order to assess the performance of our WLSDD method, we compared it with LSDD, as well as Kolmogorov-Smirnov (KS) test, log-likelihood, entropy, and Kullback-Leibler divergence. For the last 3 methods, we assumed two different approaches for estimating distributions: (1) the underlying Gaussian distributions for log-likelihood (LL), entropy (E), and Kullback-Leibler (KL) measures and (2) histograms for log-likelihood (LLH), entropy (EH), and Kullback-Leibler (KLH) measures, see [9].

Assuming the reference p_r and current p_c distributions, the log-likelihood, entropy, and Kullback-Leibler divergence measures are defined as

Table 1: Density-difference (DD) between static data from two distributions. Overall, WLSDD estimates true DD more accurately when compared to LSDD.

id	**Distribution** p_1	**Distribution** p_2	**True DD**	**LSDD**	**WLSDD**	**Wavelet**	j_0
	Gaussian	Gaussian	0.000	**0.000**	0.011	db8	1
	Gaussian	Skewed Gaussian	0.125	**0.139**	0.140	db8	1
1	Asym. Claw	Asym. Dbl. Claw	0.059	0.064	**0.063**	coif1	1
2	Sep. Bimodal	Str. Skewed Uni.	0.590	0.560	**0.591**	sym7	3
3	Bimodal	Trimodal	0.007	0.005	**0.007**	coif3	1
4	Bimodal	Skewed Bimodal	0.037	0.052	**0.050**	db5	1
5	Dbl. Claw	Claw	0.106	0.093	**0.109**	sym10	2
6	Gaussian	Kurtotic Uni.	0.257	0.209	**0.234**	coif2	2
7	Dbl. Claw	Asym. Dbl. Claw	0.004	0.000	**0.004**	sym10	3
8	Skewed Uni.	Str. Skewed Uni.	0.838	0.810	**0.841**	sym7	3
9	Bimodal	Sep. Bimodal	0.111	0.118	**0.117**	coif3	1
10	Skewed Bimodal	Sep. Bimodal	0.162	0.183	**0.179**	db5	0

$$LL = \sum_i \log_2 \frac{p_{ci}}{p_{ri}}, \tag{13}$$

$$E = -\sum_i (p_{ri} - p_{ci}) \log_2 (p_{ri} - p_{ci}), \text{ and} \tag{14}$$

$$KL = \sum_i p_{ci} \log_2 \frac{p_{ci}}{p_{ri}}. \tag{15}$$

In eqs. (13), (14), and (15), index i represents either the current sample in the case of the LL, E, and KL methods (Gaussian assumptions) or the corresponding histogram bin in the case of the LLH, EH, and KLH methods. The same two-window approach, as introduced in § 2.2, is used to calculate other density-difference measures as well.

3.1.1 Synthetic Data Set - Density-Difference between Static Data First, to assess the performance of our WLSDD method, we observed the density-difference value obtained by comparing the samples from two different distributions in the offline setup. We generated samples drawn from different densities constructed using a mixture of Gaussians, as presented in [17]. From each distribution, we drew 2,000 samples, and calculated the density-differences using the WLSDD and LSDD methods, as suggested in [11]. In Tbl. 1, we compared the density-difference (DD) values between the two methods. The table contains the optimal wavelet family, as well as the starting resolution j_0 for WLSDD, using the cross-validation method.

Overall, our method expresses more accurate results when compared to the true density-difference in the LSDD results. These are the expected results, since the wavelet-based density estimators perform better compared to the kernel-based ones (refer to [8] for more details).

3.1.2 Synthetic Data Set - Alternating Distributions In order to evaluate the change point detection performance of our method in a streaming setup, we fabricated 10 synthetic data sets as follows: (1) we draw $10,000$ samples from each pair of distributions, labeled with appropriate *id*s in Tbl. 1; (2) then, we divide the samples from each distribution into 5 different groups, called intervals (each interval contains $2,000$ random samples from a given distribution);

Table 2: Performance results for change point detection. (1) Using the bootstrap method, the estimated WLSDD accuracy A_b, false positives F_b, and delay D_b are presented for the optimal threshold thr_{opt}. The optimal window size N_w and offset N_d are obtained using cross-validation. (2) The averaged accuracy A, total number of false positives F, and delay D for different methods, are summarized. Our WLSDD method shows the best overall accuracy $\hat{A}$ and delay $\hat{D}$, while LSDD has the least number of false positives (average $\hat{F}$ and total F_t).

	WLSDD Bootstrap					**Change Point Detection Results**								
id	N_w	N_d	thr_{opt}	**Results**		**WLSDD**	**LSDD**	**KS**	**LL**	**LLH**	**E**	**EH**	**KL**	**KLH**
1	300	200	0.024	A_b	0.867	A 1.000	1.000	1.000	1.000	0.375	1.000	0.875	0.625	0.625
				F_b	1.333	F 4	1	2	1	0	1	5	2	0
				D_b	137.856	D 140.625	215.125	129.000	120.750	202.667	150.000	111.857	172.200	308.400
2	200	100	0.085	A_b	1.000	A 1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.556	1.000
				F_b	3.000	F 0	0	1	0	0	2	10	0	0
				D_b	60.178	D 49.778	77.667	37.889	60.333	110.222	51.667	100.000	130.400	64.222
3	500	400	0.008	A_b	1.000	A 1.000	1.000	0.444	0.778	0.889	1.000	1.000	0.778	1.000
				F_b	8.125	F 0	0	0	0	0	1	1	0	0
				D_b	431.844	D 315.222	383.556	399.250	327.000	500.625	596.889	326.444	262.143	423.333
4	500	400	0.010	A_b	0.889	A 1.000	1.000	1.000	1.000	0.889	1.000	1.000	1.000	0.778
				F_b	6.200	F 0	0	0	0	2	0	2	0	0
				D_b	226.600	D 147.556	202.444	181.889	229.222	599.125	224.444	406.000	254.556	416.714
5	300	200	0.048	A_b	1.000	A 1.000	0.556	1.000	0.889	1.000	1.000	1.000	0.667	1.000
				F_b	0.000	F 0	0	2	0	9	2	5	3	2
				D_b	137.222	D 144.333	152.000	143.111	183.625	165.222	140.000	111.333	123.667	354.556
6	200	100	0.056	A_b	1.000	A 1.000	0.778	0.556	1.000	1.000	1.000	1.000	0.778	1.000
				F_b	1.000	F 0	0	2	9	12	21	9	5	15
				D_b	69.311	D 84.444	94.571	149.000	163.333	220.000	137.889	191.667	103.000	153.556
7	400	300	0.047	A_b	0.889	A 1.000	0.556	1.000	1.000	1.000	0.889	0.778	0.667	1.000
				F_b	3.200	F 0	0	0	0	2	2	5	2	5
				D_b	196.175	D 205.333	260.200	148.667	149.333	478.444	270.000	305.571	199.667	335.444
8	200	100	0.100	A_b	1.000	A 1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				F_b	0.833	F 0	0	1	0	0	0	0	0	0
				D_b	49.350	D 51.778	70.778	27.333	56.444	81.667	52.000	66.778	79.333	77.000
9	200	100	0.032	A_b	1.000	A 1.000	0.889	0.889	0.778	1.000	1.000	1.000	0.667	0.889
				F_b	1.000	F 0	0	0	1	16	13	0	3	0
				D_b	97.089	D 98.000	108.125	116.875	112.571	131.222	196.889	101.778	138.000	122.250
10	200	100	1.019	A_b	1.000	A 1.000	1.000	1.000	1.000	1.000	0.889	1.000	0.556	1.000
				F_b	2.000	F 0	0	0	0	17	1	0	5	0
				D_b	68.822	D 56.889	69.000	66.444	66.000	154.444	138.375	76.667	64.000	76.667
						$\hat{A}$ **1.000**	0.878	0.889	0.945	0.915	0.978	0.965	0.729	0.929
						$\hat{F}$ 0.400	**0.100**	0.800	1.100	5.800	4.300	3.700	2.000	2.200
						F_t 4	**1**	8	11	58	43	37	20	22
						$\hat{D}$ **129.396**	163.347	139.946	146.861	264.364	195.815	179.809	152.697	233.214

(3) finally, we alternate intervals from different distributions, as illustrated in Fig. 1 (to preserve space, only 4 intervals are presented). For example, a data set with $id = 9$ in Tbl. 1 is constructed with alternating bimodal and separated bimodal intervals. Based on the designed synthetic data sets, we expect WLSDD to detect changes every $2,000$ samples.

Since we are interested only in change detection, in order to estimate the performance of our method on a relatively small number of change points available from the synthetic data sets, we suggest the following bootstrap approach. Using the original synthetic data stream (sketched on Fig. 1), we permute appropriate intervals for the given distributions, while keeping an alternating approach

between two distributions and true change points at multiples of 2,000 samples. Our approach is shown in Fig. 2 where the bootstrap data streams are obtained from the original data stream (again, only 4 intervals are presented in Figs. 1 and 2 for illustration purposes). Using the bootstrap data streams, we calculated the average accuracy, average false positives, and average delays for different threshold values (swept from 0 to true DD values, as listed in Tbl. 1 for appropriate data sets).

The average accuracy is calculated as the ratio between the number of accurately detected changes and the true number of change points across the bootstrap data sets. The average number of false positives is found as the mean of the faulty detected change points across the bootstrap data sets. Finally, the average time of delay is found based on the differences between the positively detected change instances and appropriate true change instances. For the optimal density-difference threshold, thr_{opt}, we selected a value for which the average accuracy is high and the average number of false positives and average delay values are low. The optimal threshold values for different bootstrap data sets are summarized in Tbl. 2 (we chose relatively low values for both the reference and current window sizes, N_w, as well as the offset, N_d).

Following our methodology of generating the synthetic data sets for the bootstrap approach, we created similar synthetic data sets that contain 10,000 samples in total, with 10 alternating intervals with random data from two different distributions (as per pair *id*s from Tbl. 1). Thus, each data set expresses 9 true change points every 1,000 samples. For our WSLDD approach, we used the optimal threshold values, thr_{opt}, obtained from our bootstrap approach (refer to Tbl. 2). For the LSDD method, we used the same thresholds as for WLSDD. For all other methods, we set the appropriate thresholds manually.

In Fig. 3, we illustrate the performance of all compared algorithms, using the synthetic data set that alternates between bimodal and separated bimodal intervals (id=9 distributions from Tbl. 1). The top graph, *amp*, presents the synthetic data samples. Density-difference measures, using our WLSDD method, as well as LSDD, KS, LL, LLH, E, EH, KL, and KLH, are stacked together in order to preserve time alignment with the original synthetic data set. All figures displaying the true change points are denoted by the black vertical lines. The blue vertical lines indicate the change points detected by the methods. Finally, faulty detected changes are shown as the red vertical lines. Our method reports estimated change points soon after the true points, without any false positives. Many methods (LL, LLH, E, KL) erroneously report changes, while some methods (LSDD, KS, LL, KL, and KLH) fail to detect some alterations between different intervals.

In summary, our method expresses the highest accuracy in detecting the change for the 10 synthetic data sets (with 10,000 samples) used, see Tbl. 2. Also, it shows the lowest delay in change detection making it suitable for applications that need to respond rapidly to changes. Although the LSDD method reports the least number of false positives, our method is a close second. All other methods display very inconsistent results for different data sets.

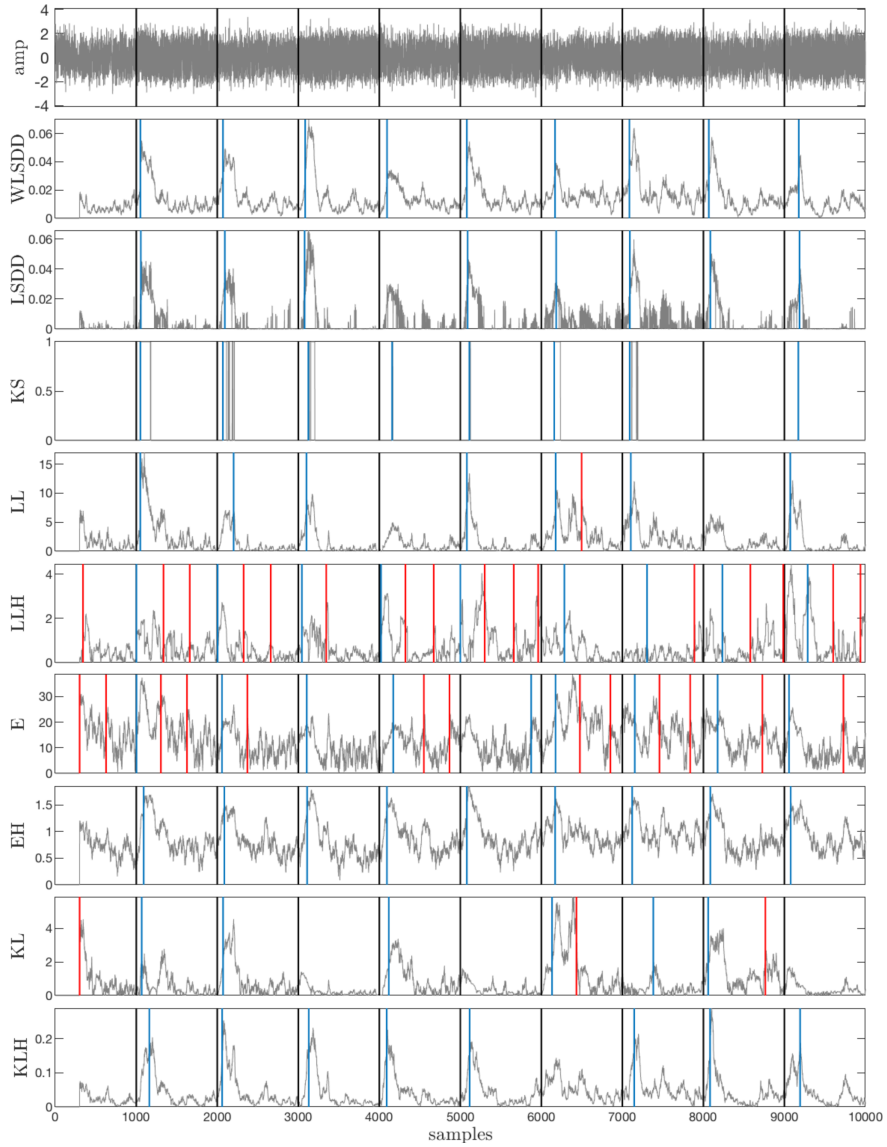


Fig. 3: Change point detection using different methods. The top figure (*amp*) presents the synthetic data samples (*id*=9, bimodal and separated bimodal alternating intervals). The true change points are indicated by the black vertical lines, true positive change points are shown as the blue lines, and the false positive change points are presented in red. Nine (9) different measures are used for comparison (in gray). *Our WLSDD method successfully detected all of the change points.* LSDD, KS, LL, KL, and KLH all missed at least one change point. LL, LLH, E and KL faultily reported non-existing change points.

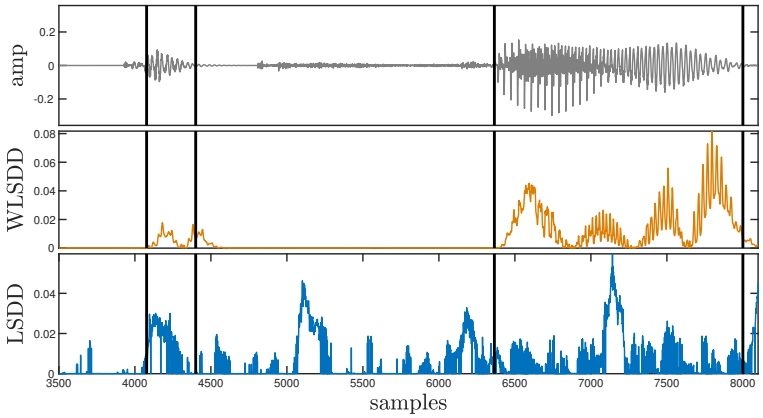


Fig. 4: Change point detection measures using the CENSCREC-1-C speech data (only subset of data is presented). The top figure (*amp*) presents the original data set (noise + silent and speech periods). The manually annotated change points are presented as the black vertical lines. Note that our WLSDD method is less sensitive to noise when compared to LSDD (around 4800 and 5730).

3.1.3 Real Data Set - CENSREC Speech Data Finally, we use the PSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition (CENSREC) data set to detect changes with our method. More specifically, we use the Noisy Speech Detection Evaluation Environment (CENSREC-1-C[1]) corpus, which consists of audio signals used for speech detection in noisy environments. Different silent and speech intervals alternate in the signals from the corpus. The change detection measures, using both WLSDD and LSDD methods, are presented in Fig. 4. The top figure presents a part of the original audio signal to allow for easier representation. The silence and speech periods are manually segregated with change points presented as the black vertical lines on all 3 figures. Our method perfectly detects the speech signals, while LSDD reports high values in the silence (low noise) intervals. Furthermore, it is important to note that the change measures reported by WLSDD follow natural speech variations during the spoken intervals (range between $6,500$ and $8,000$ samples).

4 Conclusion

In this paper, we presented a new method for determining the difference between two distributions using non-parametric wavelet expansions. Our WLSDD method is applied to a change point application, performing well on both synthetic and real data sets. While outperforming all other methods, it retains the highest point detection accuracy, lowest detection delay, and second lowest false positive detection among all competitors. In our future work, we plan to investigate regularization in order to mitigate the noise impact on our ℓ_2 density-difference measure.

[1] http://research.nii.ac.jp/src/en/CENSREC-1-C.html

References

1. Aminikhanghahi, S., Cook, D.J.: A survey of methods for time series change point detection. KAIS 51(2), 339–367 (2017)
2. Chentsov, N.N.: Evaluation of an unknown distribution density from observations. Doklady Akademii nauk SSSR 147(1), 45 (1962)
3. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Reg. Conf. Series in Applied Math., SIAM (1992)
4. Deshayes, J., Picard, D.: Off-line statistical analysis of change-point models using non parametric and likelihood methods. In: Detection of Abrupt Changes in Signals and Dynamical Systems, pp. 103–168. Springer (1985)
5. García-Treviño, E.S., Barria, J.A.: Online wavelet-based density estimation for non-stationary streaming data. CSDA 56(2), 327–344 (2012)
6. Kawahara, Y., Sugiyama, M.: Change-point detection in time-series data by direct density-ratio estimation. In: ICDM. pp. 389–400. SIAM (2009)
7. Kifer, D., Ben-David, S., G., J.: Detecting change in data streams. In: VLDB. vol. 30, pp. 180–191 (2004)
8. Peter, A.M., Rangarajan, A.: Maximum likelihood wavelet density estimation with applications to image and shape matching. TIP 17(4), 458–468 (2008)
9. Sebastião, R., Gama, J.: Change detection in learning histograms from data streams. In: PCAI. pp. 112–123 (2007)
10. Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. ACM SIGMOD 34(4), 42–47 (2005)
11. Sugiyama, M., Kanamori, T., Suzuki, T.e.a.: Density-difference estimation. Neur. Comp. 25(10), 2734–2775 (2013)
12. Sugiyama, M., Nakajima, S., et. al.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: NIPS. pp. 1433–1440 (2008)
13. Taylor, S.J., Letham, B.: Forecasting at scale. Amer. Stat. 72(1), 37–45 (2018)
14. Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D.: Detecting trend and seasonal changes in satellite image time series. Rem. Sens. of Env. 114(1), 106–115 (2010)
15. Verbesselt, J., Hyndman, R., Zeileis, A., Culvenor, D.: Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Rem. Sens. of Env. 114(12), 2970–2980 (2010)
16. Vidakovic, B.: Statistical modeling by wavelets, vol. 503. John Wiley & Sons (1999)
17. Wand, M.P., Jones, M.C.: Comparison of smoothing parameterizations in bivariate kernel density estimation. ASA 88(422), 520–528 (1993)
18. Wegman, E.J., Caudle, K.A.: Density estimation from streaming data using wavelets. CCS pp. 231–242 (2006)
19. Yamanishi, K., Takeuchi, J.: A unifying framework for detecting outliers and change points from non-stationary time series data. In: ACM SIGKDD. pp. 676–681 (2002)

Fuzzy time series applications and extensions: analysis of a short term load forecasting challenge

Guilherme Costa Silva[1], João L. R. Silva[1], Adriano C. Lisboa[2], Douglas A. G. Vieira[3], and Rodney R. Saldanha[1]

[1] Graduate Program in Electrical Engineering - Universidade Federal de Minas Gerais - Av. Antônio Carlos, 6627 31270-901, Belo Horizonte, MG, Brazil
guicosta@ufmg.br,joao@simula7.com,rodney@cpdee.ufmg.br
[2] Gaia Solutions on Demand - Rua Professor José Vieira Mendonça, 770, 31270-901, Belo Horizonte, MG, Brasil
adriano.lisboa@gaiasd.com
[3] ENACOM Handcrafted Technologies - Alameda das Latânias, 215, 31270-800, Belo Horizonte, MG, Brasil
douglas.vieira@enacom.com.br

Abstract. This article aims at studying and evaluating fuzzy time series algorithms. Some mechanisms for defining and partitioning the universe of discourse intervals, as well as their incorporation to multivariate algorithms, are discussed and applied to a short-term load forecasting challenge for testing and evaluation of their performance in hourly and daily forecasting, with interesting results.

Keywords: Fuzzy time series, Short term load forecasting, multivariate models

1 Introduction

Information on load demand in electrical systems operation can be considered of vital importance, since the operational problems and even the planning of electrical systems depend on these data, since the proper functioning of generators, substations, and lines, as well as the planning of new units, rely on the successful achievement of these data.

This load demand, according to aspects of trends and seasonalities, can be modeled through, for example, time series models of Box and Jenkins [1]. Such models may be useful and may provide more accurate decisions for other problems in electrical systems, however, some uncertainties must be considered, which is a shortcoming of conventional time series models.

The Fuzzy Time Series (FTS), the focus of this work, have some advantages, such as their easy implementation, accuracy improvements, the reduction in processing time and the simplification of calculation process, according to [2]. Other advantages of FTS are the ability to work with a very small set of data and no requirement related to the assumption of model linearity, according to [3].

In this article, we will test some mechanisms that compose evaluation of FTS, considering intervals definition and partitioning mechanisms and advanced models applied to Short Term Load Forecasting (STLF). The work accomplishes improvements in the studies done in [4], including methods for defining the universe of discourse and their intervals, as well as testing them in a recurring benchmark forecasting challenge, the GEFCom2012.

2 Fuzzy Time Series Review

The problem of short-term load forecasting is to predict demand in a range of few hours before, until few weeks before. According to [5], the choice of a daily temporal resolution means that the model must both include factors that change quickly, from one day to the next, and factors that change slowly, i.e. daily models can be much more useful for planning purposes than monthly or annual models.

According to [6], there are three goals for forecasting, such as scheduling the provision of safer and more reliable operation of power plants and the delivery of economic dispatches and greater reliability. With these goals taken into account, it is possible to model the load demand in time series some trends in electrical systems behavior.

According to [7], it is possible to obtain acceptable accuracy by performing forecasting with univariate models historical data, however, a certain level of accuracy can be improved by including weather variables.

Fuzzy time series, defined in [8], combine fuzzy set theory, defining degrees of membership rather than boolean values for each set, with techniques of time series analysis. These allow uncertainties in the model and achievement of more robust and accurate forecasts.

Some methods are reviewed in [9], including some hybrid methods, metrics, and databases. Some of the methods that can be highlighted are Chen's method in [10] and Yu's weighted method in [11].

2.1 FTS algorithm

In general, most FTS forecasting algorithms can be described by its modeling, which is presented below in the following steps summarized from [4].

Step 1. Define the universe of discourse according to a data range.

Step 2. Partition the universe of discourse into n intervals with equal length.

Step 3. Define fuzzy sets on the universe U according to the n intervals A_i, defined by linguistic values.

Step 4. Fuzzify historical data. In some FTS methods, fuzzy values are assigned automatically by using the n-interval length of U and the distance between m-th data and the mean of n-interval length.

Step 5. Establish fuzzy logical relationships (FLR), which defines the relation between the former state and the state to be forecasted, by determining the

sequence of fuzzy sets (A_i) that is equal to 1 and according to a sequence as follows:

$$F(t-\tau) \rightarrow F(t), \tag{1}$$

where τ is the instant considered before forecasting.

Step 6. Forecast all the right hand side of the fuzzy data in FLR according to the model applied.

These steps can be applied with any of the methods which are described in further subsections of this work.

2.2 UoD definition

For defining the universe of discourse according to Step 1, there are some methods. in [13], the Ratio-Based Lengths of Intervals-based method is proposed. The steps in this method define a base value to map the range and perform calculations of these ranges through the digits found considering its minimum value. The proportion percentile is defined *a priori*.

Some grouping methods can be also adopted, such as K-Means [12], which redefines the intervals, forming values adjusted by intervals.

2.3 UoD partition

After defining the universe of discourse, partition is performed as described in Step 2. The number of partitions can be defined using different methods.

The Huarng's Method, proposed in [14], defines the interval according to a base number, which is calculated from data size.

Two other methods are defined in [15] for this step, as the Sturges' method or the rule of p power, whose respective formulations are presented below:

$$p = 1 + 3.3, \log(n) \tag{2}$$

$$2^p < n, \tag{3}$$

with the second formulation being recursively defined, these recursions determines the number of partitions to be considered in the problem.

In addition to these methods, some evolutionary algorithms may be applied to find this value, according to studies in [16].

2.4 Multivariate FTS

In the Multivariate formulation of FTS, FLRs are determied according to the sequence of different variables. In this work, we consider two variables, namely F_1, to be forecasted, and F_2 as well as their fuzzy sets (A_i and B_i), within the instant τ in a sequence as follows:

$$F_1(t-\tau) \rightarrow F_2(t-\tau) \rightarrow F_1(t), \tag{4}$$

,

FLRs are established according to former states presented, then, FLRGs are defined by eliminating the recurrence of FLRs as in the example below:

$$\begin{aligned} A_1, B_1 &\rightarrow A_1, A_2 \\ A_1, A_2 &\rightarrow A_3 \\ &\vdots \\ A_{n-1}, B_n &\rightarrow A_{n-1}, A_n \\ A_n, B_n &\rightarrow A_n \end{aligned} \tag{5}$$

,
all variables in the right side are those sets whose membership degree, related to the evaluated instant, are equal to 1.

2.5 Advanced FTS models

The following advanced FTS models, used in forecasting step, are tested in this work:

Chen's Classical model - this model performs forecast is the mean of variables in FLRG, considering the midpoint M_i of fuzzy set A_i;

$$F(t+\tau) = \frac{\sum_{j=1}^{k} M_j}{k}. \tag{6}$$

Yu's Weighted model - this model forecasts data by weighting variables in FLRG;

$$F(t+\tau) = [M_1, M_2, ..., M_k] \times [w_1, w_2, \ldots, w_k]^T, \tag{7}$$

$$W(t+\tau) = \left[\frac{w_1}{\sum_{h=1}^{k} w_h}, \frac{w_2}{\sum_{h=1}^{k} w_h}, \ldots, \frac{w_k}{\sum_{h=1}^{k} w_h} \right]. \tag{8}$$

Exponential model - similar as in weighted model, but with exponential weighting, defined by the constant c;

$$W(t+\tau) = \left[\frac{1}{\sum_{h=1}^{k} w_h}, \frac{c}{\sum_{h=1}^{k} w_h}, \frac{c^2}{\sum_{h=1}^{k} w_h}, \ldots, \frac{c^{k-1}}{\sum_{h=1}^{k} w_h} \right]. \tag{9}$$

Interval model - this model defines intervals for FLR and the forecast $\mathbb{I}_f$ of this interval $[\underline{\mathbb{I}^j_{min}}, \overline{\mathbb{I}^j_{max}}]$.

$$\mathbb{I}_f = \sum_{j \in A} [\underline{\mathbb{I}^j_{min}}, \overline{\mathbb{I}^j_{max}}] \forall A_j = 1, \tag{10}$$

considering the following assumption:

$$F(t+\tau) \in \mathbb{I}_f. \tag{11}$$

This work performs forecasting by calculating the mean of intervals, by applying arithmetic or geometric means.

These methods are described in [4] with details, being part of Step 6 of the FTS algorithm.

2.6 Box-Cox Transformation

This work processes time series according to the Box-Cox transformation, which consists of obtaining the parameter λ and the vector Z from the original time series:

$$Z_t^\lambda = \begin{cases} \frac{Z_t^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(Z_t) & \lambda = 0 \end{cases} \tag{12}$$

With this, the obtained data Z is processed by FTS methods. This step is derived from Transformation FTS studied in [4]. To process the forecasted value $\hat{Z}$, we simply apply the inverse transformation in order to obtain $\hat{z}_{t+\tau}$ as follows:

$$\hat{z}_{t+\tau} = \begin{cases} \left(\lambda \hat{Z}_t^\lambda + 1\right)^{1/\lambda} & \lambda \neq 0 \\ \exp(\hat{Z}_t) & \lambda = 0 \end{cases} \tag{13}$$

note that the inverse transformation should be applied after forecasting steps and before evaluation. The above procedure normalizes data contained in time series and may provide an improved performance for all FTS methods.

2.7 The overall implementation

The system described throughout this section is implemented in Python 2.7 and can be widely described by its main steps according to Figure 1. Details about the application will be further presented in next section.

3 The GEFCom challenge

In order to apply FTS to more complex forecasting problems, the GEFCom2012 challenge was chosen for a set of tests considering several aspects for the methods presented in this paper.

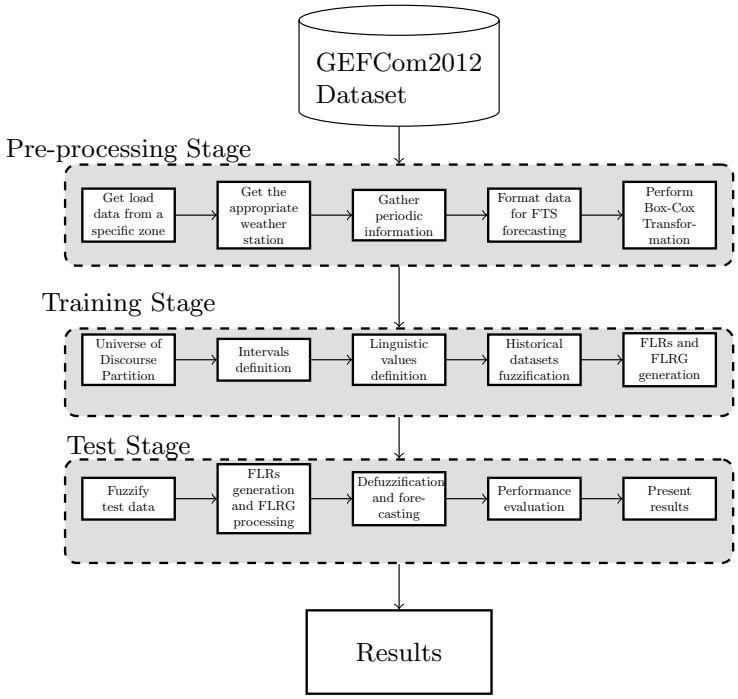


Fig. 1. Flowchart describing the framework of FTS algorithm.

3.1 Description

The GEFCom2012 is presented in [20], with some tasks required for a successful forecasting, such as data cleansing, as there are some missing values in the training data.

There are two types of values: load demand and temperature, considering this aspect, a multivariate method should be adopted in order to provide a proper forecast during test phase. There are 20 different zones and no additional information is provided regarding these zones.

The adopted methodology is the same as in [21], except for the use of a model, which is not applied in this approach, instead, we just fuzzify data as a part of processing. Thus, a weather selection technique and methodologies applied to hierarchical load data are discussed in [22], we used the mean of temperature from all zones evaluated.

The experiments consist of forecasting all zones from GEFCom2012 by analyzing load demand and weather data. Training samples are based on the data available of 2004–2006 range and test samples are based on 2007 data.

The FTS evaluation is performed in two approaches as follows:

1. Hourly, by forecasting the following hour based on the previous hour of the day ($\tau = 1$);
2. Daily, considering the 24 hours of a day being evaluated by forecasting the same hour based on the previous day ($\tau = 24$).

More about these tests are further described, considering the presented algorithms and GEFCom2012 tasks.

3.2 Challenge tasks

In order to provide proper results and more accurate forecasts, the GEFCom2012 can be evaluated through 6 different aspects as follows:

1. Data cleansing - days with missing data are ignored, and Box-Cox Transformation is performed;
2. Hierarchical forecasting - partial implementation;
3. Special days forecasting - optional step;
4. Temperature forecasting - required step, mean of zones used;
5. Ensemble forecasting - not considered;
6. Integration - performed in overall analysis.

Only tasks 1, 2, 4 and 6 are mandatory for the considered framework. Each load zone is evaluated separately, however, in order to improve results, task 2 is evaluated, and considering multivariate features of the proposed system, task 4 is a mandatory aspect.

There are some special cases, as described in some works. Zone 4 has an outage period and since Zone 9 is an industrial area, this zone is very insensitive to temperature effects. For these reasons, the analyses of both cases are ignored. The sum of all zones is also evaluated as a 21th zone of the database.

Considering these facts, we have 19 zones evaluated in the database. Tests of FTS algorithms in these presented datasets are further explained.

3.3 Tests procedure

The FTS models will be evaluated according to the percentage of absolute error (MAPE) and the interquartile of this error (IQR) that indicate the dispersion of these errors. The atypical days are replaced by the historical day of the equivalent day, except for the first days of January, whose data is omitted from the evaluation for this case.

Regarding the Universe of Discourse, the Ratio-Based Lengths of Intervals is considered as a base case. In the other cases, U is defined by K-means, while Huarng, Sturges or Rule of p power methods are used for partitioning.

For the ratio calculation in the Universe of Discourse definition, the value of 0.01 is adopted. For the Box-Cox transformation method, the value used is $\lambda = 0$, making a logarithmic normalization and avoiding calculation problems. In terms of comparisons, the methods are compared to previous results in the competition.

3.4 Results

Zone analysis Results are presented in Table 1, considering all methods using Huarng's method. Considering these results, methods based both on Classical or Weighted approaches provided better results. When there is a tie between methods, the one with lower interquartile is considered.

UoD methods analysis For a deeper analysis, Zone 21 is analyzed in further cases, the four defining and partitioning strategies applied to the Universe of Discourse are tested and presented in Figure 2 for all 5 FTS methods. Based on the results, both Ratio-Based Lengths of Intervals and K-means with Huarng's method presented better performances.

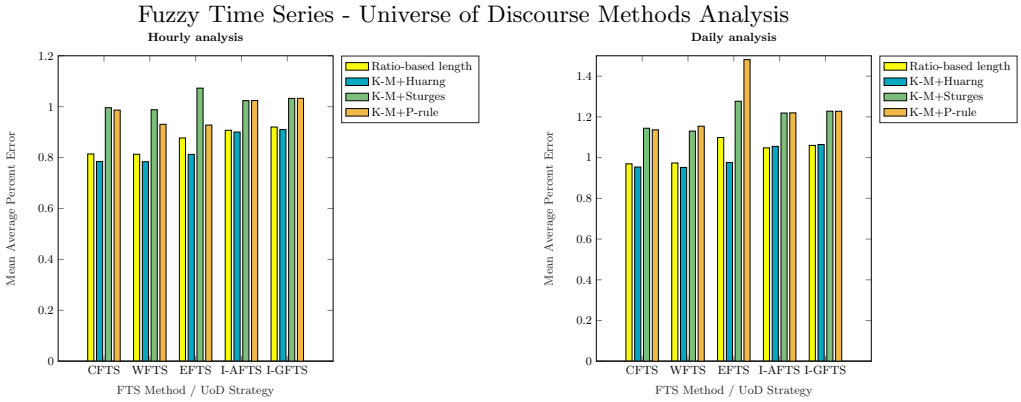


Fig. 2. Performance of FTS algorithms according to UoD mechanisms.

This analysis is relevant as the Universe of Discourse modeling is a part of FTS forecasting as the evaluated methods rely on a proper definition of fuzzy variables.

Periodic analysis Considering the daily forecasting case, we can extract more information regarding which hours and days each method can be more predictable. In this case, both strategies 1 and 2 from UoD definition are considered, due to their good performance in tests. In Figure 3 the results obtained by using the Ratio-Based Lengths of Intervals and Figure 4 presents the experiments using K-means and Huarng's method, both in daily forecasting.

By analyzing forecasting per hour and weekdays, data from daylight time tends to be more predictable than data from nighttime for most FTS methods. Regarding the daily forecast, data from Tuesdays and Wednesdays were more predictable.

Table 1. Results for all FTS methods in the base case.

Method/		Classical		Weighted		Exponential		Int.Arithmetic		Int.Geometric	
Zone	Period	MAPE	IQR	MAPE	IQR	MAPE	IQR	MAPE	IQR	MAPE	IQR
1	Hourly	**1.486**	**1.536**	1.477	1.561	1.604	1.684	1.669	1.698	1.704	1.743
	Daily	1.803	1.965	**1.796**	**1.971**	1.957	2.070	1.989	2.120	1.997	2.141
2	Hourly	1.030	0.972	**1.011**	**0.992**	1.121	1.117	1.117	1.155	1.135	1.185
	Daily	**1.143**	**1.095**	1.148	1.098	1.267	1.296	1.266	1.264	1.290	1.287
3	Hourly	**1.020**	**1.010**	1.037	0.985	1.103	1.132	1.109	1.149	1.127	1.174
	Daily	1.136	1.075	**1.125**	**1.084**	1.258	1.314	1.253	1.247	1.277	1.268
5	Hourly	**2.013**	**2.207**	**2.014**	**2.205**	2.228	2.391	2.285	2.460	2.310	2.489
	Daily	2.434	2.536	**2.429**	**2.521**	2.749	2.901	2.781	2.767	2.784	2.818
6	Hourly	1.028	0.973	**1.012**	**0.997**	1.141	1.176	1.118	1.153	1.136	1.178
	Daily	1.151	1.111	**1.143**	**1.103**	1.260	1.289	1.274	1.279	1.297	1.295
7	Hourly	**1.020**	**1.010**	1.037	0.985	1.103	1.132	1.109	1.149	1.127	1.174
	Daily	1.136	1.075	**1.125**	**1.084**	1.258	1.314	1.253	1.247	1.277	1.268
8	Hourly	**1.867**	**1.988**	1.882	1.957	1.950	2.087	2.170	2.203	2.233	2.261
	Daily	2.181	2.290	**2.176**	**2.296**	2.528	2.648	2.561	2.696	2.638	2.761
10	Hourly	**1.412**	**1.446**	1.421	1.437	1.679	1.715	2.550	2.753	2.791	2.932
	Daily	**1.555**	**1.641**	1.562	1.621	1.839	1.871	2.432	2.542	2.688	2.747
11	Hourly	1.356	1.284	**1.349**	**1.292**	1.456	1.508	1.516	1.419	1.554	1.447
	Daily	1.551	1.570	**1.541**	**1.573**	1.651	1.674	1.833	1.837	1.891	1.854
12	Hourly	1.357	1.412	**1.348**	**1.409**	1.414	1.488	1.519	1.559	1.556	1.579
	Daily	1.613	1.718	**1.604**	**1.738**	1.842	1.939	1.764	1.941	1.795	1.963
13	Hourly	1.540	1.453	**1.535**	**1.441**	1.624	1.686	1.647	1.666	1.672	1.683
	Daily	1.699	1.795	**1.700**	**1.764**	1.831	1.946	1.863	2.045	1.887	2.054
14	Hourly	**1.748**	**1.862**	1.762	1.858	1.836	1.965	2.043	2.193	2.103	2.233
	Daily	**2.141**	**2.337**	2.148	2.324	2.426	2.562	2.449	2.603	2.491	2.649
15	Hourly	**1.137**	**1.162**	1.150	1.138	1.265	1.321	1.355	1.415	1.387	1.440
	Daily	1.346	1.450	**1.346**	**1.438**	1.504	1.570	1.476	1.550	1.486	1.558
16	Hourly	**1.556**	**1.632**	1.558	1.605	1.677	1.747	1.883	1.958	1.945	1.997
	Daily	1.967	2.062	**1.959**	**2.080**	2.130	2.258	2.248	2.357	2.310	2.426
17	Hourly	1.037	1.089	**1.018**	**1.122**	1.090	1.144	1.243	1.353	1.276	1.388
	Daily	1.304	1.408	**1.301**	**1.389**	1.456	1.527	1.455	1.589	1.481	1.642
18	Hourly	1.242	1.263	**1.233**	**1.240**	1.316	1.373	1.409	1.457	1.435	1.484
	Daily	1.511	1.613	**1.509**	**1.610**	1.749	1.808	1.884	1.884	1.926	1.913
19	Hourly	**1.380**	**1.453**	1.395	1.432	1.597	1.718	1.576	1.694	1.612	1.724
	Daily	**1.683**	**1.841**	1.686	1.818	1.940	2.064	2.012	2.061	2.052	2.085
20	Hourly	1.092	1.086	**1.076**	**1.095**	1.182	1.221	1.190	1.186	1.211	1.200
	Daily	1.229	1.226	**1.225**	**1.232**	1.346	1.400	1.476	1.386	1.514	1.390
21	Hourly	0.814	0.822	**0.806**	**0.821**	0.872	0.936	0.907	0.918	0.920	0.927
	Daily	**0.969**	**1.035**	0.972	1.026	1.065	1.131	1.048	1.151	1.060	1.154

Regarding the methods, Classical and Weighted FTS have presented good performance, and for K-means and Huarng's method, presented results were more consistent, mainly considering the Exponential method.

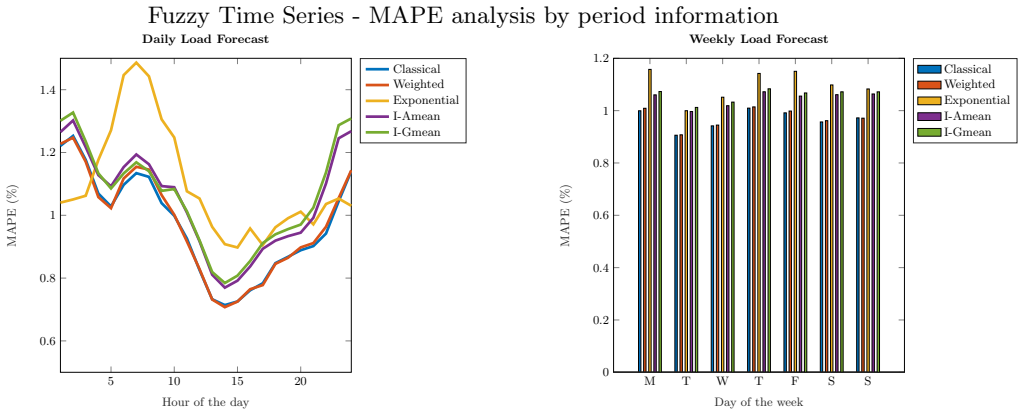


Fig. 3. Performance of FTS algorithms according to a given period (hour/day) using Strategy 1.

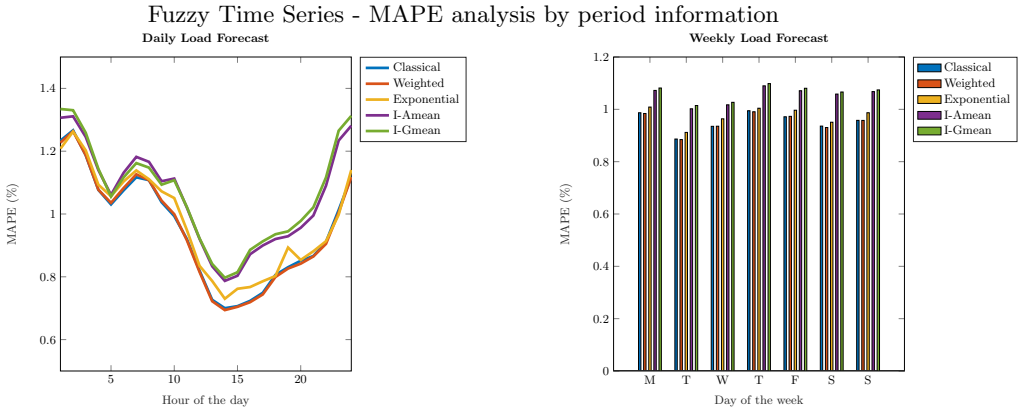


Fig. 4. Performance of FTS algorithms according to a given period (hour/day) using Strategy 2.

4 Conclusion and future work

This work presented the application of Fuzzy Time Series models in the GEFCom2012 challenge. The set of methods are based on Transformation FTS algorithm, studied in [4] and employ a multivariate model, considering two or more variables in the model, in this case, the mean temperature of all 11 zones, besides load data. These models provided interesting and competitive results.

High-order based FTS (which employs forecasting based on multiple instants) were omitted from the analysis, however, such study can be performed by including these models and verifying how good the performance of them can be and

how many instants can be used in the analysis. Another possibility is to evaluate Adaptive-Expectation strategies and their performance in FTS algorithms.

Some other challenges besides GEFCom2012 can be studied. A good example is the GEFCom2014 [23], focused on probabilistic forecasting, among other examples. In this case, the use of probabilistic models of FTS, such as in [24], can be considered.

Finally, there is a class of forecasting algorithms based on pattern similarity models [25]. Considering the daily forecasting model analyzed in this work, FTS approaches can also be incorporated to these models, providing a very interesting approach which deals with the daily forecasting with more refined results. Optimization methods may also improve these studies as well.

Acknowledgments. The authors acknowledge the support of CNPq, FAPEMIG and CAPES in this study.

References

1. G. E. P. Box, G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1970
2. M. Shah, *Fuzzy Time Series: A Realistic Method to Forecast Gross Domestic Capital of India.*, In: Melin P., Castillo O., Ramírez E.G., Kacprzyk J., Pedrycz W. (eds) Analysis and Design of Intelligent Systems using Soft Computing Techniques. Advances in Soft Computing, vol 41. Springer, Berlin, Heidelberg, 2007.
3. C. H. Aladag, U. Yolcu, E. Egrioglu, *A high order fuzzy time series forecasting model based on adaptive expectation and artificial neural networks*, Mathematics and Computers in Simulation, vol. 81, no. 4, pp. 875–882, 2010.
4. G. Costa Silva, J. L. R. Silva, A. C. Lisboa, D. A. G. Vieira and R. R. Saldanha, *Advanced fuzzy time series applied to short term load forecasting*, 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Arequipa, 2017, pp. 1-6.
5. I. M. Trotter, T. F. Bolkesjø, J. G. Féres, L. Hollanda, *Climate change and electricity demand in Brazil: A stochastic approach*, Energy, vol. 102 (2016), pp. 596-604.
6. M. Q. Raza, A. Khosravi, *A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings*, Renewable and Sustainable Energy Reviews, vol. 50 (2015), pp. 1352–1372.
7. Chapagain, K.; Kittipiyakul, S., *Performance Analysis of Short-Term Electricity Demand with Atmospheric Variables*, Energies, Vol. 111, No. 4, 818 (2018).
8. Q. Song and B. S. Chissom, *Fuzzy time series and its models*, Fuzzy Sets and Systems vol. 54, no. 3 (1993), pp. 269–277. North-Holland. Elsevier.
9. Singh, P. *A brief review of modeling approaches based on fuzzy time series*, Int. J. Mach. Learn. & Cyber. vol. 8, no. 2 (2017), pp. 397–420.
10. S.-M. Chen, *Forecasting enrollments based on fuzzy time series*, Fuzzy Sets and Systems, vol. 81, no. 3, 1996, pp. 311–319.
11. H.–K. Yu, *Weighted fuzzy time series models for TAIEX forecasting*, Physica A: Statistical Mechanics and its Applications, vol. 349, no. 3, 2005, pp. 609–624.
12. Zhang, Z. and Zhu, Q. *Fuzzy Time Series Forecasting Based On K-Means Clustering*, Open Journal of Applied Sciences, 2, 100–103, 2012.

13. K. Huarng, T. H.–K. Yu, *Ratio-based lengths of intervals to improve fuzzy time series forecasting*, in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 2, pp. 328-340, April 2006.
14. K. Huarng, *Effective lengths of intervals to improve forecasting in fuzzy time series*, Fuzzy Sets and Systems, vol. 123, no. 3, pp. 387–394, 2001.
15. R. Efendi, Z. Ismail, M. M. Deris, *Improved Weight Fuzzy Time Series As Used In The Exchange Rates Forecasting Of Us Dollar To Ringgit Malaysia*, International Journal of Computational Intelligence and Applications vol. 12, no. 1, article ID: 1350005, 19 pages, 2013.
16. S. M. Chen and N. Y. Chung, *Forecasting enrollments using highorder fuzzy time series and Genetic Algorithms*, International Journal of Intelligent Systems, vol. 21, pp. 485-501, 2006.
17. F. M. Talarposhti, H. J. Sadaei, R. Enayatifar, F. G. Guimarães, M. Mahmud, T. Eslami, *Stock market forecasting by using a hybrid model of exponential fuzzy time series*, International Journal of Approximate Reasoning, vol. 70, 2016, pp. 79-98.
18. M. H. Lee, H. J. Sadaei, Suhartono, *Improving TAIEX forecasting using fuzzy time series with Box–Cox power transformation*, Journal of Applied Statistics vol. 40, no. 11, 16 pages, 2013.
19. S.-M. Chen, *Forecasting Enrollments Based On High-Order Fuzzy Time Series*, Cybernetics and Systems vol. 33 , no. 1, pp. 1-16, 2002.
20. J. R. Lloyd, *GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes*, International Journal of Forecasting, vol. 30, no.2 (2014), pp.369–374.
21. P. Wang, B. Liu, T. Hong, *Electric load forecasting with recency effect: A big data approach*, International Journal of Forecasting, vol. 32, no. 3 (2016), pp. 585–597.
22. T. Hong, P. Wang, L. White, *Weather station selection for electric load forecasting*, International Journal of Forecasting, vol. 31, no. 2 (2015), pp. 286-295.
23. V. Dordonnat, A. Pichavant, A. Pierrot, *GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models*, International Journal of Forecasting, vol. 32, no. 3 (2016), pp. 1005-1011.
24. P. C. L. Silva, M. A. Alves, C. Alberto, S. Junior, G. L. Vieira, F. G. Guimaraes, H. J. Sadaei, *Probabilistic Forecasting with Seasonal Ensemble Fuzzy Time-Series*. XIII Brazilian Congress on Computational Intelligence, 2017.
25. G. Dudek, *Pattern similarity-based methods for short-term load forecasting - Part 2: Models*, Appl. Soft Comput., vol. 36 (2015), Pages 422–441.

Selection of neural network for crime time series prediction by Virtual Leave One Out tests

Stanisław Jankowski[1], Zbigniew Szymański[1], Zbigniew Wawrzyniak[1], Paweł Cichosz[1], Eliza Szczechla[2] and Radosław Pytlak[1]

[1]Warsaw University of Technology, Warsaw, Poland
sjankows@elektron.elka.pw.edu.pl, z.szymanski@ii.pw.edu.pl, z.wawrzyniak@ise.pw.edu.pl, p.cichosz@elka.pw.edu.pl, r.pytlak@mini.pw.edu.pl
[2]Scott Tiger S.A., Warsaw, Poland
eliza.szczechla@tiger.com.pl

Abstract . The goal of this paper is the application of the virtual leave-one-out methodology to the selection of optimal neural network structure for time series prediction. The experiments are performed on the real data set of spatiotemporal crime incidence for forecasting in the time coordinate. Due to the idea of local linearization, the estimation of generalization can be obtained in analytical form; hence the method is computationally efficient.

Keywords: neural network, virtual leave one out, crime forecasting, influential statistics

1 Introduction

Large-scale computing infrastructures and sensing technologies have produced a variety of spatially and time-dependent datasets exploiting human dynamics data [1]. The areas of urban computing [2] are intimately connected with transportation systems, social applications, and public safety and security. In the latter issue, understanding of criminal patterns within the crime datasets by the model of hotspots [3] results in the possibility of forecasting of crime [4]. The widely used assumption of hot spots and sophisticated spatiotemporal statistical models have been proposed for criminal events [5]. Temporal constraints are a severe determinant in spatial-temporal patterns of property crime in conjunction with the locations of hotspots [6]. Prediction of criminal events over time is associated with chronological sequence of observations (time series, TS). The developed TS analysis techniques for forecasting based on machine learning data-driven approaches could perform well [7].

The data are preprocessed by using the ARIMA model of time series forecasting [8]. The data are analyzed by autocorrelation and partial autocorrelation functions. It can be stated that the raw time series should be differentiated in order to remove the trend. It can be concluded that time series can be modeled as autoregression model.

Neural networks are used to improve the quality of prediction by nonlinear components [9-11].

The goal of this paper is to apply the virtual leave-one-out methodology (VLOO) to the selection of the optimal neural network structure for time series prediction. Our experiments are performed on a real data set of spatiotemporal crime incidence for forecasting in the time coordinate [12, 13].

The main task is to select the optimal network structure: the number of inputs – data from past periods and the number of hidden neurons needed to represent the nonlinear function of a predictor. The criterion of selection is the expected generalization capability usually based on test data set results. The cross-validation is considered as a very efficient method, often in the form of leave-one-out. However, this approach requires retraining of the neural network for each example of the training set [14, 15].

2 Analyzed dataset

2.1 Input data

The original data set contained 615961 records describing events from unstructured police reports. The only data fields used in the study were date and type (the data structure is shown in tab. 1. The data studied in this work cover the period 2008 to the end of 2014. The anonymized records contain information about 12 types of events such as robbery, burglary, misdeed, violence, and interventions not assigned to any of the categories listed.

Table 1. Description of data record.

Field name	Size	Description
Src	varchar(10)	Source id
Date	Datetime	Date of the event
Day of week	tinyint(4)	Day of week
Type	varchar(20)	Event type
County	varchar(30)	
Free_kids	tinyint(4)	Boolean flag indicating pupil free day
Free_aldults	tinyint(4)	Boolean flag indicating a free day
Long_weekend	tinyint(4)	Boolean flag indicating extended weekend
National_holiday	tinyint(4)	Boolean flag indicating national holiday
Religious_holiday	tinyint(4)	Boolean flag indicating religious holiday
geo_long	Double	Event location - WGS84 longitude
geo_lat	Double	Event location - WGS84 latitude
Political_assembly	tinyint(4)	Boolean flag indicating political assembly
Youth_assembly	tinyint(4)	Boolean flag indicating youth assembly
Mass_event	tinyint(4)	Boolean flag indicating other mass event
Supporter	tinyint(4)	Boolean flag indicating football match

The original incident count data were aggregated by day (fig. 1) – the aggregation was performed by summing up the total number of events and the events of individual types. The processed data were analyzed in subsequent stages of presented work.

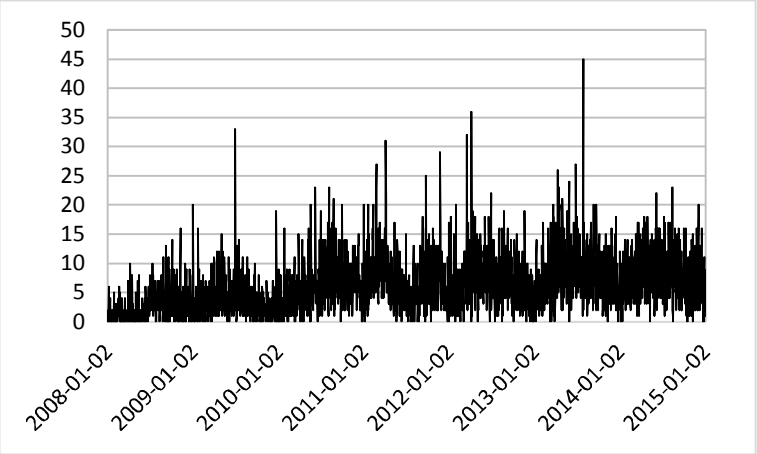


Fig. 1. Original time sequence – the number of events per day (misdeed events).

For the sake of simplicity in this project, we do not consider the crime type forecasting problem.

2.2 Data processing

Preprocessing of time series for neural network modelling involves: removing of the mean value, differentiation, and normalization. In the first stage, the mean value 6.22 was subtracted from the time sequence shown in fig. 1.

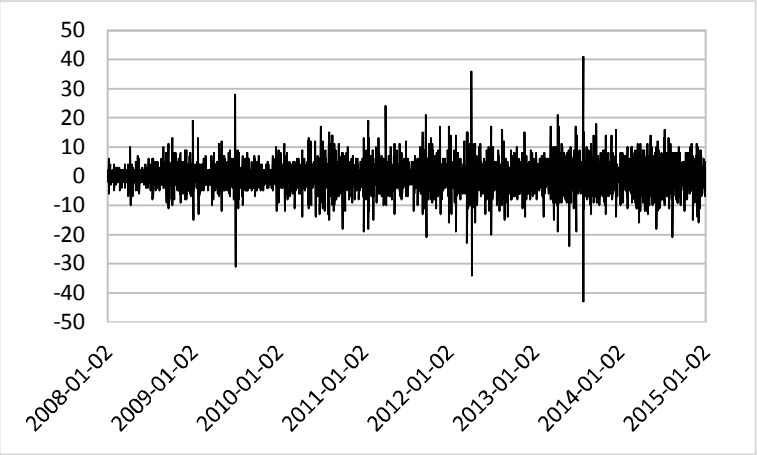


Fig. 2. Differentiated time series.

The stationary time series manifests fluctuations around the mean value and its autocorrelation function quickly decreases to 0. If the series has positive values of the autocorrelation function for many delay values (e.g., 10 or more), it means differen-

tiation should be performed. Differentiation tends to introduce negative autocorrelations. Fig. 2 shows differentiated times series according to formula

$$x_{diff}(n) = x(n) - x(n-1) \quad (1)$$

The last processing step was data normalization according to formula

$$x_{diff_{norm}}(n) = \frac{x_{diff}(n)}{std(x_{diff})} \quad (2)$$

2.3 Data for NN

The structure of the input data fed to the neural network is explained in tab. 2. The input of the neural network is a vector of k+1 consecutive values from the time series (fig. 3). The value of parameter k depends on the implemented model structure.

Table 2. The structure of learning data.

NN input				NN output
Y(0)	Y(1)	…	Y(k)	Y(k+1)
Y(1)	Y(2)	…	Y(k+1)	Y(k+2)
…	…	…	…	…
Y(i)	Y(i+1)	…	Y(k+i)	Y(k+i+1)

3 NN model of time series

Nonlinear prediction of time series can be implemented as a neural network – multilayer perceptron. The problem of neural network design is a number of inputs equal to time series delays and the number of hidden neurons necessary to represent the complexity of a given time series. The number of inputs is suggested by a linear ARMA model obtained by the autocorrelation and partial autocorrelation functions [8,16].

Fig. 4 presents the three-layered structure of a neural network consisting of inputs, hidden neurons and, an output neuron. Activation function used is hyperbolic tangent. The output neuron is linear [9-11].

At the verification stage, the properties of the model residuals are checked, which should have white noise properties. It means that the autocorrelation coefficients of residues should not differ significantly from zero. For this purpose, the graphs of autocorrelation functions for a series of model residues are analyzed.

The problem of creating neural models of time series is to avoid overfitting, which is the result of too large number of network parameters, i.e., hidden neurons. There are advanced methods of neural networks testing to help solve this problem effectively. The virtual leave one out cross-validation test is particularly recommended. The so-called sparseness characterizes the obtained models, i.e., the minimum number of parameters (weights).

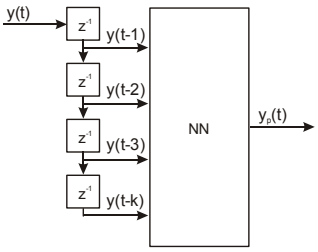


Fig. 3. Model structure

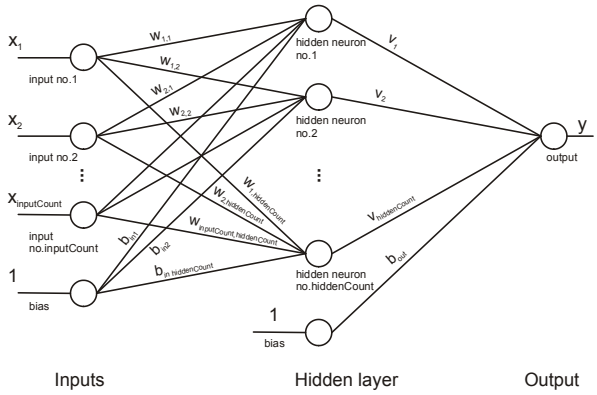


Fig. 4. Neural network structure

4 Local overfitting control - virtual leave-one-out

A locally linear model can approximate a nonlinear model. In the vicinity of the goal function minimum **w*** a nonlinear model g(**x**, **w***) can be approximated by using Taylor series expansion:

$$g(\mathbf{x},\mathbf{w}) = g(\mathbf{x},\mathbf{w}^*) + \mathbf{Z}(\mathbf{w}-\mathbf{w}^*) \tag{3}$$

where: $\mathbf{x} = [y(i), y(i-1),\ldots, y(i-k), 1]$ is the input vector, k – the maximum time delay. $\mathbf{Z}(N, q)$ is the Jacobian matrix of the nonlinear neural network of size (N, q), $N > q$, where: N – the number of training examples, q – the number of the neural network weights.

$$z_i = \left(\frac{\partial g(\mathbf{x},\mathbf{w})}{\partial w_i}\right)_{\mathbf{w}=\mathbf{w}^*} \tag{4}$$

$$\mathbf{Z}_{N\times q} = \frac{\partial J}{\partial w_i} = \left(\frac{\partial (y_p - g(\mathbf{x},\mathbf{w}))^2}{\partial w_i}\right) = -2(y_p - g(\mathbf{x},\mathbf{w}))\frac{\partial g(\mathbf{x},\mathbf{w})}{\partial w_i} \tag{5}$$

If the output error:

$$y_p - g(\mathbf{x},\mathbf{w}) = -\frac{1}{2} \tag{6}$$

then the gradient of the goal function equals the gradient of the network output.

The Jacobian **Z** can be obtained using backpropagation of the error -1/2 after the network learning is completed. The local linear approximation of the neural network weights has the least-squares form:

$$\mathbf{w}_{LS} = \mathbf{w}^* + (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T[y_p - g(\mathbf{x},\mathbf{w}^*)] \tag{7}$$

By introducing the **H** matrix of the orthogonal projection of the **Z** matrix onto solution subspace, the effect of withdrawing *k*-th example from the training set can be estimated

$$\begin{aligned} &\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T \\ &h_{kk} = \mathbf{z}^{kT}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}^k \\ &\sum_{k=1}^{N} h_{kk} = q, \quad 0 \le h_{kk} \le 1 \end{aligned} \tag{8}$$

The diagonal elements h_{kk} of **H** matrix are leverages - *k*-th components of orthogonal projections. The leverages are equal:

$$h_{kk} = \mathbf{z}^{kT}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}^k = \sum_{l=1}^{q}\sum_{j=1}^{q} Z_{kl}Z_{kj}(\mathbf{Z}^T\mathbf{Z})^{-1}_{lj} \tag{9}$$

It can be shown that the virtual leave-one-out method can estimate the effect of withdrawing one example of the training set on the network weights [12-13]:

$$\mathbf{w}_{LS}^{(-k)} = \mathbf{w}^* + (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}^k \frac{r_k}{1-h_{kk}} \tag{10}$$

where the residual r_k is:

$$r_k = y_p^k - g(\mathbf{x}^k,\mathbf{w}) \tag{11}$$

The leverages can be calculated by the singular value decomposition SVD. The virtual leave-one-out residual equals:

$$r_k^{(-k)} = \frac{r_k}{1-h_{kk}} \tag{12}$$

If all leverages were identical then:

$$h_{kk} = \frac{q}{N} \tag{13}$$

If $h_{kk} = 0$, then a model is exact for k-th example. If $h_{kk} = 1$, then the influence of k-th example is extremely high. The leverages can be interpreted as a measure of influence of each training example on the neural network structure. Therefore our goal is to select a model whose leverages are approximately equal.

Quantitative criteria of model selection [12, 13] are defined as follows. Prediction error of the virtual-leave-one-out test E_p is equal:

$$E_p = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{r_k}{1-h_{kk}}\right)^2} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(r_k^{(-k)})^2} \tag{14}$$

This value can be compared to training mean square error TMSE:

$$TMSE == \sqrt{\frac{1}{N}\sum_{k=1}^{N} r_k^2} \tag{15}$$

The quantity E_p can characterize models that overfit a subset of training examples. The relations hold:

$$\forall k: \quad h_{kk} = \frac{q}{N} \Rightarrow E_p = \frac{N}{N-q} TMSE \tag{16}$$

The leverages distribution can be characterized by the quantity μ:

$$\mu = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{N}{q}h_{kk}} \tag{17}$$

If $\mu = 1$ then $h_{kk} = \frac{q}{N}$ for $k = 1,\ldots, N$, hence this quantity can be used as an index of overfitting if $\mu \leq 1$.

As each neural network is characterized by 2 quantities, the result of model testing can be presented by E_p-μ plot. Upon the presented methodology, the selected neural network model should be described by the minimum E_p and the maximum (close to 1) μ.

The influence of training examples on the model can be also analyzed by confidence intervals. The confidence interval of the predicted output value if k-th example is virtually withdrawn is equal:

$$\pm t_{\alpha}^{N-q} s \sqrt{\frac{h_{kk}}{1-h_{kk}}} \tag{18}$$

where t_{α}^{N-q} is t Student value for N-q degrees of freedom, α is the confidence level, s is the standard deviation of training set prediction.

5 Results

The described method was applied to a real dataset discussed in section 2. The experiments were performed on several subsets of the whole dataset containing 200 samples. The first 100 samples were used as a training subset, and remaining samples were used for prediction tests. Model learning and prediction tests were performed as one step ahead predictor.

In order to design the neural network optimization, the starting point for neural network structure was the ARIMA model. The source data was preprocessed and differentiated. Previously prepared ARIMA model indicated no MA component is needed and it will be reasonable to use 3-5 delays.

The set of examined models comprises neural networks with 3 to 14 inputs and 3 to 40 hidden neurons. The network training was performed five times for each network architecture. The initial weight values were randomly selected in order to avoid local minima [9, 10]. Synthetic results of model quality assessment are shown as in the E_p-μ plot for selected neural networks (see Fig. 5). The preferred models are in the proximity of E_p equal to 0.8 and μ equal to 0.94.

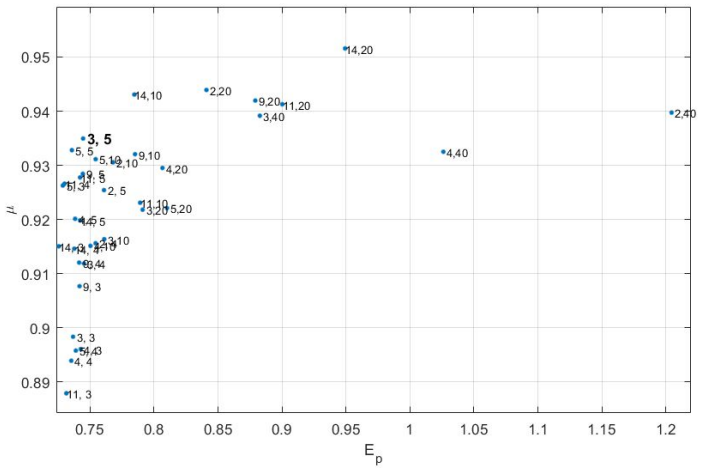


Fig. 5. Prediction error of the virtual-leave-one-out test E_p vs quantity μ - the E_p-μ plot. Each point represents one neural network, the numbers in brackets denote respectively the number of inputs, and the number of hidden neurons.

The detailed results are shown for two models:

- the model was shown a bad generalization property comprising 15 hidden neurons and four inputs – called NN (4-15).

- for comparison, the model best matching VLOO criteria comprising five hidden neurons and three inputs – called NN (3-5),

5.1 Synthetic illustrations for the NN (4-15) model according to VLOO criteria.

Fig. 6 shows the leverage histogram for the NN (4-15) model. The distribution of leverages is concentrated around value 1, so the model is too sensitive to a fraction of the training data – it is overfitted.

Prediction results of the NN (4-15) model are shown in fig. 8. The green dots indicate the influential points (leverage values greater than 0.9).

The autocorrelation function of prediction error (fig. 10) for the NN (4-15) model shows that white noise criterion is not satisfied, because some lag values exceed confidence interval.

The confidence interval (fig. 12) width for the time series prediction on learning set indicates that NN (4-15) model reliability is not very high.

5.2 Synthetic illustrations for the NN (3-5) model according to VLOO criteria.

Fig. 7 shows the leverage histogram for the NN (3-5) model. The distribution of leverages is concentrated around value close to 0, so the model is not sensitive to a fixed fraction of the training data – the overfitting does not occur.

Prediction results of the NN (3-5) model are shown in fig. 9. No influential points are present.

The autocorrelation function of prediction error (fig. 11) for the NN (3-5) model shows that white noise criterion is satisfied, because all lag values are within the confidence interval.

According to (18) the confidence interval for the time series prediction on the learning set indicate that selected NN (3-5) model is more reliable than NN (4-15) model due to more uniform leverages distribution.

The sliding window prediction is summarized in Table 3. The average MSE value of the NN (3-5) model on the test set equals 1.303 and is lower than corresponding value of the NN (4-15) model;, hence the generalization ability of the NN (3-5) model is better (despite the similar MSE values on the training set).

6 Conclusions

We applied the virtual leave-one-out approach based on an analytical approximation of the effect of withdrawing one example from the training set without withdrawing by using the local linearization of the neural network predictor [12, 13]. This approach enables the general model diagnostics by performing the influential statistics [14]. In general, this approach is attractive in the theory of statistical learning systems as e.g., new semi-supervised classifiers [17].

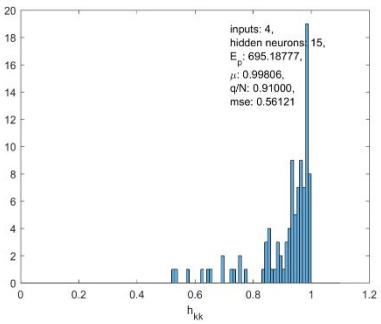


Fig. 6. Leverage histogram for the NN (4-15) model.

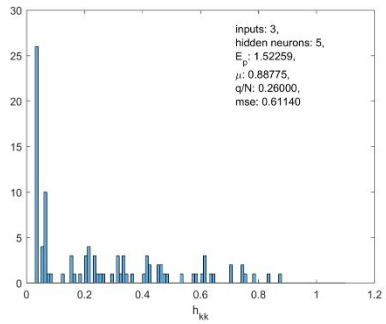


Fig. 7. Leverage histogram for the NN (3-5) model.

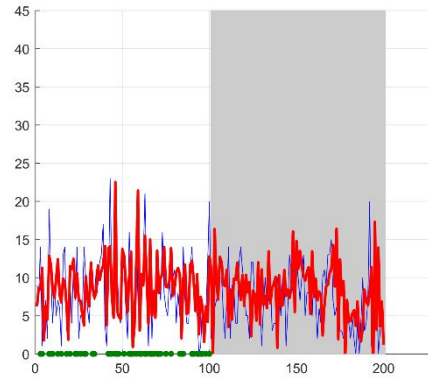


Fig. 8. Prediction results of the NN (4-15) model (blue – original data, red – prediction). Gray background indicates the testing set. White background indicates the learning set.

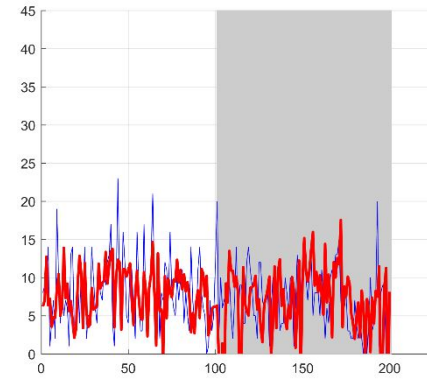


Fig. 9. Prediction results of the NN (3-5) model (blue– original data, red – prediction). Gray background indicates the testing set. White background indicates the learning set.

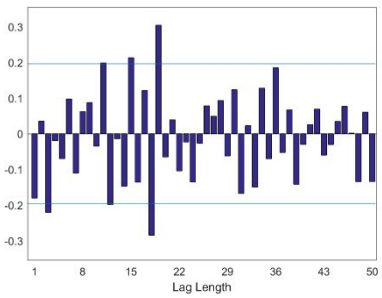


Fig. 10. Autocorrelation function of prediction error for the NN (4-15) model.

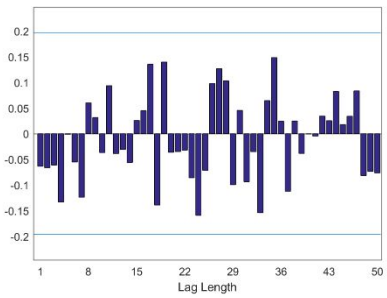


Fig. 11. Autocorrelation function of prediction error for the NN (3-5) model.

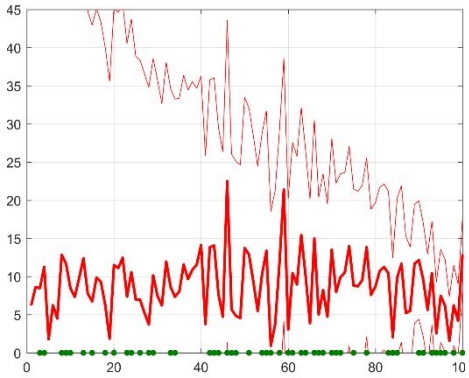


Fig. 12. Confidence interval for the time series prediction by the NN (4-15) model on the training set.

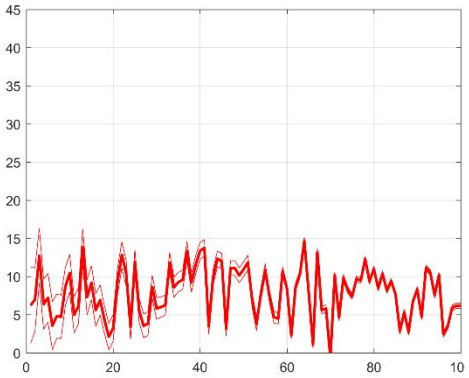


Fig. 13. Confidence interval for the time series prediction by the NN (3-5) model on the training set.

Table 3. Mean square prediction error on the training and test sets.

Subset	NN (4-15) model		NN (3-5) model	
no	training set	test set	training set	test set
1	0.32025	1.34786	0.37799	1.1517
2	0.71700	1.49739	0.56507	1.37707
3	0.63195	3.16212	0.62260	1.56501
4	0.83721	1.99718	1.16928	1.49958
5	0.57766	0.76377	0.65316	0.63076
6	0.41966	0.58519	0.51095	0.56752
7	0.44554	2.07539	0.40103	3.06500
8	1.14641	4.48117	0.94954	2.11953
9	1.06016	0.75750	0.63216	0.76378
10	0.49907	0.96839	0.54018	0.75295
11	0.62024	1.13369	0.79046	1.06503
12	0.75298	1.16708	0.79046	1.08028
average	**0.669011**	**1.661394**	**0.666907**	**1.303184**

Application of virtual leave-one-out procedure allows selection of neural models with the homogeneous influence of the training examples – thus avoiding the overfitting. This method is computationally efficient and can be also used to data set diagnostics by discovering the most influential training examples.

The prediction obtained by the selected neural network model is more accurate (low standard error deviation, low level of confidence interval) as can be stated in Table 3, and Figures 6-13.

It can be concluded that the virtual leave-one-out test is the efficient tool for the optimal selection of neural network predictors of crime time series in the application for crime prediction performed on the real dataset of spatiotemporal crime incidence.

Acknowledgments. The work was funded by the grant DOB-BIO7/05/02/2015 of Polish National Office for Research and Development.

Refrences

1. Sagl, G., Resch, B., Hawelka, B., & Beinat, E.: From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. Proceedings of the GI-Forum, 54-63 (2012)
2. Zheng, Y., Capra, L., Wolfson, O., & Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. on Intelligent Systems and Technology, 5(3),38 (2014)
3. Eck, J., Chainey, S., Cameron, J., Wilson, R.: Mapping crime: Understanding hotspots. Technical report, U.S. Department of Justice (2005)
4. Groff, E. R., & La Vigne, N. G.: Forecasting the future of predictive crime mapping. Crime Prevention Studies, 13, 29-58 (2002)
5. Wang, X., & Brown, D. E.: The spatio-temporal modeling for criminal incidents. Security Informatics, 1(1), 2 (2012)
6. Ratcliffe JH.: A temporal constraint theory to explain opportunity-based spatial offending patterns. Journal of Research in Crime and Delinquency 43(3), 261– 291 (2006)
7. Si, X. S., Wang, W., Hu, C. H., Zhou, D. H.: Remaining useful life estimation–a review on the statistical data driven approaches. European journal of operational research, 213(1), 1-14 (2011)
8. Box, G.E.P, Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, Wiley (2015)
9. Dreyfus, G.: Neural Networks - Methodology and Applications, Springer-Verlag Berlin Heidelberg (2005)
10. Bishop, C. M.: Pattern Recognition and Neural Networks, Clarendon Press (2006)
11. Zhang, G. P.: Time series forecasting using hybrid ARIMA and neural network model. Neurocomputing 30, 159-175 (2003)
12. Monari, G., Dreyfus, G.: Withdrawing an example from the training set: an analytic estimation of its effect on a non-linear parameterised model. Neurocomputing 35, 195-201 (2000)
13. Monari, G., Dreyfus, G.: Local Overfitting Control via Leverages. Neural Computation 14, 1481-1506 (2002)
14. Hastie, T., Tibshirani, R. , Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Springer (2008)
15. Zhang, Y., Yang, Y.: Cross-validation for selecting a model selection procedure. Journal of Econometrics 187, 95-112 (2015)
16. Woodward, W.A., Gray, H.L., Elliot, A.C.: Applied Time Series Analysis with R. CRC Press, Taylor and Francis Group (2017)
17. Jankowski, S., Szymański, Z., Piątkowska-Janko, E.: Semi-supervised Least-Squares Support Vector Classifier based on Virtual Leave One Out Residuals. Proc. of the 9th International Workshop on Pattern Recognition in Information Systems 2009, 71-80 (2009)

Data Mining Applied for Performance Index Prediction in Highway Long Segment Maintenance Contract

Andri IRFAN Rifai[1, a)] Susanti HANDAYANI [2, b)] and Mery LITA [3, c)]

[1] Directorate General Highway, Indonesia
[2] Greater Jakarta Transportation Authority, Ministry of Transportation
[3]Civil Engineering, Universitas Internasional Batam, Indonesia

a) Corresponding author: andrirfan@yahoo.com
b)susanhandayani707@gmail.com
c) merrytse2@gmail.com

Abstract. The dynamics of the national road service level in Indonesia was quite high. There is a significant difference between segments and different areas, even within the same area there is a different variation of level service. These conditions encourage the Directorate General Highway to make up a new concept, by doing a long segment contract that handles national road in a single integrated contract. However, the limited budget and the poor handling of distribution pattern will cause a bad implementation of a long segment contract. Two-objective optimization models consider maximum Performance Index and minimum maintenance cost. The study was conducted on the entire national road network in the Jakarta Metropolitan 1 are paved with the flexible pavement. In the proposed approach, data mining models are used to predicting the performance index over a given period of time. Preventive maintenance is chosen in this study. Multi-objective optimization models were developed based on the Simplex Method. The limited budget and effective targets are the two constraints in the developed models. Based on the R-Tools result, the optimal solutions of the two objective functions are Obtained. From the optimal solutions represented by index performance and cost, an agency more Easily Obtain the information of the maintenance planning. The result of the proposed development models can provide the optimal budget distribution for each segment in a long segment contract.

Keywords: Data Mining, Long Segment, Preventive Maintenance, Performance Index

1 Introduction

The road network is planned, constructed and maintained to facilitate transportation with safe, comfortable, and efficient. To realize these goals, within a recent decade's pavement management system continues to be developed. For example, the United States developed a pavement management system through the American Association of State Highway Officials (AASHO) in the late of 1950s [1]. At that time not only, the

adfa, p. 1, 2011.

developed countries are to develop a pavement management system. Through the help of the World Bank, pavement management systems in the developing countries still continues to be improved by developing the Highway Development and Management (HDM) since 1966. On the development progress, HDM calibrated according to the conditions and standards that apply in each country. Since 1994 until now HDM-4 was developed to respond to the global demands in the road sector which are complex, including pathway safety, environment, and energy, beside to the management aspect [2].

Over time, the purpose of a pavement management system continues to branch out. At the first stage, the government as the organizer of the road continues to increase the functional capacity of the pathway. But currently beside to increase the capacity of the pathway, pathway organizers must meet the higher expectations of road users in the form of comfort, convenience, and security. Related to the organizers of the main road and all stake holders should continue to develop themselves to maintain, expand and improve the performance of the network system of the existing road. One step to achieving this purpose need a better management system so all the existing resources can be optimized. Through the current approaches was assisted by modern mathematics and computer technology, the budget allocation as one of the resources for the improvement of the road pavement management system that can be implemented more efficiently [3].

Pavement management system is done in sustainable, begin from design, planning, construction, operation, maintenance, to control. All stages in the cycle of pavement management system have a role that all was important. Stages of pavement management system have a significant effect in maintaining the performance of the road if done continuously over a long period [4]. It is influenced by the nature and character of the road pavement structure can be patterned with different approaches of data and other historical records.

In its development through the process of testing implementations, Directorate General of Highway (DGH) Indonesia to implement the Long Segment Maintenance Contracts (LSMC) in 2015 as a way to improve the standard of maintenance and replacing the self-management approach based on direct labor applied. These contracts include elements of compensation results is fundamental in the approach of Performance Based Maintenance Contract (PBMC) but the duration is shorter. Therefore, some of the difficulties associated with the contract form like PBMC can be avoided. It is expected with this LSMC, pavement management system in Indonesia is getting better. The performance of the main road became the main parameter in measuring the sustainability of LSMC requires a special attention.

Performance may be reduced in proportion to the increasing age of the pavement and the traffic load [5]. In general, the age of the pavement is determined based on the cumulative equivalent standard axle (CESA) is expected across the road pavement, it was calculated from the start of the road pavement constructed, operated until the pavement is categorized as damaged (end ages of the plan). The decline in overall road performance of the function increase traffic volume and traffic load, changes of environmental conditions, as well as other conditions [6].

Basically, the road pavement structural function can decrease with age. However, the function of pavement on the road network often suffer a structural damage prior to

the age of the plan is achieved, due to various conditions during the operation. One of the phenomena that is common in developing countries is overloaded [7]. The condition occurs continuously without can be prevented. Due to limited modes of transport as well as the purpose of minimizing the transport cost that caused the road damage occurs sooner.

Before the damage reaches a lower layer, an indication of damage initiated by surface damage. The level of flatness and surface roughness change with road performance. The roughness of road is an important indicator because it directly affects the driver and the vehicle. The number or index flatness related to the amplitude and frequency distortion pavement, the suspension characteristics of the vehicle, and the vehicle speed. The condition of the road that is not a good flatness can reduce speed, causing potential damage to the vehicle, increasing the operating costs, and increase exhaust emissions [6].

The decline in road performance did not take place in real time, but gradually follow a function of time and a time series. The speed and shape changes in performance have certain patterns and trends. The collection of large amounts of data, it is necessary to be able to produce a good pattern and continuous [8]. Approach to new techniques and the use of the latest technology is necessary so that a set of data that has been collected through measurement Performance Index (PI) can be utilized in a structured and scalable to support pavement management system better roads through the interpretation and prediction of accurate data.

2 Study Literature

2.1 Long Segment Maintenance Contract

One of the difficulties divert the implementation of routine maintenance to the private sector in the short term is the lack of capacity of road contractors to handle the new load work. To support the contractor to take over the routine maintenance tasks, since the beginning of the LSMC was designed to reduce the risks and the financial exposure for the contractor to limit the term of the contract becomes one to three years; handling and great improvement work is determined by the DGH; using low-risk pricing structure with a combination of lump sum payments and the payments that adjust as PI for routine maintenance (similar to PBMC) and a payment schedule based on the rates for major maintenance work.

This LSMC is a milestone for achieving steady PBMC, by the way doing a new approach to the management of pavement without prejudice to the responsibility of the organizers of the road but changed the focus of the responsibilities organizers of the road radically. In a performance-based contract, the organizers do not need to set any details on how the contractor to achieve the desired results. The Road organizers will be required to be able to define the problem clearly, develop a methodology for determining acceptable performance indicators and measured in accordance with the mission of the organizers of the road, as well as developing objective performance evaluation system. Determination of performance indicators not only require engineering expertise micro multi fields but also can path the achievement of realistic macro indicators

such as implied in the mission of the organizers of the road. The large projects with strong competition, long duration and extension periods, long outsourced road sections that incorporate crack sealing, pothole repair, illumination repair/maintenance, and mowing activities, favor outsourcing under PBC [9].

LMSC requires a culture shift service provider. Technical capability and innovation service providers in order to be competitive. The pattern of construction services business will also change with the increasing integration of the stages of design, construction, operation and maintenance. LSMC also requires a change in the culture of service users, given that most risks can occur due to the behavior of service users. Many assumptions used in the design of roads and bridges being care through various traffic regulation on road transport. One important example is the traffic loading. Disobedience of road users on the rules on Heaviest Loads axis (HLA) will cause uncertainty in the design of the structural strength. As a result, the reliability of the design will be decreased which leads to the risk of premature failure. This kind of risk is borne by the contractor if it would cause a sizable premium and a burden on the budget. This would complicate the handling strategy and budgeting road, given the uncontrolled rule violations have an uncertainty broad space due to loss of control boundaries.

One of the main objectives of this LMSC is the ongoing maintenance activities in order to maintain the conditions, the capacity of the road network services that have been built so that it can meet the needs of users as well as users of the road. Through the work of maintaining the condition of existing roads to be kept in a steady state. Thus, the minimum level of service in accordance with the Minimum Service Standards, also the design life of the road can be met as well as the performance of the road will be restored to the initial condition at the time of construction. In order to maintain the service road, streamline maintenance of roads and ensuring the maintenance costs, we need a contract innovation such as performance-based contract, the expected implementation constraints that would be solved.

2.2 Artificial Intelligence

Interpretation and prediction data is one of the important things in a pavement management system. Big sets of data into information without meaning only if no interpretation and the right and accurately prediction. In connection with this, we need a model that can provide a good approach to the process of interpretation. Data mining (DM) is a widely used approach to the interpretation of data in various disciplines. Through an approach to artificial intelligence (AI), DM has a huge potential to assist in the interpretation and prediction [10]. Utilization of AI in a group of civil engineering science has been done by Terzi [11] to compile a predictive model pavement serviceability index (PSI) and surface distress index (SDI). Furthermore, Zhou [12] to develop models of geographic information system (GIS) road maintenance; the development of predictive models of jet grouting [13]. In literature searches were performed, until now the approach to AI and DM techniques have not been developed for predictive modeling PI on LSMC.

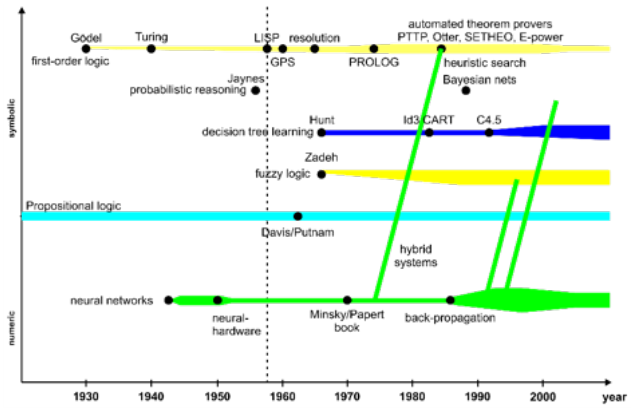


Fig. 1. Development of Artificial Intelligence [13]

Soft computing method performed by imitating the processes that found in nature, such as brain and natural selection [14]. Soft computing techniques allow the processing of data with uncertainly, imprecise and ambiguous. In the early mid-1960s, a new branch of computer science began to attract the attention of much of scientists. This new branch, known as AI, can be defined for the study of how to make computers able to push the quality of people jobs to get better. To achieve these objectives, the computer developed by imitating human behavior. In 1970 AI is more focused on the development of expert systems are designed to support decision-making through opinions of experts computed. Then, in the 1990s there was a shift of AI development, that is studying the various issues directly from the data [15]. Until now AI continues to grow and includes several methods and solutions across a science. In figure 1 we can see the development of AI in various areas of science. AI development began in 1970 is growing, characterized by the melting of numeric and symbolic approaches are complementary.

The development of the information technology industry is very fast, scientific data collection was growing rapidly. Databases in large size is not a problem if it can take advantage of computer technology with a range of major applications and supporters. All data has been collected and stored in a database that can either be a very valuable knowledge that can be used to support making decision and optimization of an action. Classical statistics have limitations to doing the data analysis with a large number or when the function of the complex relationship between the variable data. To overcome these limitations, need to develop tools of computer-based data analysis with greater capabilities and automatic [16]. This field is formally defined as knowledge discovery from databases (KDD). Wang [17] mentions in its development is increasingly recognized by the term KDD DM. Furthermore, in this dissertation, DM terminology is often used as a synonym of KDD.

2.3 Data Mining

Understanding of scientific fields plays an important role in the success of designing algorithms DM. The database was only a set of data without meaning if it is not done with the right algorithm approach [18]. Furthermore, Fu also said that the results of a review conducted in the last few years, the ability of DM growing in a particular domain and depends on the number of researchers who continuously develop specific algorithms. In the simple case, science can help identify the right features for modeling the underlying data compilation and database Scientific knowledge can also help design the business objectives that can be achieved using in-depth analysis of the data base.

One of the steps in developing the model prediction performance of the road pavement management system is processing the road condition data in a process to establish a data mining KDD poverty. DM is combined logically with the knowledge of the data, and statistical analyzes were developed in the knowledge business or a process that uses statistical techniques, mathematics, artificial intelligence, imitation and machine-learning to extract and identify useful information to the associated knowledge of a variety of large databases, In the stages of KDD, DM algorithm equipped with the dataset that used for learing-phase, to be developed into a data-driven models. The model can be described as the relationship between inputs and outputs, which can provide useful information. DM operation in developing the model hereinafter referred to DM task.

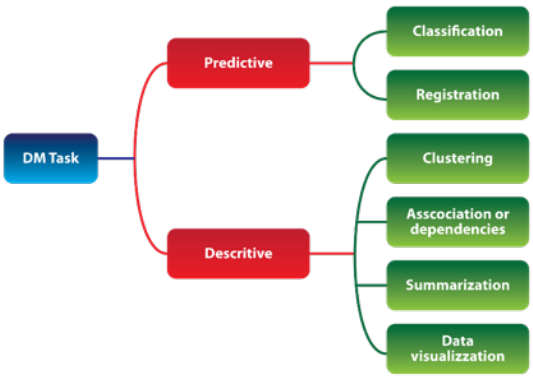


Fig. 2. DM Task

DM task is based on the ability of DM in solving various problems with interpretation and other statistical operations on the data [19]. Depending on the type of pattern is found, DM task usually classified into two categories, predictive and descriptive. Predictive Approach doing inference on the data to predict the values of the unknown variables of output, considering the known values of the input variables [20]. While the descriptive approach to characterize and summarize the general nature of the data in order to improve understanding and provision of information. The ability of DM task depends on the ability for users to do initial identification of a problem and goal completion.

DM classification task is one of the most frequently to used and have the purpose of finding a model that can classify data into each class. Trained model must be able to classify the data into groups based on the attributes of certain data [21]. The model used for classifying usually constructed using a set of supervised learning. Figure 1 summarizes the DM task that is used today. Various task developed in DM largely based statistical approach applicable in general, so it is easy to understand description. Some DM algorithms are mostly used in the classification task on a decision tree, neural networks and support vector machines. The model ability to perform his function to classify with classification metric [22].

2.4 Pavement Maintenance Optimization

Optimization approach in a pavement management system is needed to optimize the limited resources to meet the needs of road pavement maintenance that continues to grow. In a simple understanding, optimization involves a variety of resources to maximize or minimize the objective function of several binary, integer decision variables considering the inequality constraints. Modeling results generated through engine modifications iteration still need improvement so that the number of input variables formed by fuzzy approach is more accurate. Model optimization is not only used in the maintenance of pavement but is also used to get the optimal planning, one of the models that exist today is the integration of genetic algorithm with a geographic information system the way to get the alignment is optimal [23] and [24].

An obstacle to the single-objective function is rare in road pavement management problems. In a pavement management system is precisely the various objectives and constraints to be resolved at the same time. The objective can be achieved more than one and contradictory, so it needs to be optimized simultaneously or by minimizing some objective function. Single-objective optimization approach, the idea of optimization with the goal of minimizing or maximizing a certain objective value [25]. While the approach of Multi-Objective Optimization (MOO) is consists of two or more objective that needs to be optimized. In development MOO approach is more developed, due to various constraints and objectives more dynamic. MOO approach can be used to perform flexible pavement optimization with overload and can be developed with a variety of the others approaches [26].

3 Methodology

The basic principle of LSMC is to maintain the serviceability level of the pavement with the available resources and budget. To obtain a good result, the policy maker efficiently utilizes the existing resources, by optimizing the equipment, materials, personnel, methods, and costs. Moreover, policy maker can also plan a cost-effective and efficient method by considering the priorities and the schedule of routine maintenance, major rehabilitation and or reconstruction. But when faced with the extensive national road network and its constraints, this method of optimizing is considered as a difficult task. A systematic and well-concept effort is needed so that business-process

maintenance management can run smoothly and measurable. The systematic process can be started with the identification of the problem, pavement performance prediction, deterministic formulation, and optimization process itself.

3.1 Performance Index Prediction

One of the basic issue in pavement management system is the development of performance or deterioration prediction models. Several performance prediction models have been proposed over the years, some of which are simple and others more complex. The success of a maintenance management process in LSMC depends on the performance prediction, executed by the system. To enhance the performance of PMS, successful prediction of pavement performance is of primary importance. The researcher will conduct the PI prediction models based on Support Vector Machines (SVM), which are empirical (data-driven) methods, while the occurrence of the fatigue cracking was predicted by a mechanistic-empirical procedure. SVM is one of them popular method approach in Data Mining (DM). DM aims at the extraction of useful knowledge from raw data and it is receiving an increasing attention by the both the research community and industry. Indeed, many case studies suggest that companies are increasingly investigating the potential of DM technology to deliver competitive advantage [27]. The success of the development of the IRI model can be used to use the existing data to support the management of the road network with the Pavement Management System in the road network of West Java [26]. Then, with SVM approach, IRI prediction is implemented to separate the road networks which affected by the normal load and overload. The model shows a great influence of the truck overloads identified on the road network evaluated [26]. DM in R-Project for statistical computing (R-Tools) is an open-source computational environment and high-level language that integrates powerful statistical and graphical features for data. R-Tools adopts a very flexible and object-oriented design [10]. The tool can be easily extended by the creation of packages. PI prediction model was developed from *rminer* library with the inclusion of several variables such as crack, pothole, and rutting.

3.2 Deterministic Formulations.

This study has a multi-objective, namely maximizing level of service as measured by the PI and minimize the cost of maintenance. Further, the model developed by including the variables of overload and life cycle cost in a multi-year budget.

3.3 Performance Index Maximization and Maintenance Cost Minimization

As described in pavement condition section, a smaller value of PI is indicated better road performance, as written in (Eq. 1). In addition to maximizing serviceability level road with a target value of PI biggest, the agency also should minimize the budget that will be used for maintenance costs. In the (Eq. 2), it is described that the budget used for the maintenance of any treatment depends on the unit cost multiplied by the length of the segment roads.

$$\text{Maximize:}\left(\frac{1}{\sum_{i=1}^{n} d_i}\right) * \sum_{i=1}^{n}(d_i * \sum_{t=1}^{m} PI_{it}^{1} * x_{it}) \geq average_{PI} \quad (1)$$

$$\text{Minimize: } \sum_{i=1}^{n} 0 \sum_{t=1}^{m} C_{it} * x_{it} \leq B \quad (2)$$

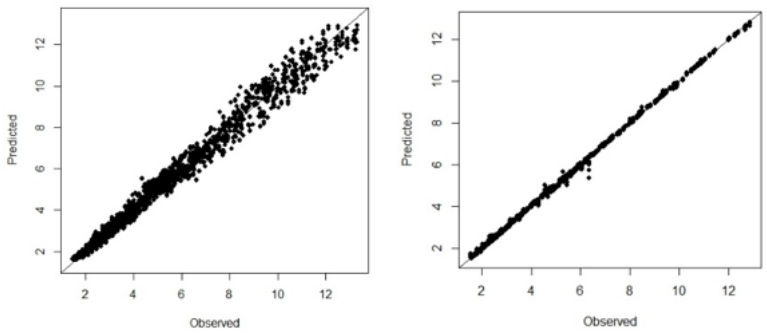
Where
d_i= distance weight parameter to pavement segment i
n= total number of pavement segment of the network
m= total number of pavement management treatment options
t= type of treatment options
PI_{it}^{1}= PI value one year later for treatment t applied to pavement segment i
x_{it}= if treatment t selected to pavement segment i
$average_{PI}$= predefined pavement network average PI level
c_{it}= cost parameter of treatment t selected to pavement segment i
B= budget level for pavement LSMC

4 Experiment and Discussion

As the case study, the national road network in the Metropolitan 1 Jakarta – Indonesia is selected. The road network in Metropolitan 1 has a fairly complete characteristic. The northern part is characterized by the presence of the northern corridor of Java Island that serves as the main transportation lines and this corridor is passed by all types of vehicles. West Java's northern coast line connecting the port city of Jakarta with other cities in Java such as Cirebon, Semarang and Surabaya. In addition, there is south corridor marked by big city transportation character.

4.1 PI Prediction Model

Using DM with SVM model will result in predictive PI value obtained for each road segment on the national road network in Metropolitan 1 Jakarta. In this work, we used the rminer package of the R tool to train the SVM model. For each model, a total of 1.000 runs of a 20 cross validation procedure were applied. The predictive results (measured over unseen data) are shown in terms of observed versus predicted scatterplots. In such scatterplots, the better the predictions, the closer they are to the diagonal line (perfect model). Figure 3 the scatterplots of PI predictive models, revealing a good fit.



a. Learning b. validation stage

Fig. 3. Target PI values versus SVM PI outputs

Figure 3 (a) is some scatterplots showing the results of the learning stage modelling learning stage with total amount of 1100 data, and Figure 3 (b) is an iteration for validation stage. The computed regression error metrics, in terms of the Mean Absolute Deviation (MAD) 0.62 ± 0.01, Root Mean Squared Error (RMSE) 0.72 ± 0.02 and coefficient of determination (R2) 0.89 ± 0.02. The lower the MAD and RMSE values, the better the predictive model, while a perfect model should have an R2 value close to 1.0. The results are presented in terms of the average of the runs and with the respective 95% confidence intervals according to a t-student distribution. Analyzing the results it is clear that a good fit was achieved by the SVMs model.

4.2 The Developed Optimization Model

Stochastic Optimization approach can be used to determine the model of Pareto Solution to obtain the optimization of PI value and the maintenance costs. The post-optimization decision making, or the methods used to choose the final solution are also illustrated by the model application. In this research, the optimization is conducted for various maintenance scenarios. The optimal maintenance programs are selected by using the Pareto approach. Pareto approach is an approach to choosing the pattern of maintenance with the closest distance to the axis 0 (Fig. 4).

In the optimization phase, the maintenance scenario is performed by iteration, utilizing the tools provided by the R-Tools by performing simulations tiered generation. Maintenance Scenario is conducted gradually refers to the DGH standard scenario, in sequence and then combined to achieve the optimum point is called the Pareto optimality and the shortest normalized distance.

In this optimization stages, the scenario maintenance is performed by iterating utilizing the tools provided by R-tools by conducting simulation tiered generation. Scenario maintenance done gradually consecutively then combined to achieve the optimum point is called Pareto optimality. The pattern of Pareto optimization approach is done by transformation with the first generation. This is consistent with Pareto's theory, that

a small percentage (20%) of the causes of the problems giving a potential settlement of the majority (80%) issues. In the optimization of road maintenance, known indicator value is the IRI. Expected by selecting the maintenance scenario that became a group-based index of Pareto, the pattern of maintenance costs being able to read the movement of the overall value of PI.

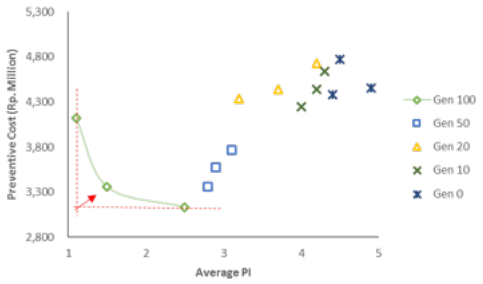


Fig. 4. Pareto Optimality

To simplify the optimization scenario in this research approach road conditions with 4 types of handling LSMC. Maintenance chose to achieve the best PI and use the available budget. With Pareto approach provided by optimx on R, obtained pattern maintenance activities of each segment and the prediction value predicted PI on each segment. Estimated value of PI obtained in LSMC period before (original) and after optimization (Opt1, Opt2, Opt3) can be seen in figure 5.

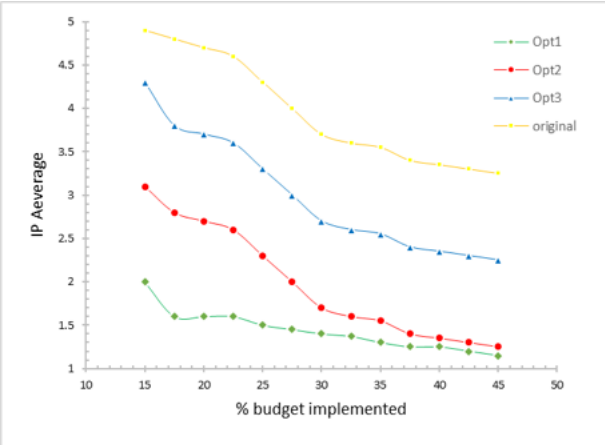


Fig. 5. PI optimization

Simulations carried out by iterating dynamically linked with the performance prediction model part way through iteration SVM models. The second main part is mutually connected and controlled with the subject equation (1) and (2). PI value to be come to a target in the simulation is the average value of PI most optimal road network with due

regard to the minimum limit value PI on each segment. Iteration models show that the necessary steps to achieve this jump.

5 Conclusion

This study developed a model of multi-objective optimization in LSCM to generate an optimal scenario of pavement maintenance. A two-objective optimization model considers maximum PI and minimum maintenance cost. Both of these objectives are considered to be achieved simultaneously. Constraints faced is road deterioration that can accelerate the decline in the level of PI. Through the DM approach to obtain predicted PI, maintenance optimization is then performed by maintenance type that is received in each group of highway networks. The results showed the load factor group with preventive pavement maintenance scenario produces the most optimal financing.

6 Reference

1. Akofio-Sowah, M. A. and A. A. Kennedy. "A Critical Review of Performance-Based Transportation Asset Management in United States Transportation Policy." Proceedings of the 17th IRF World Meeting & Exhibition. International Road Federation, 2014.
2. Martinaz Diaz, M. and I. Perez. "Mechanistic-empirical pavement design guide: features and distinctive elements." Journal of Construction (2015): 32-40.
3. Santos, J. and A. Ferreira. "Pavement design optimization considering costs and M&R interventions." Social and Behavioral Sciences 53 (2012): 1184 – 1193.
4. Ding, T., L. Sun and J. Chen. "Optimal Strategy of Pavement Preventive Maintenance Considering Life-Cycle Cost Analysis." Social and Behavioral Sciences 96 (2013): 1679 – 1685.
5. Yuhong Wang, P.E, et al. "Evolution and locational variation of asphalt binder aging in long-life hot-mix asphalt pavements." Construction and Building Materials 68 (2014): 172–182.
6. Hass, R. "Good Technical Foundations Are Essential for Successful Pavement Management." Maintenance and Rehabilitation of Pavements and Technology Control. Guimaraes: Universidade do Minho-Escola de Engenharia, 2003. 3-28.
7. Ede, A. Nkem. "Cumulative Damage Effects of Truck Overloads on Nigerian Road Pavement." International Journal of Civil & Environmental Engineering IJCEE-IJENS Vol: 14 No: 01 (2014): 21-26.
8. Varela-González, M., et al. "A semi-automatic processing and visualisation tool for ground-penetrating radar pavement thickness data." Automation in Construction 45 (2014): 42-49.
9. P. Ch. Anastasopoulos, B. G. McCullouch, K. Gkritza, F. L. Mannering, and K. C. Sinha, "Cost Savings Analysis of Performance-Based Contracts for Highway Maintenance Operations", Journal of Infrastructure Systems Dec 2010, Vol. 16, No. 4, pp. 251-263
10. Cortez, P. "Data mining with neural networks and support vector machines using the r/rminer tool." Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining, 83,. Berlin, Germany: J In P. Pemer, editor,, 2010.
11. Terzi, S. "Modeling the Pavement Present Serviceability Index of Flexible Highway Pavements Using Data Mining." Journal of Applied Sciences (2006): 193-197.
12. Zhou, G., et al. "Integration of GIS and Data Mining Technology to Enhance the Pavement Management Decision Making." Journal of Transportation Engineering (2010): 332-341
13. Ertel, W. (2009). Introduction to Artificial Intelligence. Springer, ISBN 9780857292988.

14. Tinoco, J., A., G. Correia and P. Cortez. " Support vector machines applied to uniaxial compressive strength prediction of jet grouting columns." Computers and Geotechnics 55 hal. 132–140 (2014): 132-140.
15. Liao, S, P Chu and P Hsiao. "Data mining techniques and applications – A decade review from 2000 to 2011." Expert Systems with Applications 39 (2012): 11303–11311.
16. Rahman, F. A., et al. "Knowledge Discovery Database (KDD)-Data Mining Application in Transportation." Proceeding of the Electrical Engineering Computer Science and Informatics, 1(1). 2014. 116-119.
17. Wang, X. Z. Data mining and knowledge discovery for process monitoring and control. Springer Science & Business Media., 2012.
18. Fu, T. C. "A review on time series data mining." Engineering Applications of Artificial Intelligence, 24(1), (2011): 164-181.
19. Freitas, A. A. Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media, 2013.
20. Wu, X., et al. "Data mining with big data. Knowledge and Data Engineering." , IEEE Transactions on, 26(1) (2014): 97-107
21. Kantardzic, M. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011
22. Mukhopadhyay, A., et al. "A survey of multiobjective evolutionary algorithms for data mining: Part I." Evolutionary Computation, IEEE Transactions on, 18(1) (2014): 4-19.
23. Kang, Min-Wook, Manoj K. Jha and Paul Schonfeld. "Applicability of highway alignment optimization models." Transportation Research (2012): 257-286.
24. Beg, M. A. and A. Banerjee. "Developing Optimized Maintenance Work Programs for an Urban Roadway Network using Pavement Management System." 9th International Conference on Managing Pavement Assets. 2015.
25. Di Mino, G., et al. "An Advanced Pavement Management System based on a Genetic Algorithm for a Motorway Network." Proceedings of the Third International Conference on Soft Computing Technology in Civil, Structural and Environmental Engineering. Stirlingshire, Scotland. : Civil-Comp Press, 2013. 26.
26. Rifai, A., Hadiwardoyo, S. P., Correia, A. G., & Pereira, P. Genetic Algorithm Applied for Optimization of Pavement Maintenance under OverloadTraffic: Case Study Indonesia National Highway. Applied Mechanics and Materials ISSN: 1662-7482, Vol. 845, Trans Tech Publications, Switzerland, 2016 pp 369-378
27. Rifai, A. I., Hadiwardoyo, S. P., Correia, A. G., Pereira, P., & Cortez, P. Data Mining Applied for The Prediction of Highway Roughness under Overloaded Traffic. International Journal of Technology 5, (2015) pp 751-761

Novel order patterns recurrence plot-based quantification measures to unveil deterministic dynamics from stochastic processes

Shuixiu Lu[1,2], Sebastian Oberst[1⋆], Guoqiang Zhang[1], and Zongwei Luo[2]

[1] Centre for Audio, Acoustics and Vibration (CAAV),
University of Technology Sydney, NSW 2007, Australia
[2] Department of Computer Science and Engineering,
Southern University of Science and Technology, China

Abstract. Forbidden ordinal patterns are known to be useful to discriminate between chaotic and stochastic systems. However, while uncorrelated noise can be separated from deterministic signals using forbidden ordinal patterns, correlated noise exhibits apparently forbidden ordinal patterns, which can impede distinguishing between a chaotic system from correlated noise. We introduce order patterns recurrence plots to visualise the difference among chaotic systems, uncorrelated and correlated noise. Our numerical results show that an optimal embedding dimension in the plot preserves the diagonal lines of a chaotic system. While the uncorrelated noise shows up as thinly isolated points in the plot, the correlated noise forms clusters. We propose two measures, the mean and the median of relative frequencies of order patterns that appear in a time series to characterise the nonlinear dynamics. The effectiveness of the two measures are analysed through the bifurcation diagrams of the logistic map, the tent map, the delayed logistic map and the Hénon map. Both, the mean and the median can distinguish chaos from quasiperiodicity in the delayed logistic map. Our results show that the mean of relative frequencies is a reciprocal of the number of order patterns that occur in a given time series, and thus can be a measure of forbidden ordinal patterns. While the mean is robust to the change of parameters in the bifurcation diagrams, the median is sensitive to the change, which has significance on the characterisation of nonlinear dynamics.

Keywords: forbidden ordinal patterns, noise, optimal embedding dimension

1 Introduction

Given an irregular time series, it is often necessary to distinguish between stochastic dynamics (random noise, high dimensional) and chaotic dynamics (deterministic, here assumed to be of low dimension). However, both the stochastic system

⋆ sebastian.oberst@uts.edu.au

and the chaotic system generate a broadband power spectrum in the spectral domain, which often hinders easy distinctions between the two [25]. The situation gets extremely complicated when it comes to real-life data since natural information is inextricably contaminated by omnipresent dynamical noise [16, 21, 25].

To discriminate chaotic systems with observational noise from stochastic systems, Bandt and Pompe [6] develop a symbolisation scheme, Bandt-Pompe(BP) methodology, which encodes a time series to order patterns. An order pattern with embedding dimension m is a permutation of the set $\{0, 1, ..., m-1\}$. The scheme exhibits an invariant under the process of monotonous transformations, and is robust with regards to observational Gaussian noise [6]. Owing to this robustness, order patterns have attracted a growing interest [7, 15, 23, 24, 33].

To quantify order patterns and their degree of determinism, many studies focus on two streams of measures that are related to the BP method. The first stream is based on observations by making use of *forbidden ordinal patterns* [3, 4, 27, 28, 31, 32, 34]. Forbidden ordinal patterns are order patterns (permutations) that persistently do not occur in a time series. For a deterministic map $x_{i+1} = f(x_i)$, the intersections of $f^0(x) = x$,$f^1(x)$,...,$f^{m-1}(x)$ determines the number of order patterns. When the number of intersections can't outgrow $m!$ as m is large enough [4], forbidden ordinal patterns occur. For the logistic map ($x_{i+1} = 4x_i(1-x_i)$), as $m = 3$, the number of intersections of f^0, f^1, f^2 is 6 and hence the number of order patterns is 5. The permutation 210 never occurs whatever long the time series is [4]. The interested reader is directed to [4]. However, a stochastic system has no persistently forbidden ordinal patterns [2, 4, 9, 16, 26]. Also, the number of forbidden ordinal patterns grows exponentially as m increases [4, 5, 27], yet there is a minimal length to detect the outgrowth of forbidden ordinal patterns [27]. In addition, the number of forbidden ordinal patterns is robust to low degrees of sampling irregularities where the time series is from irregular time intervals [15]. To visualise forbidden ordinal patterns, Kulp and Smith [14], Kulp and Zunino [16] introduce permutation spectrum test, which counts the frequencies of each permutation of $\{0, 1, ..., m-1\}$ and plots the count of each permutation. A permutation that counts 0 is a forbidden ordinal pattern and indicates determinism. However, correlated noise (i.e., fractional Brown motion and pink noise) can show apparently forbidden ordinal patterns when the length of a time series is not long enough [9, 26]. As a result, permutation entropy, as information entropy of order patterns, is suggested for effective identification of determinism from the count of each permutation [22, 26, 27].

The abovementioned literature focuses on the properties of order patterns that do not occur in a times series. In contrast, the second stream is to visualise and estimate recurrent behaviour of order patterns that appear in a time series, using order patterns recurrence plot (OPRP). OPRP was developed by Groth [12] to visualise the dependencies between two time series. Different from a conventional recurrence plot (RP), which is based on the spatial closeness to define recurrent behaviour of two points in the time series on the same trajectory, OPRPs allow studying the dependency between two time series [12, 19] and nonlinear dynamics of the system [8] using the recurrence of the same order

patterns. Hence, the quantifications of recurrence plots of RP can be applied to the OPRP. However, literature is scarce on interpreting recurrence quantification analysis (RQA) measures extracted from an OPRP [8]. Schinkel et al. [29] and Marwan et al. [18] employ the RQA to detect determinism of event-related potential, which is a measure of brain response via electroencephalography. Donner et al. [10] used RQA to investigate short-term dynamics of discrete-valued data. McCullough et al. [20] regenerate a time series using ordinal network and compare the dynamics of the original time series and a surrogate time series designed from ordinal network using the RQA and OPRP. Recently, different from Groth [12] where the order patterns are based on ordering the elements in a time series, Caballero-Pintado et al. [8] provides a new OPRP in terms of symbolic correlation integral, which is an order pattern based on the elements in a time series but orders the time sequence rather than the elements in a conventional OPRP. Lu et al. [17] applies the OPRP to uncover determinism buried in noise. While Groth [12] uses the OPRP to examine the dependencies of two time series, Caballero-Pintado et al. [8] applies an OPRP to a time series and study the change in dynamics. However, all studies on RQA are unrelated to the forbidden ordinal patterns.

Our goal is to connect these two streams of studies and show that OPRP can visualise the effect of exponential increase of forbidden ordinal patterns through the increase of embedding dimension m in an OPRP. Also, we investigate how the correlated noise having forbidden ordinal patterns influence the result of RQA extracted from the OPRP.

To unveil the change of dynamics as parameters change, we introduce and validate two new measures related to forbidden ordinal patterns, the mean and the median of the relative frequencies of order patterns that occur in a time series, as quantification measures using white, pink and Brown noise. As test models, we apply the logistic map, the delayed logistic map, the tent map and the Hénon map to examine the evolution of the two measures.

2 Methodology

Underlying a one-dimensional time series ($\{x_t\}_{t=1}^N$) with length N, order patterns of $\{x_t\}_{t=1}^N$ depends on the time delay τ and embedding dimension m [6]. We encode x_t to its order pattern through ordering the vector $\overrightarrow{x(t)} = (x_t, x_{t+\tau}, ..., x_{t+(m-1)\tau})$ [12]. Here, we study order patterns in the case of $\tau = 1$, and then rank $\overrightarrow{x(t)} = (x_t, x_{t+1}, ..., x_{t+(m-1)})$, where 0 corresponds to the smallest element in $\overrightarrow{x(t)}$. To guarantee the uniqueness of order patterns, when $x_{t+i} = x_{t+j}, i < j$, we assume $s_i < s_j$, where s_i is the rank of x_{t+i} in $\overrightarrow{x(t)}$. As a result, order patterns of $\{x_t\}_{t=1}^N$ are permutations of $\{0, 1, ..., m-1\}$, and the length of order patterns equals to the embedding dimension m. Also, since m consecutive elements define an order pattern, the number of well-defined order patterns in $\{x_t\}_{t=1}^N$ is $N - m + 1$.

An order pattern of $\{4, 3, 7, 5\}$ with $m = 4$ corresponds to a permutation 1032 since 7 is the largest number and symbolised by 3.

2.1 Order patterns recurrence plot

Given a time series, an OPRP [12, 19] visualises its order patterns and analyses recurrent order structures via the following recurrence matrix

$$R_{i,j}(m) = \begin{cases} 1, \pi_i = \pi_j \\ 0, \pi_i \neq \pi_j \end{cases} \quad i, j = 1, 2, ..., N - m + 1 \tag{1}$$

where π_i and π_j are the order pattern of the i^{th} and j^{th} element of $\{x_t\}_{t=1}^N$,respectively. m is the embedding dimension.

2.2 Forbidden ordinal patterns

Let $\Pi_m = \{\pi_{\xi_1}, ..., \pi_{\xi_n}\}$ be the set of order patterns that occur in $\{x_t\}_{t=1}^N$ with embedding dimension m and C_i be the frequency of $\pi_{\xi_i} \in \Pi_m$, then

$$\sum_{i=1}^{n} C_i = N - m + 1 \tag{2}$$

where n is the number of order patterns that appear in $\{x_t\}_{t=1}^N$. The maximal value of n is $m!$ when $N \geq m! + m - 1$, that is, all possible order patterns can appear and the number of forbidden ordinal patterns is zero. If the time series is long enough to show every possible order patterns ($N \gg m!$) and $n < m!$, then the number of forbidden ordinal patterns (FOP) is

$$\text{FOP} = m! - n \tag{3}$$

And the relative frequency (RF) of π_{ξ_i} is provided by

$$\text{RF}_{\pi_{\xi_i}} = \frac{C_i}{N - m + 1} \tag{4}$$

The mean of relative frequencies (MRF) is then given by

$$\text{MRF} = \sum_{i=1}^{n} \text{RF}_{\pi_{\xi_i}}/n = \sum_{i=1}^{n} \frac{C_i}{n(N - m + 1)} = \frac{1}{n} \tag{5}$$

Gottwald and Melbourne [11] took the median value of their measures derived from many experiments to reduce the influence of outliers, and Kulp and Zunino [16] showed that while a few order patterns of the fractional Brown noise have high frequencies, most of order patterns have low frequencies. Motivated by Gottwald and Melbourne [11] and Kulp and Zunino [16], we investigate the median of relative frequencies to envisage the evolution of the systems without

the influence of outliers. The median of relative frequencies (MDRF) can be expressed as

$$\mathrm{MDRF} = \mathrm{median}\{\mathrm{RF}_{\pi_{\xi_1}}, ..., \mathrm{RF}_{\pi_{\xi_n}}\} = \frac{\mathrm{median}\{C_1, ..., C_n\}}{N - m + 1} \tag{6}$$

Bandt and Pompe [6] examined permutation entropy (information entropy of order patterns) through the negative value of the logarithm of a probability of permutations. Consistent with their approach [6], we take the negative value of the logarithm of the mean and the median of relative frequencies to measure the mean and the median of relative frequencies of order patterns that occur in a given time series.

$$\mathrm{mean}^l = -\log \mathrm{MRF} = \log n \tag{7}$$

$$\mathrm{median}^l = -\log \mathrm{MDRF} = \log(N - m + 1) - \log(\mathrm{median}\{C_1, ..., C_n\}) \tag{8}$$

where mean^l and median^l are the negative value of the logarithm of the mean and the median of relative frequencies, respectively.

3 Results

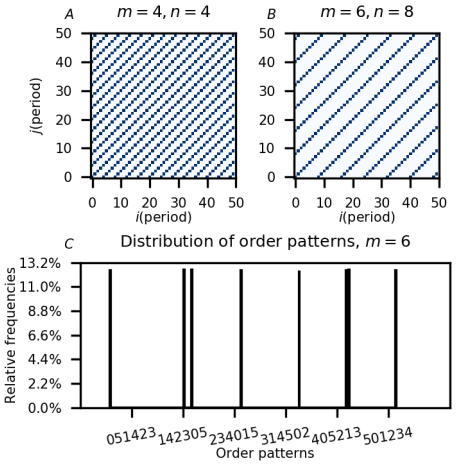
3.1 Order patterns recurrence plot

For Gaussian noise (Fig I3), the increase of m makes the OPRP sparse. In Fig I3 (B,C), we use Gaussian noise and see that the OPRP is denser for $m = 4$ than $m = 6$. However, the OPRP of Brown noise (Fig I4) preserves many points indicating that the system has the order pattern of which the relative frequency is high. For the periodic or the chaotic system under study (Fig I1 and Fig I2), the OPRPs show distinctly more diagonal lines than the cases of Gaussian or Brown noise. As highlighted by Rosso et al. [27], a minimal embedding dimension allows the occurrence of forbidden ordinal patterns and their exponential growth. An OPRP can then have an optimal embedding dimension that rules out Gaussian noise.

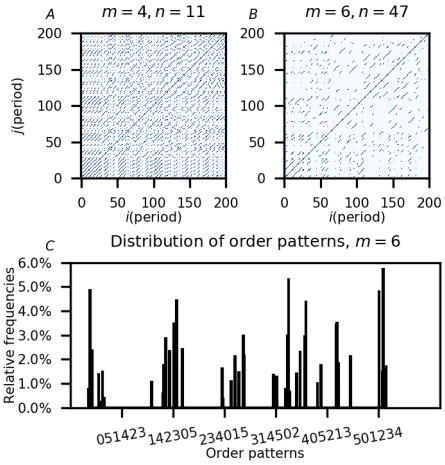
The reasons for having a sparse OPRP in case of Gaussian noise is that no forbidden ordinal patterns exist [4] and that the relative frequencies of every order pattern is close [1]. Therefore as m increases (Fig I3(C)), the number of order patterns that occurs is up to $m!$, and thus the probability of an order pattern that recurs in a given time window is $\frac{1}{m!}$, which is low for large m, leading to a spare distribution of points in a given time window.

However, for Brown noise, in a given time window, the time series shows an upward (downward) tendency (Fig II(D)). As the upward (downward) tendency lasts for six time intervals, it forms the order pattern 012345 (543210). If the tendency lasts for over six time intervals, the order pattern does not change and forms a cluster in the OPRP (Fig I4(C)).

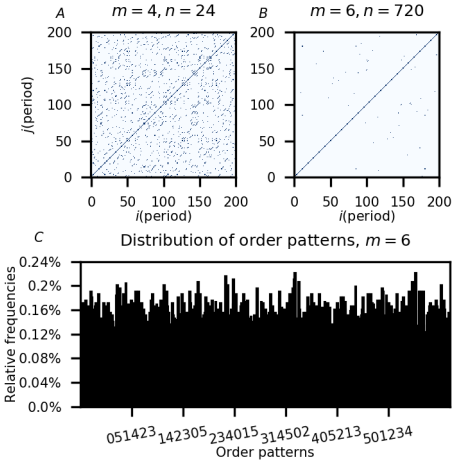
Due to the tendencies in the time series, Brown noise shows high relative frequencies (over 4%) of order patterns (012345 and 543210) (Fig I4(C)), compared with Gaussian noise (below 0.24%) (Fig I3(B)). As consistent with Rosso et al. [26], Brown noise shows apparently forbidden ordinal patterns ($n = 710 < 6! = 720$) (Fig I4(B-C)).



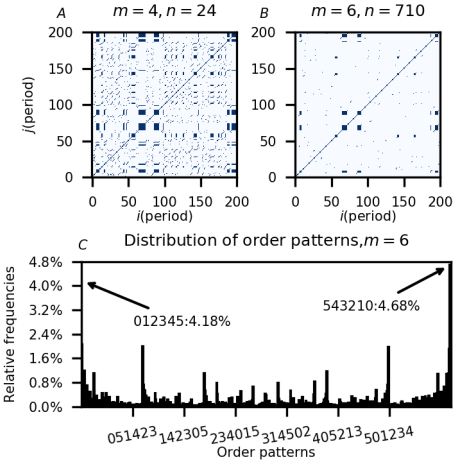
(1) OPRP of a periodic system



(2) OPRP of a chaotic system



(3) OPRP of Gaussian noise($N(0,1)$)



(4) OPRP of Brown noise

Fig. I: OPRPs as m increases.(1) logistic map ($x_i = 3.55x_{i-1}(1 - x_{i-1})$);(2) logistic map ($x_i = 3.91x_{i-1}(1 - x_{i-1})$);(3) Gaussian noise;(4) Brown noise.

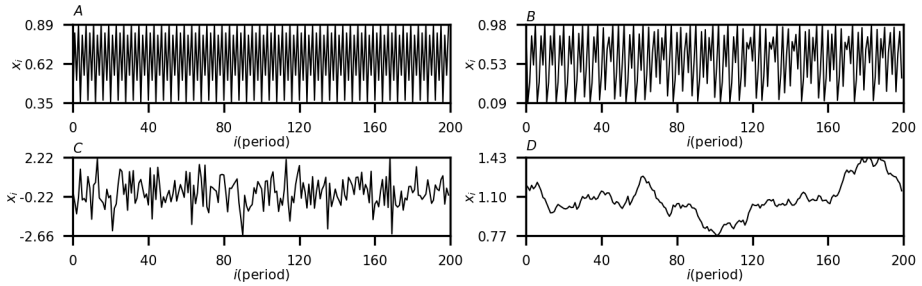


Fig. II: Time series. (A) logistic map (periodic,$r = 3.55$); (B) logistic map (chaotic,$r = 3.91$); (C) Gaussian noise;(D) Brown noise.

3.2 Bifurcation diagrams

Caballero-Pintado et al. [8] showed that RQA in the OPRP can detect the change in dynamics. Now, we show how the mean^l and median^l unveil a change in dynamics and are able to uncover details in bifurcation diagrams. In this section, we study the behaviour of the mean^l and median^l for $m = 6$.

Fig III shows that the change of mean^l is like a staircase as bifurcation parameter changes. That is, the number of forbidden ordinal patterns is robust to the change in dynamics, which is complementary to the results that forbidden ordinal patterns are robust to irregular sampling time [15] and the order patterns that occur is robust against Gaussian noise [1].

From Fig III1B, Fig III2B, Fig III3B and Fig III4B, we find that median^l is sensitive to a change in dynamics. In a periodic window, the median^l is close or equal to the mean^l. However, in a chaotic window, the difference between median^l and mean^l is large and visible. Also, the median^l shows an abruptly decrease to fit with the mean^l in a periodic point or window. In contrast, the abruptly increase of the mean^l lies at the points where the system switch between chaotic and periodic dynamics.

In a periodic window, the length of a time series may cause the number of some order patterns to become higher by one than the remaining order patterns. For example, in our numerical experiments, $N = 20000, m = 6$, for the logistic map with $r = 3.55$, eight order patterns appear, so that the number of the eight order patterns is $N - m + 1 = 19995$. However, 19995 cannot be divided by 8, so the count of some order patterns is 2500, the others is 2499, resulting in this difference of unity between the mean and the median of $\{C_1, \cdots, C_n\}$. For the same length of the time series, as the periodicity increases, this difference is narrowed. However, the difference may be visualised in a low periodic window such as the period-3 window. This is possible since a low periodicity leads to a high relative frequency for each order pattern. The unit difference of low periodicity induces a higher difference between the median^l and the median^l than that of a high periodicity (Fig III1B).

Whereas in a periodic window the order patterns are evenly distributed without consideration of the unit difference, in a chaotic window, the order patterns are not necessary to be evenly distributed [16]. An unevenly distribution of order patterns that occurs can result in the difference of median^l and mean^l. Fig III1 and Fig III2 show that the mean^l changes step by step even in a chaotic region. Therefore, our results support that the number of forbidden ordinal patterns show some degree of robustness to the changes of dynamics since the mean^l is related to the number of forbidden ordinal patterns.

As shown by Sprott [30], for the delayed logistic map, it is difficult to distinguish chaos from quasiperiodicity from the bifurcation diagram. However, fig III3B shows different structures of the change of the median^l and mean^l under the two kinds of dynamics. In the window of quasiperiodicity (see fig III3A), where the maximal Lyapunov exponents remain at zero [30], the mean^l as well as the median^l show no perturbation. Also, different from periodic windows, the evolution of the median^l does not fit with that of the mean^l, showing a different quality in changes.

4 Conclusion

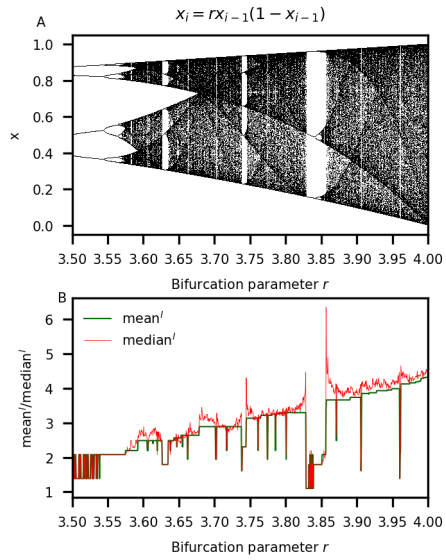
Monitoring the embedding dimension can be used to qualify chaotic versus stochastic dynamics in an OPRP. An optimally embedded OPRP preserves diagonal lines for a chaotic system, leads in the case of uncorrelated noise to thinly distributed isolated points and correlated noise to small clusters.

We illustrate that the mean of the relative frequency of order pattern, which occurs in a time series and accounts for the OPRP, is equal to the reciprocal value of the number of order patterns that show up. Both, the set that contains the appearing order patterns and the set that includes forbidden ordinal patterns are complementary. Therefore, we conclude that for the analytical systems tested the mean^l as average number of appearing order patterns, is an effective measure of the number of forbidden order patterns. However, whether this holds true for any system needs to be verified in the future.

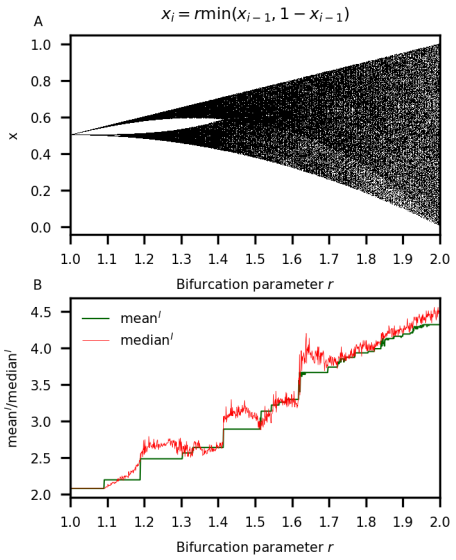
While the mean of the relative frequencies of order patterns that occur is robust to changing bifurcation parameters, the median of the relative frequencies of order patterns that occurs is sensitive to the change in dynamics. Using the difference between the mean^l and the median^l and their fluctuations, we can distinguish chaos from quasiperiodic dynamics. That is, in a quasiperiodic window, the mean^l has no perturbation, and the median^l changes abruptly and then makes a smooth change, not showing evident perturbations.

From a practical consideration, the calculation of the median^l and the mean^l is easily accomplished and computationally inexpensive, which makes the median^l and the mean^l a cost effective analysis tool to characterise the underlying dynamics of the system.

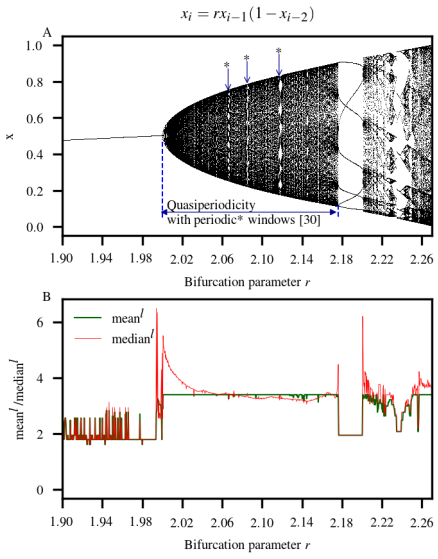
Since the RFs of each order pattern of Gaussian noise are close in value to each other, the mean^l and the median^l of the deterministic signals, which are buried in Gaussian noise, are likely to be different from that of pure noise. In a



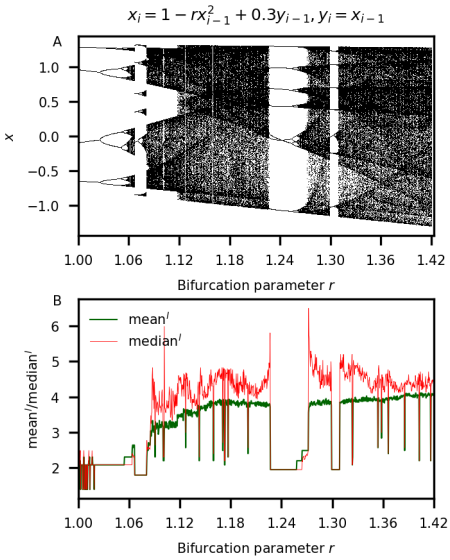
(1) Logistic map



(2) Tent map



(3) Delayed logistic map



(4) x component of Hénon map

Fig. III: Bifurcation diagrams and the evolution of the mean^l and the median^l

real application, one may use the two measures to test the null hypothesis that the system oscillates due to Gaussian noise alone. However, the two measures may fail to detect the change of dynamics if used to identify chaotic systems which are contaminated by noise.

In the future, several directions are worth pursuing. Firstly, we will analyse the sensitivity of the mean^l and the median^l to correlated noise and we will investigate the use of the two measures to unveil determinism buried in correlated noise. Correlated noise exhibits apparently forbidden ordinal patterns, which hinders us effectively discriminating between chaotic systems and stochastic systems. A second direction is to identify the tipping points in the change of dynamics, especially for systems undergoing boarder collision bifurcations(BCB), and examine their robustness with regards to forbidden ordinal patterns and changing dynamics. It is common that a real system has piecewise smooth functions that can induce BCB, as evidenced in dynamic pricing processes [13].

References

1. Amigó, J., Zambrano, S., Sanjuán, M.A.: Combinatorial detection of determinism in noisy time series. EPL (Europhysics Letters) 83(6), 60005 (2008)
2. Amigó, J.: Permutation complexity in dynamical systems. Springer-Verlag, Berlin (2010)
3. Amigó, J.M., Kocarev, L., Szczepanski, J.: Order patterns and chaos. Physics Letters A 355(1), 27–31 (2006)
4. Amigó, J.M., Zambrano, S., Sanjuán, M.A.: True and false forbidden patterns in deterministic and random dynamics. EPL (Europhysics Letters) 79(5), 50001 (2007)
5. Amigó, J.M., Zambrano, S., Sanjuán, M.A.: Detecting determinism in time series with ordinal patterns: a comparative study. International Journal of Bifurcation and Chaos 20(09), 2915–2924 (2010)
6. Bandt, C., Pompe, B.: Permutation entropy: a natural complexity measure for time series. Physical Review Letters 88(17), 174102 (2002)
7. Barreiro, M., Marti, A.C., Masoller, C.: Inferring long memory processes in the climate network via ordinal pattern analysis. Chaos: An Interdisciplinary Journal of Nonlinear Science 21(1), 013101 (2011)
8. Caballero-Pintado, M.V., Matilla-García, M., Ruiz Marín, M.: Symbolic recurrence plots to analyze dynamical systems. Chaos: An Interdisciplinary Journal of Nonlinear Science 28(6), 063112 (2018)
9. Carpi, L.C., Saco, P.M., Rosso, O.: Missing ordinal patterns in correlated noises. Physica A: Statistical Mechanics and its Applications 389(10), 2020–2029 (2010)
10. Donner, R., Hinrichs, U., Scholz-Reiter, B.: Symbolic recurrence plots: A new quantitative framework for performance analysis of manufacturing networks. The European Physical Journal Special Topics 164(1), 85–104 (2008)
11. Gottwald, G.A., Melbourne, I.: Testing for chaos in deterministic systems with noise. Physica D: Nonlinear Phenomena 212(1), 100 – 110 (2005)
12. Groth, A.: Visualization of coupling in time series by order recurrence plots. Physical Review E 72(4), 046220 (2005)
13. Hu, Z., Chen, X., Hu, P.: Dynamic pricing with gain-seeking reference price effects. Operations Research 64(1), 150–157 (2016)

14. Kulp, C.W., Smith, S.: Characterization of noisy symbolic time series. Physical Review E 83(2), 026201 (2011)
15. Kulp, C., Chobot, J., Niskala, B., Needhammer, C.: Using forbidden ordinal patterns to detect determinism in irregularly sampled time series. Chaos: An Interdisciplinary Journal of Nonlinear Science 26(2), 023107 (2016)
16. Kulp, C., Zunino, L.: Discriminating chaotic and stochastic dynamics through the permutation spectrum test. Chaos: An Interdisciplinary Journal of Nonlinear Science 24(3), 033116 (2014)
17. Lu, S., Luo, Z., Zhang, G., Oberst, S.: Order pattern recurrence plots: unveiling determinism buried in noise. In: University of Technology Sydney, FEIT Research Showcase. Sydney, NSW, Australia (14 Jun 2018)
18. Marwan, N., Groth, A., Kurths, J.: Quantification of Order Patterns Recurrence Plots of Event Related Potentials. Chaos and Complexity Letters 2, 301–314 (2007)
19. Marwan, N., Romano, M.C., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. Physics Reports 438(5-6), 237–329 (2007)
20. McCullough, M., Sakellariou, K., Stemler, T., Small, M.: Regenerating time series from ordinal networks. Chaos: An Interdisciplinary Journal of Nonlinear Science 27(3), 035814 (2017)
21. Oberst, S.: Nonlinear dynamics: Towards a paradigm change via evidence-based complex dynamics modelling. In: NOVEM 2018. Ibiza, Spain (7-9 May 2018)
22. Olivares, F., Plastino, A., Rosso, O.A.: Contrasting chaos with noise via local versus global information quantifiers. Physics Letters A 376(19), 1577 – 1583 (2012)
23. Parlitz, U., Berg, S., Luther, S., Schirdewan, A., Kurths, J., Wessel, N.: Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics. Computers in Biology and Medicine 42(3), 319 – 327 (2012)
24. Porfiri, M., Marín, M.R.: Symbolic dynamics of animal interaction. Journal of theoretical biology 435, 145–156 (2017)
25. Rosso, O.A., Larrondo, H.A., Martin, M.T., Plastino, A., Fuentes, M.A.: Distinguishing noise from chaos. Physical Review Letters 99(15), 154102 (2007)
26. Rosso, O.A., Carpi, L.C., Saco, P.M., Ravetti, M.G., Plastino, A., Larrondo, H.A.: Causality and the entropycomplexity plane: Robustness and missing ordinal patterns. Physica A: Statistical Mechanics and its Applications 391(1), 42 – 55 (2012)
27. Rosso, O.A., Olivares, F., Zunino, L., De Micco, L., Aquino, A.L., Plastino, A., Larrondo, H.A.: Characterization of chaotic maps using the permutation bandt-pompe probability distribution. The European Physical Journal B 86(4), 116 (2013)
28. Schindler, K., Gast, H., Stieglitz, L., Stibal, A., Hauf, M., Wiest, R., Mariani, L., Rummel, C.: Forbidden ordinal patterns of periictal intracranial eeg indicate deterministic dynamics in human epileptic seizures. Epilepsia 52(10), 1771–1780 (2011)
29. Schinkel, S., Marwan, N., Kurths, J.: Order patterns recurrence plots in the analysis of erp data. Cognitive neurodynamics 1(4), 317–325 (2007)
30. Sprott, J.C.: Chaos and time-series analysis, vol. 69. Oxford: Oxford University Press (2003)
31. Zanin, M.: Forbidden patterns in financial time series. Chaos: An Interdisciplinary Journal of Nonlinear Science 18(1), 013119 (2008)
32. Zanin, M., Zunino, L., Rosso, O.A., Papo, D.: Permutation entropy and its main biomedical and econophysics applications: a review. Entropy 14(8), 1553–1577 (2012)
33. Zhang, J., Zhou, J., Tang, M., Guo, H., Small, M., Zou, Y.: Constructing ordinal partition transition networks from multivariate time series. Scientific reports 7(1), 7795 (2017)

34. Zunino, L., Zanin, M., Tabak, B.M., Pérez, D.G., Rosso, O.A.: Forbidden patterns, permutation entropy and stock market inefficiency. Physica A: Statistical Mechanics and its Applications 388(14), 2854–2864 (2009)

On Statistical Inference for Independent Colored Sources Analysis

Rachel C. Nethery[1] and Young Truong[2]

[1] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Harvard University

[2] Department of Biostatistics, Gillings School of Global Public Health,
The University of North Carolina at Chapel Hill

Abstract. Blind source separation (BSS) is usually carried out by independent component analysis (ICA) which in turn utilizes only the marginal information. Motivated by many medical brain imaging data analysis and signal processing problems, this paper considers an approach by exploiting the temporal or spatial correlation structure. Both temporal and spatial ICA will be addressed using the R package `coloredICA`. The main objective of the present paper is on the topic of statistical inference for extracting independent colored sources (ICS). Some bootstrap method will be highlighted and a new method based on stratification or partition will be described. The usefulness of this method will be illustrated by conducting a comparison with `fastICA`.

Keywords. Independent component analysis, blind source separation, power spectrum, maximum likelihood, resting state, EEG, fMRI, connectivity, networks.

1 Introduction

The main objective of the *Blind Source Separation* (BSS) problem is to isolate the latent sources by decomposing the observations into several components that attempt to identify independent sources of variations in the data. There are many BSS applications and perhaps the most well known of all is the *cocktail party problem* in which there are many voices talking with microphones scattered around the room. Through weighted combination of microphones recordings, one can separate individual voices. A popular approach to solving this problem is known as *Independent Component Analysis* (ICA).

It is important to know that ICA is an *unsupervised learning* problem and has many interesting useful applications ranging from medical signal analysis to text mining. When coupling with *Singular Value Decomposition* (SVD), the method has been used as *dimension reduction* and *feature transformation* techniques. For example, in EEG data analysis, ICA has been very useful for identifying and removing blink artifacts as these are large in amplitude, have a discrete source and are extremely reliable from blink to blink. Moreover, by viewing ICA as a matrix factorization method, many important features can be extracted through the interpretion of matrix multiplication in terms of the *outer products*. This viewpoint is very useful for temporal and spatial feature extraction in many applications (Hyvärinen et al., 2001; Stone, 2004; Cichocki et al., 2009; Comon and Jutten, 2010; Yu et al., 2014). For more detailed background and history, see `https://en.wikipedia.org/wiki/Independent_component_analysis`.

Note that most of the existing algorithms have been developed by ignoring the auto-correlated or temporal structures of the latent sources. Some have referred to them as *instantaneous ICA* (Comon and Jutten, 2010). From the viewpoint of statistical theory, these methods can be synthesized as an application of *maximum likelihood estimation* (MLE) to the problem of estimating the *probability density function* (pdf) of each latent source. The aforementioned algorithms all employed various parametric functional forms of the source pdf. The desire to estimate these densities with greater flexibility has led to the problem of *nonparametric density estimation* (see Kawaguchi and Truong (2011) and the references therein). Using pdf, however, is not the only way to characterize the sources. In fact, some sources may be better described using their physiological or biological temporal or spatial features.

For example, in EEG data analysis, various rhythms have distinct spectral characteristics that are easier to identify than their counterparts using pdf. These types of sources will be referred to as colored sources and the process of separating them is called the *independent colored source analysis* (ICSA). A R-package `coloredICA` is available for extracting the independent colored sources (ICS). Some statistical inference about the underlying sources is still not widely accessible. The present paper attempts to describe issues related to assessing the latent features.

We proposed a semi-parametric bootstrapping algorithm which invokes ICA estimates in order to create bootstrap samples of either single subject or group EEG scalp data (Nethery et al, 2018). By bootstrapping from the independent auto-regressive (AR) time series residuals of each estimated IC and reconstructing the data using the ICA parameter estimates, we are able to preserve the cross-correlation between EEG channels

and the auto-correlation within EEG scalp channels, critical features for assessing connectivity, in these bootstrap samples. Bootstrapping from the ICs in a manner that preserves temporal structure was proposed by Meinecke et al. (2002) to gauge the separation performance of the ICA algorithm. The bootstrap method was demonstrated to preserve their temporal structure, and the bootstrap samples of EEG data produced by re-mixing the bootstrapped ICs can be used to form confidence intervals and perform hypothesis tests on connectivity-related parameters. The main objective of this paper is to focus on the issue of computing the bootstap estimates, and the problem of identifying ICS by correlation analysis.

The rest of the paper is organized as follow. Section 2 describes the method of ICSA, which includes some more common time series models for the latent sources. The section ends with a study for comparing a selected set of ICA algorithms. Section 3 highlighted the bootstrap method for assessing the sources, followed with a new method based on segmenting or partitioning the data matrix. Section 4 illustrates the proposed procedure using a simulated example and a EEG daa set. This data set has multiple trials that can be used as segments for the proposed procedure. The final section contains some discussions, remarks and thoughts for further research.

2 Colored Independent Component Analysis

Let $\mathbf{X}$ and $\mathbf{S}$ denote the matrices constructed from the observation and source vectors, respectively. Then ICA can be viewed as a matrix factorization method given by

$$\mathbf{X} = \mathbf{AS}.$$

Note that many important features can be extracted by interpreting the matrix multiplication as the sum of columns of $\mathbf{A}$ times rows of $\mathbf{S}$ (also know as outer product). In neuroimaging data analyses such as *functional magnetic resonance imaging* (fMRI) and *electroencephalogram (EEG)* (Truong, 2016), $\mathbf{A}$ is usually viewed as the topographical map of component weights (spatial) and $\mathbf{S}$ as the component time course (temporal), respectively.

Suppose that $\mathbf{X}(t)$ and $\mathbf{S}(t)$, $t = 0, 1, \ldots, T-1$, are the M channel observed time series and the latent sources. Suppose that the sources can be extracted so that the ICA equation can be written as

$$\mathbf{S}_{M\times T} = \mathbf{W}_{M\times M}\mathbf{X}_{M\times T}, \qquad \text{where} \qquad \mathbf{W} = \mathbf{A}^{-1}.$$

The ICA problem will be solved by finding the unmixing matrix $\mathbf{W}$. The method we employ is based on the statistical sampling properties of the discrete Fourier transform (DFT) of $\mathbf{X}$. Before describing this further, we make a digression to discuss some models to be considered for the latent sources.

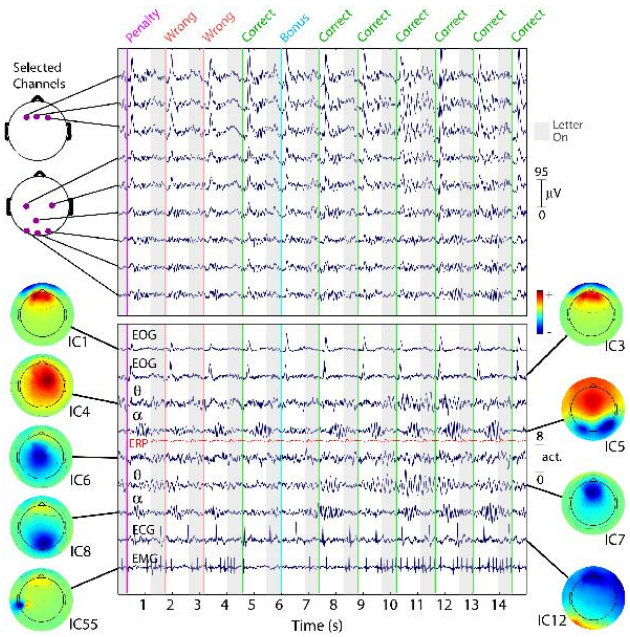


Fig. 1. Fifteen seconds of EEG data at 9 (of 100) scalp channels (top panel is $\mathbf{X}$) with activities of 9 (of 100) independent components (ICs is $\mathbf{S}$ in the center of bottom panel, topographical maps are columns of $\mathbf{A}$ depicted on both sides of bottom panel). Source: http://sccn.ucsd.edu/eeglab.

2.1 Models for the Latent Sources

Our discussion will be based on a general stationary time series Y_t, $t \in \mathbb{Z} \equiv \{0, \pm 1, \pm 2, \ldots\}$. See also books on time series analysis (Brillinger, 2001; Brockwell and Davis, 1991). In the case of the source signal, simply replace Y by S.

White Noise The time series Y is said to be a *white noise* series if Y_t, $t \in \mathbb{Z}$, are stochastically independent. If each Y_t has a Guassian distribution, we say Y is a Gaussian white noise.

In defining more general time series models, it is convenient to denote the white noise by ϵ_t, $t \in \mathbb{Z}$. This is also referred to as the *innovation*. Moreover, we will assume the innovation ϵ_t is stationary with mean zero and variance σ^2.

Moving-Average Processes The time series Y is said to be a *moving-average process* of order q, abbreviated by MA(q), if

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad t = 0, \pm 1, \ldots, \tag{1}$$

where q is a non-negative integer, $\theta_1, \theta_2, \ldots, \theta_q$ are parameters and ϵ_t is the innovation.

Autoregressive Processes The time series Y is said to be an *autoregressive process* of order p, abbreviated by AR(p), if Y can be written as

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \epsilon_t, \quad t = 0, \pm 1, \ldots, \tag{2}$$

where p is a non-negative integer, $\phi_1, \phi_2, \ldots, \phi_p$ are parameters and ϵ_t is the innovation.

Before describing the next process, it will be convenient to introduce a more general way to express the above processes. Let $\theta(\cdot)$ and $\phi(\cdot)$ denote, respectively qth and pth polynomials

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q, \tag{3}$$

and

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p, \qquad z \in \mathbb{C}. \tag{4}$$

Let B be the backward shift operator defined by

$$B^j Y_t = Y_{t-j}, \qquad j \in \mathbb{Z}. \tag{5}$$

Thus the MA(q) process can be represented by

$$Y_t = \theta(B)\epsilon_t, \qquad t \in \mathbb{Z}. \tag{6}$$

Similarly, the AR(p) process is given by

$$\phi(B)Y_t = \epsilon_t, \qquad t \in \mathbb{Z}. \tag{7}$$

Autoregressive and Moving-Average Processes The time series Y is said to be an *autoregressive and moving-average process* of orders p and q, abbreviated as ARMA(p), if

$$\begin{aligned} \phi(B)Y_t &= \theta(B)\epsilon_t, \\ Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} &= \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad t \in \mathbb{Z}, \end{aligned} \tag{8}$$

where p, q, ϕ's, θ's and ϵ_t are given as before. Also, Y_t is an ARMA(p, q) with mean $\mu \in \mathbb{R}$ if $Y_t - \mu$ is ARMA(p, q).

An ARMA(p, q) process is *causal* iff the polynomial $\phi(z)$ does not admit any roots inside the unit disc of the complex plane (Brockwell and Davis, 1991):

$$\phi(z) \neq 0, \qquad |z| \leq 1. \tag{9}$$

Harmonic Processes The time series Y_t, $t \in \mathbb{Z}$ is a harmonic process if

$$Y_t = A\cos(\lambda t + \phi) + \epsilon_t, \qquad t \in \mathbb{Z}, \tag{10}$$

where A is the amplitude, λ is the frequency, and ϕ is the phase.

Spectral Properties of the Sources In EEG signals as well as the neuronal activity inside the brain, an effective way to describe the latent sources is based on their frequency or spectral properties. The dynamics of the above models can also be examined this way. Specifically, the parameters included in the above models become an integral part of the so called power spectral density, which is the Fourier transform of the auto-covariance function of the time series. For example, the power spectrum of the ARMA(p, q) is

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2, \qquad \lambda \in \mathbb{R}. \tag{11}$$

Another useful feature of the frequency interpretation of the signal is it becomes very easy to identify the harmonic processes. For instance, the power spectrum of the harmonic process is

$$f(\omega) = \frac{A^2}{4}[\eta(\omega - \lambda) + \eta(\omega + \lambda)], \qquad \omega \in \mathbb{R}, \tag{12}$$

where $\eta(\cdot)$ is the delta comb function. This type of power spectra is called the *line spectra* and is very useful for modeling brain rhythms or waveforms in EEG data analysis.

2.2 Colored Source Models

It is now ready to model the latent colored sources. For example, if the jth source follows some stationary ARMA(p_j, q_j) model so that $\Phi_j(B)S_j(t) = \Theta_j(B)\epsilon_j(t)$, $\epsilon_j(t) \sim WN(0, \sigma_j^2)$, where B is the backshift operator described above, $\Phi_j(z) = 1 - \phi_{j,1}z - \cdots - \phi_{j,p_j}z^{p_j}$, and $\Theta_j(z) = 1 + \theta_{j,1}z + \cdots + \theta_{j,q_j}z^{q_j}$. Then the power spectrum of this source is given by

$$f_{jj}(\lambda) = \frac{\sigma_j^2}{2\pi}\frac{|\Theta_j(e^{-i\lambda})|^2}{|\Phi_j(e^{-i\lambda})|^2}, \quad \lambda \in \mathbb{R}. \tag{13}$$

A very useful model for the sources is the AR(p) processes described above. In this case, the polynomial $\Theta_j \equiv 1$. The likelihood function in our method has fewer parameters to estimate, and that is the leading case in our software implementation. Another important feature is the automatic selection of the order p for each hidden source. Thus, this approach is flexible in modeling the sources, auto-correlation structures. Moreover, in some applications such as EEG analysis, it will be extremely useful to model the periodic activity as part of the latent sources by adding the line spectra to the AR ones (Nethery et al, 2018).

A more flexible approach to model the spectral density is to express its logorithm as a spline function or smooth piecewise polynomal function. The function `lspec` in the `R` package `polspline` (R Core Team, 2016; Kooperberg el al , 2018) allows one to model the source by a mixture of continuous and line spectra together. See Lee et al. (2018).

2.3 Whittle Likelihood Function

A key element in many statistical data analyses is the estimation of some parameters in the model. An efficient method is based on the likelihood function which, from the conventional approach, is obtained through some specification of probability distributions of the random processes. We consider that is inadequate as the results can be biased if the distributional models are misspecified.

To see how this works, we begin by computing the DFT of $\mathbf{X}(t)$ and its second-order periodogram according to

$$\tilde{\mathbf{X}}_k = \sum_{t=0}^{T-1} \mathbf{X}(t) \exp(-i2\pi kt/T), \qquad \tilde{\mathbf{p}}_k = \frac{1}{2\pi T} |\tilde{\mathbf{X}}_k|^2, \qquad k = 0, 1, \ldots, T-1.$$

Set $\mathbf{f_{SS}} = \text{diag}\{f_{11}, \ldots, f_{MM}\}$, where f_{jj} is the spectral density of the jth source. Using the fact that each component of the periodogram $\tilde{\mathbf{p}}$ is a chi-square random variable distributed according to $f_{jj}(2\pi k/T)\chi^2_2/2$, independently of the other variates for $k = 0, 1, \ldots, T-1$ and $j = 1, \ldots, M$, the latent spectral densities and the unmixing matrix will be estimated by maximizing the *Whittle likelihood* (Whittle, 1952):

$$L(\mathbf{W}, \mathbf{f_{SS}}; \mathbf{X}) = -\frac{1}{2} \sum_{j=1}^{M} \sum_{k=0}^{T-1} \left\{ \frac{\mathbf{e}_j^\top \mathbf{W}_j^\top \tilde{\mathbf{p}}_k \mathbf{W}_j \mathbf{e}_j}{f_{jj}(2\pi k/T)} + \ln f_{jj}(2\pi k/T) \right\} + T \ln |\det(\mathbf{W})|, \tag{14}$$

where $\mathbf{W}_j$ is the j-th column vector of $\mathbf{W}$ and $\mathbf{e}_j$ is the j-th M-dimensional unit vector.

We will illustrate how to extract the latent source spectral information (such as the EEG waveforms) using the ARMA type models described previously.

2.4 Maximum Likelihood Estimation

The unmixing matrix and nuisance parameters related to spectral densities are estimated iteratively by maximizing (14) (Lee et al., 2011). The orthogonality of the unmixing matrix $\mathbf{W}$ can be imposed in two different ways: 1) performing the minimization of the objective function according a Newton–Raphson method with Lagrange multiplier (Lee et al., 2011); and 2) performing eigenvalue decomposition on $\sum_{k=0}^{T-1} \left\{ \frac{\tilde{\mathbf{p}}_k}{f_{jj}(2\pi k/T)} \right\}$ and the eigenvector corresponds to the smallest eigenvalue is the estimates of $\mathbf{W}_j$. Both methods have been implemented in the `coloredICA` `R`-package (R Core Team, 2016; Lee and Zanini, 2015). To determine convergence, Amari's distance (Amari et al., 1996) is used as the convergence criterion due to this scale and permutation invariance.

2.5 Remarks on ICA Algorithms

EEGLAB is widely used by neural scientists and it allows users to try different ICA decomposition algorithms. Only `Infomax` and `jader` are a part of the default EEGLAB distribution. Other algorithms (such as `fastICA`) are also available by installation. According to EEGLAB, the physiological significance of any differences in the results or different algorithms (or of different parameter choices in the various algorithms) have not been tested. However, all three algorithms return near-equivalent components in low-dimensional simulated experiments.

To see how they perform in the presence of colored sources, five independent stationary ARMA time series were simulated under four different sample sizes ($T = 128,\ 256,\ 512,\ 1024$). The sources $\mathbf{S}$ are:

- $S_1(t) = S_1(t-1) - 0.21\ S_1(t-2) + \epsilon_1(t), \quad \epsilon_1(t) \sim$ i.i.d. uniform$(-\sqrt{3}, \sqrt{3})$;
- $S_2(t) = 0.3\ S_2(t-1) + \epsilon_2(t), \quad \epsilon_2(t) \sim$ i.i.d. $N(0,1)$;
- $S_3(t) = 0.8\ S_3(t-1) + \epsilon_3(t), \quad \epsilon_3(t) \sim$ i.i.d. t(3);
- $S_4(t) = \epsilon_4(t) + 0.5\ \epsilon_4(t-1), \quad \epsilon_4(t) \sim$ i.i.d. Weibull$(0.5, 0.5)$;
- $S_5(t) \sim$ i.i.d. double exponential(1).

We then generated a 5×5 orthogonal mixing matrix $\mathbf{A}$ randomly. The data matrix $\mathbf{X}$ was obtained as $\mathbf{X} = \mathbf{AS}$. The simulation was replicated 100 times. The performance of cICA was compared with several popular existing ICA methods, including Infomax (Bell and Sejnowski, 1995), fastICA (Hyvärinen et al., 2001), Kernel ICA (KICA) (Bach and Jordan , 2001), Prewhitening for Charateristic Function based ICA (PCFICA) (Chen and Bickel , 2005), and AMICA (Palmer, et al. , 2010). (Infomax and jader use 4th-order moments (or higher in Infomax depends on the number of channels), we choose Infomax in our simulation study.) The results are given in 2.

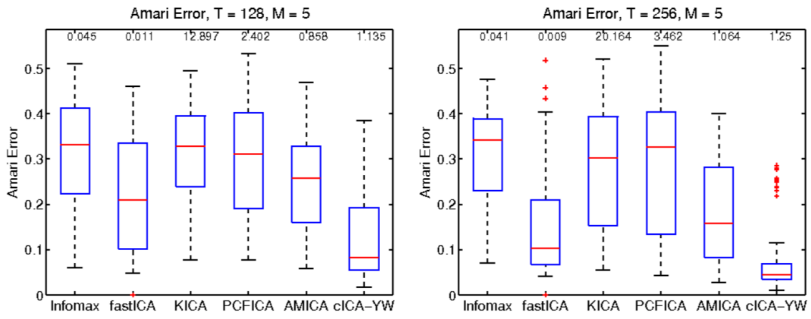
We remark that there is also another algorithm for detecting colored sources and is based on the model:

$$\mathbf{X}_t = \mathbf{A}_1\mathbf{S}_{t-1} + \cdots + \mathbf{A}_p\mathbf{S}_{t-p}. \tag{15}$$

One issue with this model is the number of parameters can be much higher than our cICA models, and that there is some sparsity situation in the sequence of mixing matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$.

The above study employed Amari distance for comparisons. There are other metrics to be considered as well. For example, one may prefer the false-positive and false-negative rates for evaluating the performance. This is particularly useful for detecting brain activation. A study was reported in (Lee et al., 2011).

With increasing number of ICA algorithms, it is very challenging to compare their performance, both in simulated and neural physiological studies. Here we present a small comparative study using algorithms that are highly relevant to human brain imaging studies. Our cICA performs well in various settings of the simulated experiments (Lee et al., 2011). The real challenge is how significance this approach is in real brain studies. Can this work with sources that are not stationary? In fact, the method to described later is motivated by this issue. For example, non-stationarity is often detected in EEG studies, but this occurs only



(a) $T = 128, M = 5$ (b) $T = 256, M = 5$

(c) $T = 512, M = 5$ (d) $T = 1024, M = 5$

Fig. 2. Simulation Study I-1: Performance Comparison for ARMA Sources. One hundred simulation runs are performed under different sample sizes. In each simulation run, five ARMA sources are generated and are mixed through a 5×5 mixing matrix. The boxplots show the Amari error between the true unmixing matrix and the estimated unmixing matrix obtained by the various ICA methods. The median computation time of each method is at the top of the corresponding boxplot. The cICA provides more accurate estimates than its competitors in a fairly short time.

when tasks are carried out. So the data can be analyzed by breaking into seqments for which cICA can still be applicable.

3 Methods for Assessing the IC Estimates

While the conventional ICA algorithms assume that the ICs contain no auto-correlation (Bell and Sejnowski, 1995; Cichocki et al., 2009; Hyvärinen et al., 2001), Lee et al. (2011) developed a semi-parametric colorICA (CICA) algorithm that models the ICs as AR time series processes which was described in Section 2. The `R` `coloredICA` package will produce estimates of the mixing matrix and ICs, $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$, a collection of estimated AR coefficients and the time series variances for each IC. It will also provide a matrix of smoothed power spectra estimates for each IC, $\hat{\mathbf{G}}$. Allowing for auto-correlation within ICs is critical in the analysis of resting state EEG data, because the electrical signals emitted by resting state brain activity are known to be cyclic processes. Hence, we focus the development of our bootstrapping algorithm around CICA.

The data, such as the recorded EEG scalp channels, contain both cross-correlation and auto-correlation, and each of these features are critical to properly characterizing connectivity. Thus, any useful bootstrapping algorithm must preserve both the channel cross-correlation and auto-correlation in the bootstrapped datasets. In order to do so, our algorithm must take into consideration both the mixing of signals, which induces the cross-correlation between the channels, and the temporal correlation in these signals, which induces the auto-correlation within the channels.

The following semi-parametric ICA-based procedure for creating a bootstrapped dataset is similar to the semi-parametric procedure recommended for bootstrapping in a linear model framework. First, CICA should be applied to the matrix of resting state EEG data, say $\mathbf{X}$. Because the ICs are independent, we can construct a bootstrap sample of each of one and mix them to obtain a bootstrap sample of the original data that preserves the channel cross-correlations. To create a bootstrap sample of each IC that retains its time series structure, its estimated AR model residuals should be resampled with replacement and plugged into the estimated AR model (Efron and Tibshirani, 1986), initializing the bootstrapped time series using the block initialization method (Stine, 1987). Finally, to construct the bootstrap sample of the data, $\mathbf{X}^*$, the bootstrapped ICs should be concatenated into a matrix $\mathbf{S}^*$ and multiplied by $\hat{\mathbf{A}}$, i.e. $\mathbf{X}^* = \hat{\mathbf{A}}\mathbf{S}^*$. A large number, B, of bootstrapped datasets can be constructed by repeating this process B times. A summary of this method motivated by an application to EEG study is described in (Nethery et al, 2018).

One of the issues to this method is the computing resources it required. In this paper, we assess the variance of the IC estimates by segmenting the data into blocks. Then apply `coloredICA` to each block and treat this block as the bootstrapped sample. With the stationary nature of the time series data, this works very well for large data sets (such as EEG) and it is much faster than the above bootstrap approach.

Start by applying `coloredICA` to the original data $\mathbf{X}$, which yields the estimates of the mixing matrix $\hat{\mathbf{A}}$ and the latent source $\hat{\mathbf{S}}$. Suppose number of columns of $\mathbf{X}$ is $B \times T$. We then partition $\hat{\mathbf{S}}$ into B blocks: $\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_B$ with T columns in each block. Now construct $\mathbf{X}_b^* = \hat{\mathbf{A}}\hat{\mathbf{S}}_b$, $b = 1, \ldots, B$ and invoke `coloredICA`

to these blocks to produce $\tilde{\mathbf{A}}_b$ and $\tilde{\mathbf{S}}_b$, $b = 1, \ldots, B$. Then aggregate these results along with the Amari distance for each estimate of the mixing matrix: amari$(\tilde{\mathbf{A}}_b, \hat{\mathbf{A}})$, $b = 1, \ldots, B$. We then arrange these distances in ascending order so as to form order statistics. This is similar to the percentile method in bootstrap (Efron and Tibshirani, 1986) for setting the confidence limits. We will see how it works in the next section.

4 Numerical Examples

This section presents studies based on simulation and real EEG data analysis. The main objective is to illustrate how the data can be used to evaluate statistical properties of varous estimates derived from cICA.

4.1 A simulated study

In this section, we will use the R-package `coloredICA` to illustrate the idea of assessing the IC estimates. For simplicity we will consider a setup involving only two ICs and these are simulated from MA(2) and AR(1), respectively. Namely,

```
library(coloredICA)
library(fastICA)

T <- 1024
B <- 10
S1  <-  arima.sim(list(order=c(0,0,2),ma=c(1,0.25)),T*B)
S2  <-  arima.sim(list(order=c(1,0,0), ar=-0.5),T*B,
            rand.gen = function(n, ...) (runif(n)-0.5)*sqrt(3))
S <- rbind(S1, S2)
A  <- matrix( c(0.9239171, 0.9216859, -0.3825927, 0.3879370), 2, 2)
X  <-  A %*% S
cica  <-  cICA(X, tol=0.001)
fica  <-  fastICA(t(X), 2)

amari_distance(t(A), t(cica$A))
## [1] 0.004102551
amari_distance(t(A), fica$A)
## [1] 0.007137048
```

We can see fastICA has an Amari distance almost double of the cICA's.

Estimates of the mixing matrix The estimated mixing matrix by cICA is

```
cica$A
```

```
##           [,1]       [,2]
## [1,] -1.309830 -0.2225984
## [2,] -1.310516  0.2224818
```

The estimated mixing matrix by fastICA:

```
fica$A
```

```
##           [,1]       [,2]
## [1,] 0.2258657 -0.2192114
## [2,] 1.3092701  1.3110669
```

Note that how similar they are after a transpose and permutation of columns in fastICA. (Recall that Amari distance is permutation invariant.)

The latent sources can be recovered via:

```
Shat1 <- cica$S     # cICA
Shat3 <- t(fica$S)  # fastICA
```

We remark that cICA nearly reproduced the latent sources indicated by

```
cor(S[1,], Shat1[1,])
## [1] -0.9999993
```

```
cor(S[2,], Shat1[2,])
## [1] 0.9999625
```

The negative correlation is a trademark of ICA, where things can be flipped. In our case, this can be corrected later when we consider resampling methods.

Estimates of ARMA parameters from cICA Estimates of the ARMA parameters based on the true source and `Shat1` via cICA are given below. Note that the reconstructed ICs produced almost identical ARMA estimates as the latent sources.

```
##
## Call:
## arima(x = S[1, ], order = c(0, 0, 2))
##
```

```
## Coefficients:
##          ma1     ma2  intercept
##       0.9840  0.2371    -0.0034
## s.e.  0.0096  0.0096     0.0219
##
## sigma^2 estimated as 0.9947:  log likelihood = -14503.29,  aic = 29014.58

##
## Call:
## arima(x = Shat1[1, ], order = c(0, 0, 2))
##
## Coefficients:
##          ma1     ma2  intercept
##       0.9840  0.2370     0.0024
## s.e.  0.0096  0.0096     0.0154
##
## sigma^2 estimated as 0.4935:  log likelihood = -10914.49,  aic = 21836.98

##
## Call:
## arima(x = S[2, ], order = c(1, 0, 0))
##
## Coefficients:
##           ar1  intercept
##       -0.4933     0.0021
## s.e.   0.0086     0.0033
##
## sigma^2 estimated as 0.2525:  log likelihood = -7482.36,  aic = 14970.71

##
## Call:
## arima(x = Shat1[2, ], order = c(1, 0, 0))
##
## Coefficients:
##           ar1  intercept
##       -0.4933     0.0037
```

```
## s.e.   0.0086     0.0058
##
## sigma^2 estimated as 0.7566:  log likelihood = -13102.23,  aic = 26210.47
```

Assessing the variability using block/segment approach We now partition the reconstructed source `Shat1` into $B = 10$ blocks. We remark that B depends on the dimension of the data. For EEG data, this can be easily hundreds of blocks. Now carry out the algorithm for cICA on each of these blocks as described in the previous section.

The resulting Amari distances are listed below with the first being the whole sample.

```
# cICA
##  [1] 0.004102551 0.003274755 0.029627172 0.007178622 0.005465980
##  [6] 0.018813174 0.003426712 0.004873029 0.005445852 0.010608094
## [11] 0.017459276
mean(a1[2:11])
## [1] 0.01061727

# fastICA
##  [1] 0.007137048 0.124032848 0.109251357 0.039868824 0.118512387
##  [6] 0.035403977 0.056871647 0.075247401 0.074473983 0.054886045
## [11] 0.030177283
mean(a2[2:11])
## [1] 0.07187258
```

Note that fastICA has a larger values in general which means that it has a poorer performance compared to cICA.

The average of the mixing matrix from cICA is given by

```
##           [,1]       [,2]
## [1,] -1.308776 -0.2216407
## [2,] -1.308943  0.2231364
```

This is in close agreement with the whole sample given above.

4.2 EEG Data Analysis

In EEG data analysis, ICA has been very useful for identifying and removing blink artifacts as these are large in amplitude, have a discrete source and are extremely reliable from blink to blink.

The EEG data (M. Cohen) has 64 channels, 640 time points and 99 trials. The trials are for obtaining Event Related Potential (ERP). The left panel of Figure 3 depicts the 64 channel locations. The topographical location at 500 ms is shown on the right panel.

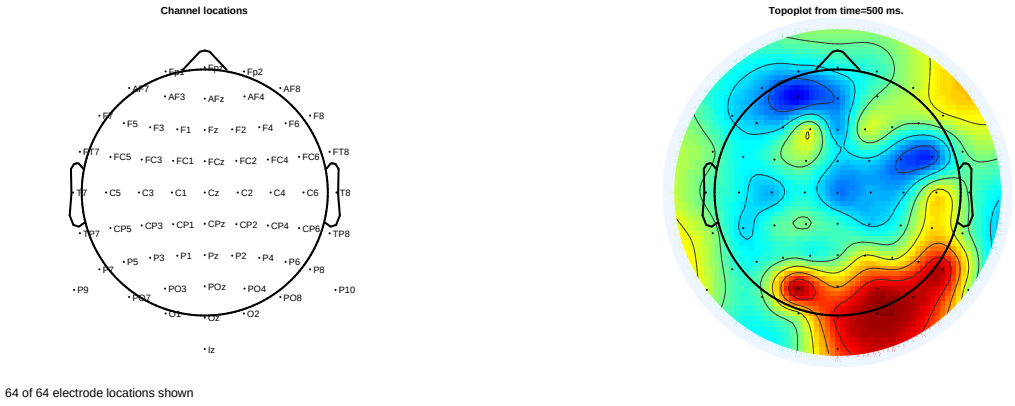


Fig. 3. Left panel: 64 channel locations. Right panel: Topographical localization at 500 ms. Source: http://sccn.ucsd.edu/eeglab.

While this is useful to see the task related scalp potential, it conveys very little how the brain actually works because there may be millions of neurons from different areas of the brain forming a network to yield the electrode response. To see the dynamics of the brain, we apply ICA or cICA to identify various independent physiological processes. Some of these processes (delta, theta, alpha, beta rhythms) have well-defined spectral properties (oscillations with banded frequencies). To make the visualization easy, we apply singular value decomposition (SVD) then select a few ICs while the reduced still accounts for a high percentage of the original variations. In this case, 4 components will account for 98% of the variations. Thus the dimension reduction step is given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' \approx \mathbf{U}_4\mathbf{D}_4\mathbf{V}_4'.$$

Now apply cICA to yield $\mathbf{V}_4' = \mathbf{AS}$. This procedure is repeated for each trial so that there will be 99 sets of $\mathbf{A}$ and $\mathbf{S}$. This is the segmenting step in the previous section. Now each trial data will be used to evaluate the variance of the cICA estimates of the brain localization map $\mathbf{A}$.

The table below shows the mean of $\mathbf{A}$ over the first ten trials.

```
> Ahatmean
              [,1]         [,2]          [,3]         [,4]
[1,] -0.0079284215  0.005583165 -0.0076123458 -0.003882935
[2,]  0.0024007673  0.007016976  0.0006320704 -0.003452728
[3,]  0.0025362729  0.008831092 -0.0059501678  0.003315204
[4,]  0.0007322852 -0.001218710  0.0020414141  0.001956753
```

The table below shows the SD of $\mathbf{A}$ over the first ten trials.

```
> Ahatsd
            [,1]        [,2]        [,3]       [,4]
[1,] 0.020402953 0.006344184 0.009374807 0.03171105
[2,] 0.031275928 0.018740173 0.006289630 0.01709043
[3,] 0.015374616 0.031796542 0.009962712 0.01563880
[4,] 0.004612701 0.012490078 0.037424755 0.01214102
```

Figures 4 and 5 depict $\mathbf{A}$ averaged over 99 trials and the 9th trial, respectively. From this analysis, one can see that the data is very noisy and this has reflected in the cICA results. For example, IC1 has the largest variation among the trials. IC4 of a single trial has a topomap intensity ten times the average. We can also see that the brain dynamics is a very complicated process, there is a high variation of trial-to-trial dynamics even it is under the same experimental setting. This explains why ERP should be estimated using multiple trials design.

5 Discussion

In this paper, we present a method for assessing the performance of two ICA algorithms using a sample of data. The idea is based on partitioning the data into blocks so as to maintain the correlated structures of the temporal data, and study if the ICA algorithm would be able to reproduce these structures. The choice of `fastICA` is just for convenience, and it seems to have very good numerical properties (Lee et al., 2011). See also EEGLAB (`http://sccn.ucsd.edu/eeglab`) for the comparison to `Infomax`.

Through the simple experiment above, we see that methods such as `coloredICA` are capable of detecting auto-correlated structures. This should be expected as its algorithm was derived for that purpose and it should work well with correlated data. One interesting result is that it also recovers the mixing matrix (the

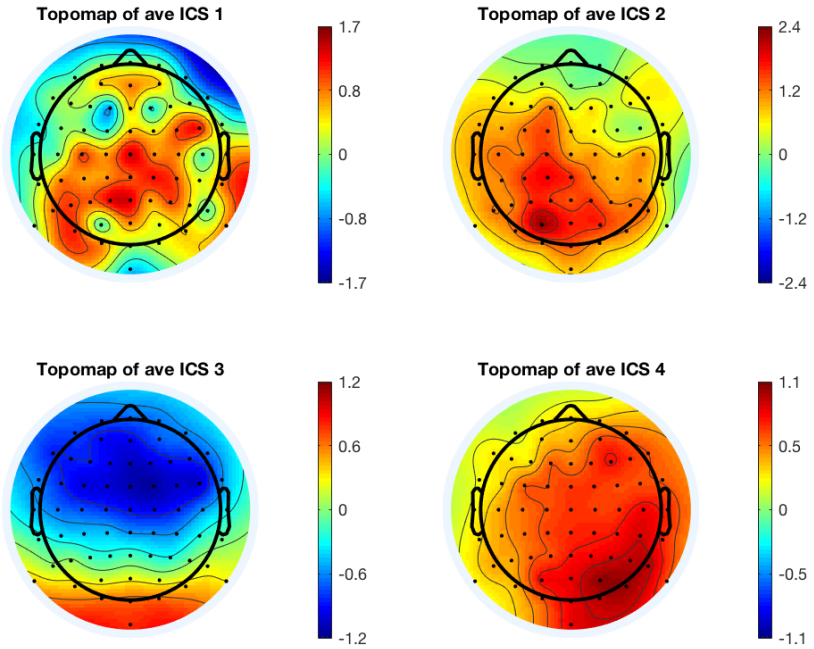


Fig. 4. Brain localization map as an average over 99 trials.

spatial feature) very well as demonstrated by the analysis based on block-samples. The implication here is that the cross-correlations should be well-preserved via a good estimate of the mixing matrix.

The sign or flip problem encountered in practical applications of ICA has been addressed in our experiment by simply calculating the correlations among the block-samples. We applied this idea when the average of the block mixing matrices was computed.

A more extensive numerical experiment is being conducted to investigate and compare this with other bootstrap methods. Furthermore, the other side of the temporal ICA is the spatial ICA, which is part of `coloredICA`. It will be important to examine if the same idea can be extended to such setting.

In the EEG data analysis, it is observed that the data is much noisier than our simulated examples. This is shown by the contrasts between trial average and individual trial results. It is conceivable that the dimension reduction via SVD may have contributed some, but is it really true that a small perturbation of the singular value may have caused such a huge impact on the variance of the ICA estimates?

Aside from the brain localization map, it will be important to study the temporal features closely as the EEG data should have some of the well known physiological oscillations or the banded frequency components. This should be naturally addressed through the application of `coloredICA` as the estimating procedure is based on the spectral information of the latent sources. Some of these have been addressed by Nethery et al (2018) and the multiple trials design of the data should make a good comparison with the bootstrap, both computationally and statistically.

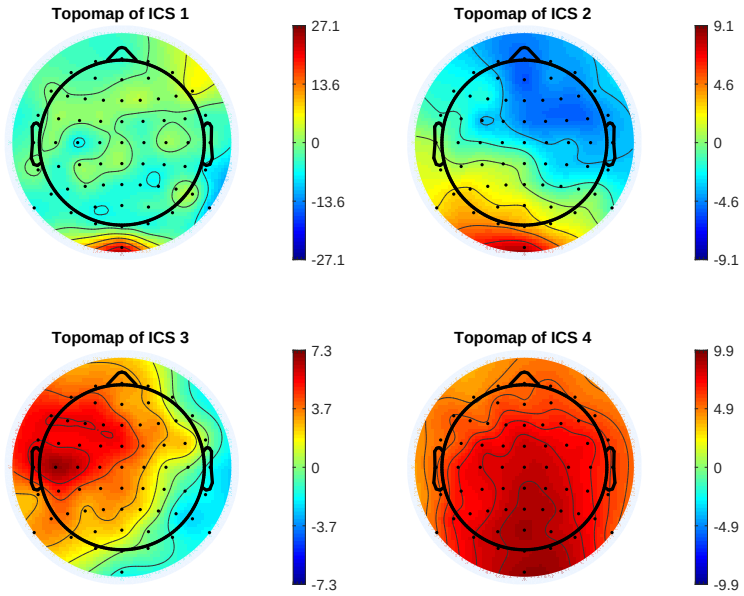


Fig. 5. Brain localization from trial 9.

Bibliography

S. Amari, A. Cichocki, H.H. Yang, et al. A New Learning Algorithm for Blind Signal Separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1996.

Bach, F.R. and Jordan, M.I. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3: 1–48, 2003.

Bell, A. J. and T. J. Sejnowski (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation 7*(6), 1129–1159.

D.R. Brillinger. *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, 2001.

P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, second edition, 1991.

Chen, A. and Bickel, P. J. Consistent Independent Component Analysis and Prewhitening. *IEEE Transactions on Signal Processing*, 53: 3625–3632, 2005

A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Hyvarinen, A. and E. Oja (2000). Independent component analysis: Algorithms and applications. *Neural Networks 13*(4-5), 411–430.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

Kachenoura, A., L. Albera, L. Senhadji, and P. Comon (2008). ICA: A potential tool for BCI systems. *IEEE Signal Processing Magazine 25*(1), 57–68.

A. Kawaguchi and Y. K. Truong. Logspline independent component analysis. *Bulletin of Informatics and Cybernetics*, 43:83–94, 2011.

C. Kooperberg, C. Stone and Y. Truong. *polspline: Polynomial Spline Routines*, 2018. R package version 1.1.13.

Laufs, H. (2010). Multimodal analysis of resting state cortical activity: What does EEG add to our knowledge of resting state BOLD networks? *Neuroimage 52*(4), 1171.

S. Lee, H. Shen, Y.K. Truong, M.M. Lewis, and X. Huang. Independent Component Analysis Involving Autocorrelated Sources with an Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association*, 106(495):1009–1024, 2011.

S. Lee, P. Zanini, H. Shen and Y. Truong. *coloredICA: Implementation of Colored Independent Component Analysis and Spatial Colored Independent Component Analysis*, 2015. R package version 1.0.0.

Meinecke, F., A. Ziehe, M. Kawanabe, and K. Müller (2002). A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE Transactions on Biomedical Engineering 49*(12), 1514–1525.

Nethery, R., A. Campbell and Y. Truong (2018). Assessing EEG Resting State Connectivity using Independent Component Analysis. *7th Annual International Conference on Cognitive and Behavioral Psychology* (CBP 2018), 114–124.

Palmer, J.A. and Kreutz-Delgado, K. and Makeig, S. An Adaptive Mixture of Independent Component Analyzers with Shared Components. http://sccn.ucsd.edu/$\sim$jason/amica_r.pdf, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Shumway, R. and D. Stoffer (2011). *Time Series Analysis and Its Applications, 3rd Edition*. Springer.

Stine, R. (1987). Estimating properties of autoregressive forecasts. *Journal of the American Statistical Association 82*(400), 1072–1078.

J. V. Stone. *Independent Component Analysis: A Tutorial Introduction*. MIT Press, Cambridge, MA, 2004.

Y. Truong and M. Lewis. *Statistical Techniques for Neuroscientists*. Taylor & Francis, NY, USA, 2016.

P. Whittle. Some Results in Time Series Analysis. *Skand. Aktuarietidskr*, 35:48–60, 1952.

X. Yu, D. Hu, and J. Xu. *Blind Source Separation: Theory and Applications*. John Wiley & Sons, 2014.

Relevance analysis in spatio-spectral components based on Permutation Entropy supporting MI discrimination

C. López-Montes, D. Cárdenas-Peña, and G. Castellanos-Dominguez

Signal proccesing and recognition group, Universidad Nacional de Colombia, Manizales, Colombia
{julopezm,dcardenasp,cgcastellanosd}@unal.edu.co

Abstract. Nowadays, the motor imagery (MI) has become an important paradigm for developing and testing Brain-Computer Interfaces (BCI). Conventionally, BCI systems rely on extracting powerful discriminating patterns from the electroencephalogram (EEG) using the Common Spatial Pattern (CSP) algorithm. An important improvement on the BCI emerged from the inclusion of filter-banked CSP variants. Nonetheless, its successful application depends on an appropriate selection of frequency bands and EEG channels for each subject. To solve this issue, this work introduces a new supervised spatio-spectral relevance analysis (termed PFB) from EEG data. The proposed PFB ranks each channel and sub-band according its MI discrimination measured by means of an accomplished statistical test on the trial-wise permutation entropy (PeEn). The proposed approach holds two parameters that rule the number of selected spatio-spectral components and CSP features. The experimental results evidence an improved accuracy in comparison with CSP, FB and SFB assessed in the BCI competition III IVa, competition IV IIa, and GIGASCIENCE datasets. Therefore, focusing on the discriminative channels and sub-bands, we enhance the MI classification with the benefit of a better neurophysiological interpretation of the extracted components.

Keywords: spatio-spectral relevance · Permutation entropy · Brain-computer interface.

1 Introduction

Nowadays, there are many challenges in rehabilitation in the development of systems that enhance communication between the brain and the exterior environment. Conventional Brain-Computer Interfaces (BCI) have been widely used to assist disabled people to re-establish their capabilities of environmental control decoding the brain activity [10]. Electroencephalogram (EEG) is an easy to implement and noninvasive method to measure the brain activity [6]. Since the EEG amplitude is contaminated with artifacts as muscle and eye movements, common spatial patterns (CSP) provide a feature extraction procedure that reduces the influence of the most common noise present along the recordings [10]. However, CSP features are highly sensitive and yield wrongly the frequency band selection. Therefore, either unfiltered or inappropriately filtered EEG is performed [7].

Several approaches have been employed to address the problem of correctly select a specific frequency band for each subject performing either optimization of the spectral filter applied before CSP or selecting significant features from multiple frequency bands [4]. Sub-band Common Spatial Pattern (SBCSP) try to find the frequency band with the highest performance focusing on the relationship between CSP features and the frequency [5], Filter Bank Common Spatial Pattern (FBCSP) estimate the mutual information of the CSP features in each frequency bands highlighting the most discriminatory bands [2], Discriminative Filter Bank Common Spatial Pattern (DFBCSP) selects a specific frequency band for each subject comparing the Fisher ratio of filtered EEG signals [9], Sliding Window Discriminative CSP (SWD-CSP) uses affinity propagation to detect discriminative CSP features [7], and Sparse Filter Band Common Spatial Pattern (SFBCSP) perform a sparse regression to select the most discriminant bands [11].

For overcoming the above issues, we propose a spatio-spectral relevance analysis simultaneously selecting channels and bands in a supervised scheme. The approach, termed P_λFB, characterizes each channel within the set of bandpass-filtered EEG recordings trough the Permutation entropy. Then, we test whether trials from two classes have the same mean entropy, so that resulting p-values rank the spatio-spectral. According to a subject-dependent significance level, spatio-spectral components (channels in a frequency band) are selected to feed the CSP feature extraction and LDA classifier. Base opun the obtained results, it is proven that the introduced relevance analysis improves the performance and enhances the representation of the motor imagery paradigm, while reducing the space of the spatio-spectral components.

2 Materials and Methods

2.1 Sub-band Common Spatial Patterns

Let $\{\boldsymbol{x}_n^c \subset \boldsymbol{X} \in \mathbb{R}^T : n \in N, c \in C\}$ be a set of N acquired EEG recordings of length T and C the number of channels to be further bandpass filtered, adjusting two main parameters for each subject: elemental bandwidth $B \subset F$ and their band overlapping $\delta_B \subset B$. Therefore, the following set of bandpass-filtered EEG data is obtained: $\{\tilde{\boldsymbol{x}}_{n,b}^{\boldsymbol{c}} \subset \tilde{\boldsymbol{X}} : b \in B\}$. For the above-described MI dataset, as carried out in [8], we use 17 band-pass filters with a bandwidth of $B = 4\,Hz$ as to cover the whole frequency EEG band $F \in \mathbb{R}^+$, ranging from 4 to 40 Hz and fixing the overlap between each other at $\delta_B = 2\,Hz$.

In binary classification tasks, conventional CSP finds a spatial filter matrix $\boldsymbol{W}_b \in \mathbb{R}^{C \times 2K}$ to linearly map the bandpass-filtered EEG data $\tilde{\boldsymbol{X}}_b \in \mathbb{R}^{N \times C}$ onto a space $\tilde{\tilde{\boldsymbol{X}}}_{\boldsymbol{b}} = \boldsymbol{W}_{\boldsymbol{b}}\, \tilde{\boldsymbol{X}}_{\boldsymbol{b}}$, so that the variance of the mapped signal is maximized for one class while the variance of another class is minimized. The spatial filters $\boldsymbol{w}_b^* \in \mathbb{R}^C$ are the solution of maximizing the Rayleigh quotient:

$$\boldsymbol{w}_b^* = \max_{\boldsymbol{w}_b} \frac{\boldsymbol{w}_b^\top \boldsymbol{\Sigma}_b^- \boldsymbol{w}_b}{\boldsymbol{w}_b^\top \boldsymbol{\Sigma}_b^+ \boldsymbol{w}_b}, \text{ s.t.: } \|\boldsymbol{w}_b\|_2 = \boldsymbol{I}_C \tag{1}$$

where $\boldsymbol{I}_C$ is the identity matrix size $C \times C$, and the spatial covariance matrix of the class $l \in \{-,+\}$ is estimated as $\widehat{\boldsymbol{\Sigma}}_b = \mathbb{E}\left\{ \tilde{\boldsymbol{X}}_b^r \, \tilde{\boldsymbol{X}}_b^{r\top} : \forall r \in N_l \right\}$, being N_l the number of trials in

class l. Notations $\|\cdot\|_2$ and $\mathbb{E}\{\cdot\}$ stand for ℓ_2-norm and expectation operator, respectively.

In practice, the optimization framework in Eq. (1) is equivalently transformed into the generalized eigenvalue problem $\widehat{\boldsymbol{\Sigma}}_b^- \boldsymbol{w}_b^* = \lambda \widehat{\boldsymbol{\Sigma}}_b^+ \boldsymbol{w}_b^*$ with $\lambda \in \mathbb{R}^+$. Thus, a set of spatial filters $\boldsymbol{W}_b^* = [\boldsymbol{w}_{b,1}^* \dots \boldsymbol{w}_{b,2K}^*]$ are obtained by collecting eigenvectors that correspond to the K largest and smallest eigenvalues of the generalized eigenvalue problem. Therefore, the CSP feature vector that accounts for the bandpass filtered components (termed sub-band CSP – SBCSP) is formed as $\boldsymbol{\xi}_b = [\xi_{b,k} : k \in 2K]$ with $\boldsymbol{\xi}_b \in \mathbb{R}^{2K}$, where $\boldsymbol{\xi}_b$ across all trials is defined as follows:

$$\boldsymbol{\xi}_b(\boldsymbol{X}) = \ln(\text{var}\{\boldsymbol{W_b}^{*\top} \widetilde{\boldsymbol{X}}_b\}) \tag{2}$$

where $\text{var}\{\cdot\}$ stands for the variance operator.

2.2 Improved Data-driven Sub-band Spatial Relevance

Mainly, the plain SBCSP method in Eq. (2) does not reflect the contribution of $2K$ selected eigenvalues to increase the discriminating power of the spatial filter designing, encouraging the use of data-driven approaches to enhance the relevance estimation of each bandwidth. Besides, the CSP-based feature extraction methods poorly behave in the presence of nonstationarity, without mentioning its degradation due to outlier and artifact trials. To overcome such an issue, we propose to select the most discriminating spatio-spectral components based on the Permutation Entropy $PeEn(\cdot) \in \mathbb{R}^+$, instead of further using variance-based methods as a measure of uncertainty. Therefore, we introduce an entropy-based relevance criterion for enhancing the discriminative power of CSP features, extracted from b-th temporal-frequency component, termed P_λFB:

$$\rho_b^c(\tilde{\boldsymbol{x}}_{n,b}^{\boldsymbol{c}}|l) = \text{d}\{\mathbb{E}\left\{PeEn(\tilde{\boldsymbol{x}}_{r,b}^{\boldsymbol{c}}) : \forall r \in N_-\right\} - \mathbb{E}\left\{PeEn(\tilde{\boldsymbol{x}}_{r,b}^{\boldsymbol{c}}) : \forall r \in N_+\right\}\} \tag{3a}$$

$$\text{s.t.: } \rho_b^c(\tilde{\boldsymbol{x}}_{n,b}^{\boldsymbol{c}}|l) \leq \varepsilon \in \mathbb{R}^+; \forall b \in B, c \in C \tag{3b}$$

where $\text{d}\{\cdot\} \in \mathbb{R}^+$ denotes the distance between class-related measures of entropy. Note that the smaller the value $\rho_b^c(\cdot)$, the more significance the component holds. For purposes of implementation, the supervised distance $\text{d}\{\cdot\}$ is estimated as the significance of statistical t-test value, being ε a minimal threshold of accepted relevance computed for any bandpass filtered component that is fixed empirically. As regards computation of the measure of uncertainty in Eq. (3a), we use the Permutation Entropy estimated as below:

$$PeEn(\tilde{\boldsymbol{x}}_b^{\boldsymbol{c}}) = \sum_{i=1}^{m!} p(\pi_i) \log \frac{1}{p(\pi_i)} \tag{4}$$

where $\boldsymbol{p}(\boldsymbol{\pi_i})$ is the discrete probability distribution of $m!$ ordinal patterns, which is built by mapping the time series $(\tilde{x}(t)_b^c, \tilde{x}(t+1)_b^c, ..., \tilde{x}(T-1)_b^c, \tilde{x}(T)_b^c)$ into a sequence of ordinal patterns. The particular pattern π_t at time t is obtained form the m-tuple of values:

$$(\tilde{x}(t)_b^c, \tilde{x}(t+\tau)_b^c, \tilde{x}(t+2\tau)_b^c ..., \tilde{x}(t+(m-1)\tau)_b^c) \rightarrow \pi_t \in \Omega_m \tag{5}$$

being $m \in \mathbb{N}$ the order of the PeEn and $\tau \in \mathbb{R}^+$ the time delay, and the probability distribution is computed by counting the pattern occurrences. Therefore, the feature set in Equation (2) results from the subset of channels and bands that agree the relevance criterion, $\boldsymbol{X}_\varepsilon = \left\{ \tilde{\boldsymbol{x}}^{\boldsymbol{c}}_{n,b} : \rho^c_b(\tilde{\boldsymbol{x}}^{\boldsymbol{c}}_{n,b}|l) \leq \varepsilon \right\}$, as $\boldsymbol{\xi}_b(\boldsymbol{X}_\varepsilon)$, enhancing the class discrimination in the new feature space.

As a result, by measuring the inter-class distance from the trial-wise Permutation entropy, we handle a better accounting for time-variant dynamics of each channel and sub-band thanks to the construction of distributions based on ordinal pattern encodings.

3 Experimental set-up

3.1 Description of Datasets

The BCI Competition III dataset IVa[1] contains the EEG signals recorded from five subjects performing two different (right hand and left hand) MI tasks. 118 channels at the positions of the extended international $10/20$ system were used for measuring the EEG signal, sampled at a rate of 1000Hz. The dataset contains a total of 280 trials for each subject with equal number of trials for each task. The data provided were interval filtered with a passband of $0.05 - 200$Hz the down-sampled data at 100Hz is used. Visual cues, lasting 3.5s indicated which of the following three motor imageries the subject should perform: (L) left hand, (R) right hand, (F) right foot. The presentation of target cues were intermitted by periods of random length, 1.75 to 2.25s, in which the subject could relax.

The BCI competition IV dataset IIa[2] contain a collection of EEG signals recorded using a 22-electrode montage from nine subjects that compose the dataset. Each subject performs two sessions on different days with four motor imagery tasks, namely left hand, right hand, both feet, and tongue on two sessions. The session includes six runs with twelve trials per task, obtaining 144 trials for each class. The signals were sampled with 250Hz and bandpass-filtered between 0.5Hz and 100Hz. The subjects were sitting in a comfortable armchair in front of a computer screen. At the beginning of a trial ($t = 0$s), a fixation cross appeared on the black screen. In addition, a short acoustic warning tone was presented. After two seconds (t=2s), a cue in the form of an arrow pointing either to the left, right, down or up (corresponding to one of the four classes left hand, right hand, foot or tongue) appeared and stayed on the screen for 1.25s. This prompted the subjects to perform the desired motor imagery task. No feedback was provided. The subjects were asked to carry out the motor imagery task until the fixation cross disappeared from the screen at $t = 6$s.

The GIGASCIENCE[3] dataset contains an EEG signal recorded conducting a BCI experiment for the motor imagery movement of the left and right hands with 52 subjects [3] of which we selected 44 subject for the current experiment because the remainder subject hold a large amount of trials with the artifact. EEG data were collected using 64 Ag/AgCl active electrodes $64-$channel montage based on the international $10-10$

[1] http:// www.bbci.de/competition/iii/
[2] http://www.bbci.de/competition/iv/
[3] http://gigadb.org/dataset/100295

system of sampling rate 512Hz. The dataset contains 100 or 120 trials for each MI class. Each trial starts with the monitor showed a black screen with a fixation cross for $t = 2$s then one of two instructions ("left hand" or "right hand") appeared randomly on the screen for $t = 3$s, and subjects were asked to move the appropriate hand depending on the instruction given. After the movement, when the blank screen reappeared, the subject was given a break for a random 4.1 to 4.8 seconds. These processes were repeated 20 times for one class (one run), and one run was performed

Being the most used MI paradigm in the literature, this work considers the binary classification of left and right movement for the three datasets. The raw signal from EEG is segmented within the time MI task is performed. Common Spatial Patterns (CSP) and Linear Discriminant Analysis (LDA) achieve the feature extraction and classification stages respectively, for the three datasets.

3.2 Parameter tuning

The proposed method P_λFB requires two procedures of parameter tuning; one is the significance threshold for the t-test, $\varepsilon \in (0,1]$ and another parameter is the Lasso regularization factor, $\lambda \in \mathbb{R}^+$. The first controls the amount of spatio-spectral components selected and the second controls the number of features extracted in the frequency bands that are used for the classification. The parameters ε, λ are set up based on a grid search for the best classification accuracy. Fig. 1 illustrates the best pair of parameters found in each subject in the three evaluated datasets, proving the need for tuning ε and λ for each subject to properly decode the different brain activities. Besides, it is observed a directly proportional relationship between ε and λ. This result can be explained because both parameters regulate the amount of information in component selection and feature reduction stages of the proposed framework. A high value of λ means that Lasso will remain with none or few features while a very low value of ε yields few spatio-spectral components. Therefore, the suitable trade-off between ε and λ retains relevant information in the space of the components and features without compromising the performance of the classifier.

3.3 Performance results

We compare the P_λFB against the conventional CSP, Filter-bank (FB) CSP [2] and Sparse Filter-bank (SFB) CSP [11] methods for the considered datasets. Fig. 2 presents the accuracy rate attained for each subject by the evaluated approaches. Note that subjects are sorted according the CSP accuracy, aiming to highlight the performance improvement by each approach. Subject-wise accuracies evidence that P_λFB improves the performance of the worst subjects while maintaining the accuracy of the best subjects in comparison with CSP. Regarding the other filter-banked approaches, our method outperforms by achieving the highest accuracy rates for most of the subjects in the three datasets. Therefore, the proposed component selection in both, space and in frequency, removes irrelevant and redundant information under the t-test criterion applied over the permutation entropy.

Further, Table 1 displays the overall fivefold cross validation classification accuracy, component reduction rate (*crr*) and the p-value of the proposed approach against

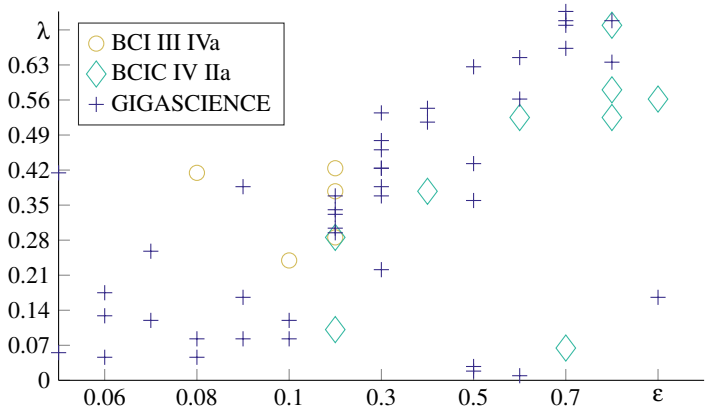


Fig. 1: Distribution of the tuned parameters λ and ε for each subject in the three datasets.

the competing methods for the three datasets. PFB corresponds to P_λFB without the feature selection stage, which only uses the component selection. Note that the general accuracy of PFB enhances CSP and FBCSP in the three datasets. Moreover, PFB outperforms SFB in GIGASCIENCE that holds a large number of subjects. In comparison with the baseline, P_λFB improves the accuracy up to 4.2%, 10.7% and 8.4% for BCICIII IVa, BCICIV IIa, and GIGASCIENCE, respectively. In addition, PFB and P_λFB maintains a percentage of components used for the feature extraction stage in 14.5% (BCICIII IVa), 65.2% (BCICIV IIa), and 32.4% (GIGASCIENCE), while maintaining a superior performance compared to FBCSP and SFBCSP that require all channels and sub-bands. Furthermore, the statistical test with 3% significance level has been conducted by comparing the proposed method against CSP, FBCSP, PFB and SFB to exhibit the significance of the proposed method.

Lastly, Fig. 3 illustrates the sub-bands and channels that were mainly selected for each dataset. Particularly for BCICIII IVa and GIGASCIENCE, frequency bands in the range $[8-16]$Hz predominate as they are related to higher mental functions (i.e. emotional expressions or to what extend subjects focus on) [1]. Further, the observed clusters imply that there exists relevant information concentrated in a specific combination of bands and channels.

4 Discussion and concluding remarks

Finding stimulated areas of the brain is very important to develop a highly performing BCI system. Since it is known that different parts of the brain are involved in different MI, actions, and feelings. Nonetheless, addressing this problem can be very hard due to inter-subject variability. In this study, an automatic method based in spatio-spectral relevance analysis supporting EEG discrimination is presented to find the bands and channels that contain the most discriminating and relevant information for MI task.

The proposed approach holds two free parameters, namely, the significance level of the statistical test and the λ regularization that trades-off the label regression and the

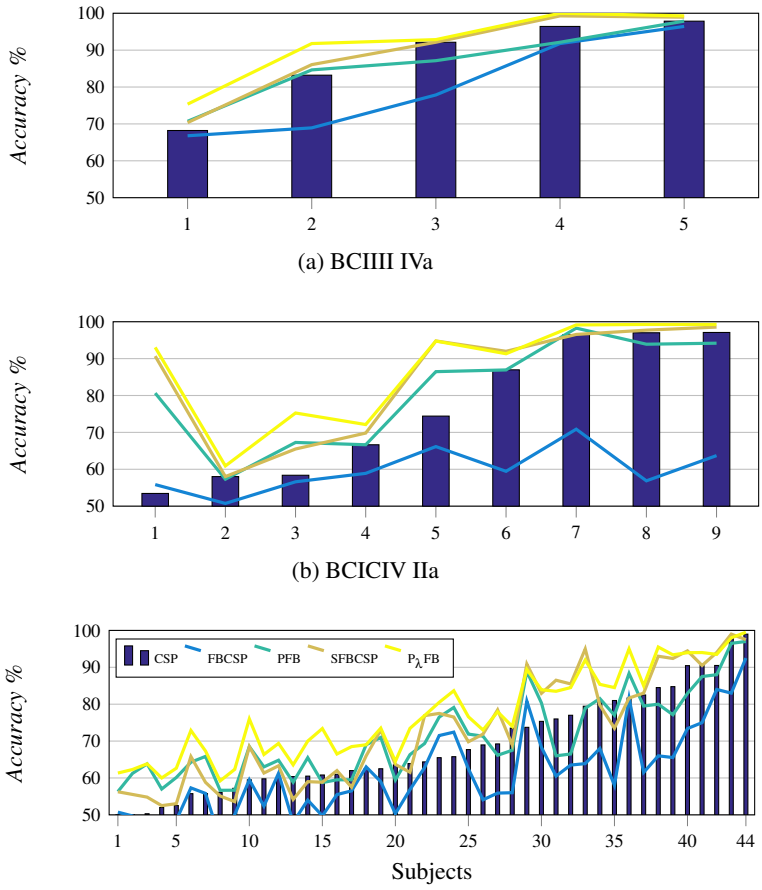


Fig. 2: Performed accuracy by each tested approach on the three datasets. Average of a five-fold cross validation is depicted. Subjects are sorted according the performed accuracy, note that CSP is the baseline approach.

component sparsity. The significance level sets a threshold to reject the null hypothesis that the difference of the mean entropies is zero. Hence, the smaller the ε - the more restrictive the spatio-spectral component selection. Subsequently, λ controls the number of selected CSP features, so that large regularization values exclude useful features, while no regularization yields overfitted high dimensional spaces. Results in Fig. 1 reveal that each subject demands its own parameter to decode the subject inherent variability. Therefore, the parameter tuning allows extracting only the most discriminative components for each subject according to its own training difficulties.

In order to evidence the reduction of spatio-spectral components, the Fig. 3 depicts the bands and channels selected for the three datasets evaluated, where the largest com-

	CSP	FBCSP	PFB	SFB	P_λFB
			BCICIII IVa		
$\mu \pm \sigma$	87.6 ± 4.2	80.4 ± 4.5	86.5 ± 4.9	89.3 ± 2.8	91.8 ± 2.5
p-value	$7.2e^{-4}$	$9.9e^{-7}$	$5.9e^{-4}$	$1.7e^{-2}$	-
crr	118/118	2006/2006	54/2006	2006/2006	298/2006
			BCICIV IIa		
$\mu \pm \sigma$	76.5 ± 5.6	59.9 ± 9.3	81.3 ± 5.7	84.8 ± 5.8	87.2 ± 4.3
p-value	$4.6e^{-6}$	$1.3e^{-15}$	$1.0e^{-6}$	$2.8e^{-2}$	-
crr	22/22	374/374	248/374	374/374	244/374
			GIGASCIENCE		
$\mu \pm \sigma$	68.9 ± 7.6	61.3 ± 7.3	71.0 ± 6.7	72.6 ± 6.3	77.3 ± 6.5
p-value	$1.2e^{-24}$	$2.6e^{-51}$	$7.9e^{-19}$	$2.4e^{-12}$	-
crr	64/64	1088/1088	382/1088	1088/1088	353/1088

Table 1: Classification accuracy for the considered approaches. Mean and standard deviation for five-fold cross-validation is computed, p-value is the statical test with 3% significance level between P_λFB and the other methods and crr is the component rate selection

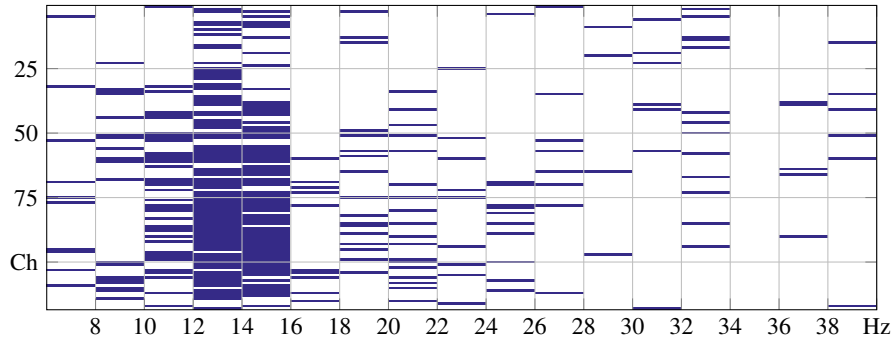
ponent rate reduction is achieved for the BCICIII IVa (14.5%) that originally holds 118 channels. In addition, P_λFB localizes the relevant information on physiologically interpretable bands and channels for the MI task. This is due to the spatio-spectral relevance analysis and the benefits presented by the PeEn for decoding the EEG dynamics.

The performance of PFB is compared against the baseline CSP, FB and SFB. The Fig. 2 presents the accuracy by subject and Table 1 summarizes the global performance achieved by each method where it is observed the statistical superiority of PFB over CSP and FB, while using less information. In order to compare with SFB, a feature selection based on Lasso regression was included after the feature extraction, termed PλFB. The proposed method improves the feature reduction, only demanding a few spatio-spectral components. These results reveal the importance of eliminating information that does not contribute to the task or that can be redundant, so validating the proposed methodology.

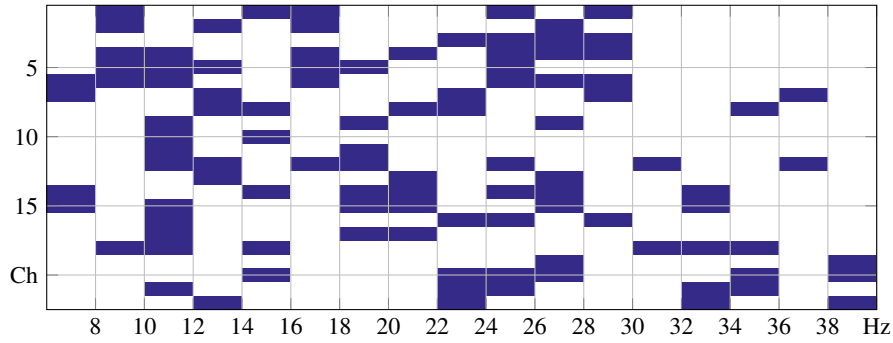
As a future work, we plan to construct sub-bands that can be adapted to each subject aiming to tackle the computational burden. Also, optimization of free parameters has to be studied using more elaborated strategies to reduce the training

Acknowledgements

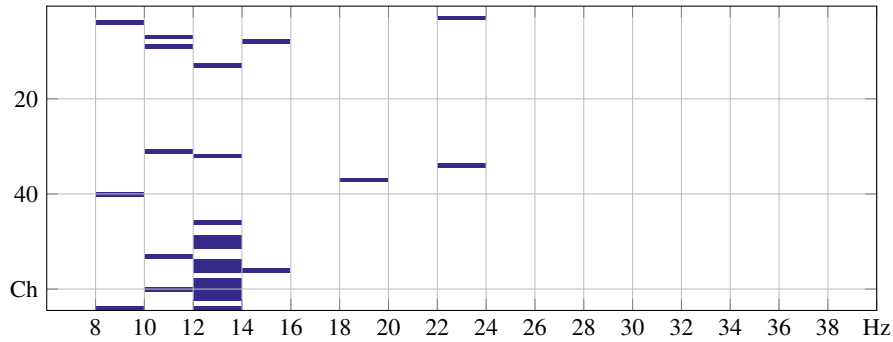
This work is supported by the project 111974454838 funded by COLCIENCIAS.



(a) BCIIII IVa



(b) BCICIV IIa



(c) GIGASCIENCE

Fig. 3: The most selected bands and channels for each dataset, the y-axis represents the channels and the x-axis the frequency bands.

References

1. Ahn, M., Jun, S.C.: Performance variation in motor imagery brain–computer interface: a brief review. Journal of neuroscience methods **243**, 103–110 (2015)

2. Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C.: Filter bank common spatial pattern (fbcsp) in brain-computer interface. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 2390–2397 (June 2008). https://doi.org/10.1109/IJCNN.2008.4634130
3. Cho, H., Ahn, M., Ahn, S., Kwon, M., Jun, S.C.: Eeg datasets for motor imagery brain–computer interface. GigaScience **6**(7), 1–8 (2017). https://doi.org/10.1093/gigascience/gix034
4. Higashi, H., Tanaka, T.: Simultaneous design of fir filter banks and spatial patterns for eeg signal classification. IEEE Transactions on Biomedical Engineering **60**(4), 1100–1110 (April 2013). https://doi.org/10.1109/TBME.2012.2215960
5. Novi, Q., Guan, C., Dat, T.H., Xue, P.: Sub-band common spatial pattern (sbcsp) for brain-computer interface. In: 2007 3rd International IEEE/EMBS Conference on Neural Engineering. pp. 204–207 (May 2007). https://doi.org/10.1109/CNE.2007.369647
6. Ortiz-Rosario, A., Adeli, H.: Brain-computer interface technologies: from signal to action. Reviews in the Neurosciences **24**(5), 537–552 (2013)
7. Sun, G., Hu, J., Wu, G.: A novel frequency band selection method for common spatial pattern in motor imagery based brain computer interface. In: The 2010 International Joint Conference on Neural Networks (IJCNN). pp. 1–6 (July 2010). https://doi.org/10.1109/IJCNN.2010.5596474
8. Thomas*, K.P., Guan, C., Lau, C.T., Vinod, A.P., Ang, K.K.: A new discriminative common spatial pattern method for motor imagery brain x2013;computer interfaces. IEEE Transactions on Biomedical Engineering **56**(11), 2730–2733 (Nov 2009). https://doi.org/10.1109/TBME.2009.2026181
9. Thomas, K.P., Guan, C., Lau, C.T., Vinod, A.P., K., K.: A new discriminative common spatial pattern method for motor imagery brain-computer interfaces. IEEE Transactions on Biomedical Engineering **56**(11), 2730–2733 (Nov 2009). https://doi.org/10.1109/TBME.2009.2026181
10. Zhang, Y., Wang, Y., Jin, J., Wang, X.: Sparse bayesian learning for obtaining sparsity of eeg frequency bands based feature vectors in motor imagery classification. International Journal of Neural Systems **27**(02), 1650032 (2017). https://doi.org/10.1142/S0129065716500325, pMID: 27377661
11. Zhang, Y., Zhou, G., Jin, J., Wang, X., Cichocki, A.: Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface. Journal of Neuroscience Methods **255**, 85 – 91 (2015). https://doi.org/https://doi.org/10.1016/j.jneumeth.2015.08.004

Entropy-based relevance selection of independent components supporting motor imagery tasks

No Author Given

No Institute Given

Abstract. Brain-Computer Interfaces provide an alternative control of devices through the human brain activity. This paper proposes a trial-wise channel filtering by selecting the subset of independent components with the largest entropy. The present work holds two free parameters: The order for the Renyi entropy weighs the component quantization according to its probability, and the percentage of retained entropy that rules the number of independent components to reconstruct the spatially filtered EEG channels. Both free parameters are tuned using a subject-dependent grid search for the best classification accuracy. The proposed approach outperforms against heuristic channel selection in a binary classification task using GigaScience BCI database. Relying on the obtained results, we prove that using ICA as a spatial filtering allows the feature extraction stage to build more discriminating spaces, reducing the influence of non-informative components. As an advantage, the resulting spatial filtering maintains the physiological interpretation of the EEG channels.

Keywords: Component selection; Renyi Entropy; Brain Computer Interface

1 Introduction

Brain-Computer Interfaces (BCI) provide an alternative device control through the human brain activity, aiming to improve life quality of people suffering from severe disabilities. Typical BCI systems demand the subjects under testing to perform specific mental tasks to produce particular electroencephalographic (EEG) responses, which are generated voluntarily (slow cortical potentials and sensorimotor rhythms) or by stimulation (event-related potentials and steady-state evoked potential). However, the task-unrelated noise and redundant information hamper the performance of BCI systems in practical applications [4]. Furthermore, commercial EEG devices with low spatial resolution cannot guarantee a near-optimal channel number and location [2].

Aiming to solve the above issues, heuristic algorithms select an optimal subset channel more related to the task with a higher noise-to-signal ratio. Some of the recent proposals are Glow Swarm Optimization algorithm followed by a naïve Bayes classifier [8], Sequential Floating Forward Selection by locally grouping EEG channels [13], Non-dominated Sorting Genetic Algorithm II for multi-objective optimization [10], and the backtracking search optimization algorithm by the binary encoding the selected channels [7]. Despite increasing the accuracy rates in comparison to the full channel set, heuristic and evolutionary algorithms heavily depend on the initialization of the algorithm due to the large hyperparameter set to be tuned. Besides, the high computational cost of the training stage restraints their use in practical applications.

On the contrary, information measures provide less costly and more accessible to optimize approaches for selecting channels. For instance, the mutual information between the Laplacian derivation of the average channel power and the task labels significantly enhances the performance of the feature extraction by common spatial patterns (CSP) [16]. The wavelet-based maximum entropy criterion in the spatio-temporal domain selects channels considerably decreases the number of channels with a minimum accuracy loss. This kind of selection approaches uses Shannon or quadratic entropy as the fundamental information measures, which correspond to the particular case of Renyi's entropy of order 1 and 2, respectively. Then, neglecting the influence of the entropy order reduces the flexibility and intensifies the biasing of the estimator [12]. Further, channel selection reduces the discriminating information that is hidden in a surrogate channel combination [3].

To overcome the channel selection shortcomings, this paper proposes to select a subset of independent components based on a Renyi entropy relevance analysis. The independent components are estimated using the WASOBI algorithm that takes advantage of the temporal data structure. Therefore, reconstruction of the EEG signal using the most informative components is expected to reduce the noise influence and improve the discrimination performance of the feature extraction and classification stages using CSP and Linear Discriminant Analysis.

2 Methods

2.1 Independent Component Analysis for MI Tasks

In MI tasks, Independent Component Analysis (ICA) separates spatial components from their measured mixture of scalp EEG signals, without accessing the activity sources themselves or knowing the mixing system. The linear ICA model is given by the model $\boldsymbol{x}(t)=\boldsymbol{A}\boldsymbol{s}(t)+\boldsymbol{\eta}(t)$, where $\boldsymbol{s}(t)\in\mathbb{R}^D$ are D independent components that vary along time $t\in\mathbb{R}$, $\boldsymbol{x}(t)\in\mathbb{R}^C$ are C mixed, measured EEG channels, and $\boldsymbol{A}\in\mathbb{R}^{C\times D}$ is the mixing matrix. Mainly, the ICA model assumes second-order stationarity, mutually uncorrelated components, and additive Gaussian noise, $\boldsymbol{\eta}(t)\in\mathbb{R}^D$.

Estimation of mixing matrix can be reformulated within a weighted least-squares problem, so that instead of computing $\boldsymbol{A}$ from D vectors, a small number of the time-lagged sample correlation matrices are used. This algorithm (termed WASOBI) exploits the time structure of the components as follows:

$$\begin{aligned}\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}}(\tau) &= \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\bar{\boldsymbol{x}}(t)\bar{\boldsymbol{x}}(t-\tau)^{\top} \\ &= \boldsymbol{Q}\boldsymbol{\Sigma}_{\boldsymbol{x}}(\tau)\boldsymbol{Q}^{\top},\ \tau\in[0,M-1] \end{aligned} \tag{1}$$

where $\boldsymbol{Q}\in\mathbb{R}^{C\times C}$ is an orthogonalization matrix, M is the number of time windows, and T is the time series length. In order estimate the spatially independent components $\hat{\boldsymbol{s}}(t)$, the WASOBI algorithm jointly diagonalizes the matrices $\boldsymbol{\Sigma}_{\boldsymbol{x}}(\tau)$ by non-linear weighted least squares, resulting in an asymptotically optimal weighting matrix for Auto-Regressive Gaussian sources that improve the ICA stability and reduces the computational cost [5].

2.2 Entropy-based relevance analysis

Given an estimate of the independent components $\hat{\boldsymbol{s}}(t)$ of a given EEG signal $\boldsymbol{x}(t)$, the Renyi entropy H_α of the d-th independent component $\hat{s}_d(t)$ is defined as:

$$H_\alpha(\hat{s}_d) = \frac{1}{1-\alpha}\log\int p^\alpha(\hat{s}_d)d\hat{s}_d \tag{2a}$$

$$\approx \frac{1}{1-\alpha}\log\sum_{b\in B} p_{db}^\alpha \tag{2b}$$

where $\alpha \geq 0$ stands for the entropy order. The probability distribution $p(\hat{s}_d)\in[0,1]$ is computed by quantization of the estimated components [11]:

$$p_{db} = \mathbb{E}\{\kappa(s_b, q(\hat{s}_d(t);\theta)) : \forall t \in T\} \tag{3}$$

where $\kappa(\cdot,\cdot)\in\mathbb{R}^+$ is a kernel function, and $q{:}\mathbb{R}\mapsto\mathbb{R}_\theta$ (provided with parameter set θ) is a given quantization or projection function, mapping $\hat{s}_d(t)$ to the nearest event in the subset of B bins $\{s_b\in\mathbb{R} : b\in[1,B]\}$. Notation $\mathbb{E}\{\cdot:\xi\}$ stands for the expectation operator across ξ.

Therefore, the entropy order allows weighting the events according to their probability [15]: H_0 only accounts the number of events with $p_{db} > 0$, whereas H_∞ only accounts the largest probability of p_d.

Upon the assumption of independence between components, the total amount of information held by the EEG signal can be computed as the summation of component entropies, that is, $\hat{H}_\alpha(\boldsymbol{x}){=}\sum_{d\in D}\hat{H}_\alpha(p_d)$. As a result, a relevance measure $\rho_d\in[0\%,100\%]$ is introduced that ranks each component according as:

$$\rho_d = 100\hat{H}_\alpha(p_d)/\hat{H}_\alpha(\boldsymbol{x}) \tag{4}$$

Consequently, the relevance measure enables reconstructing the EEG channels with a subset of components as:

$$\tilde{\boldsymbol{x}}(t) = \hat{\boldsymbol{A}}\boldsymbol{W}_\gamma\hat{\boldsymbol{s}}(t)$$

where $\tilde{\boldsymbol{x}}(t)\in\mathbb{R}^C$ is the reconstructed version of $\boldsymbol{x}(t)$, and $\boldsymbol{W}_\gamma\in\{0,1\}^{D\times D}$ is a diagonal matrix, indicating whether or not a component is selected to build $\tilde{\boldsymbol{x}}(t)$. The selection rule is as follows: $w_\gamma(d,d){=}\{1{:}\rho_d \geq \gamma; 0{:}\rho_d < \gamma\}$ where $\gamma\in(0,100]$ is a predefined relevance threshold. As a result, using a subset of the most relevant components, the performed reconstruction of the signal channels tends to regret the non-informative electrodes, enhancing the further processing stages.

3 Experimental Set-Up

EEG dataset and processing pipeline

The proposed approach of relevance-based component selection is evaluated on the online available datasets for motor imagery brain-computer interface with 52 subjects [6]. The dataset holds EEG data recorded by 64 Ag/AgCl active electrodes using the 10-10

montage at a sampling rate of 512,*Hz*. The BCI2000 recording system allows collecting EEG data and present the MI instructions. Each subject performed five or six runs of 20 trials per class. Each trial starts with a black screen with a fixation cross for 2,*s* as a ready instruction. Then one of the two classes ("left hand" or "right hand") randomly appear on the screen for 3,*s*, asking the subject to imagine the movement of the appropriate hand. Finally, the blank screen shows up as a break for a random period between 4.1 to 4.8,*s*. At the end of the recording, trials are labeled as *bad_trial* according to the voltage magnitude, its correlation with electromyographic activity, and the subject comments about the run. Due to above labeling critically reduces the amount of trials of seven subjects, we validate our proposed approach on the resting 45 subjects.

The trials recorded from each subject train a subject-dependent BCI system with four stages, namely, preprocessing, component selection, feature extraction, and classification. In the preprocessing stage, all EEG recordings are band-pass filtered between [8–30],*Hz* and downsampled at 100,*Hz* using a fifth order Butterworth filter to remove noise and slow baseline signal. We further extract the period between [0.5–2.5],*s* after the cue to focus on the learning part of the MI instruction. Then, each trial is decomposed according SEC2.1 and reconstructed using the resulting components from SEC2.2. Further, Common Spatial Patterns, one of the most efficient algorithms for extracting discriminating patterns from MI, maps each EEG covariance into a feature vector. Resulting trial-wise feature vector lastly feed an LDA-based classifier.

Parameter Tuning

Since the proposed component selection approach relies on an ICA model, we consider the WASOBI algorithm that takes advantage of the temporal structure of EEG data. Regarding the component selection, the following parameters are settled as the free parameters: *i*) Entropy order $\alpha \in \mathbb{R}^+$ that weighs the bin probability, in such a way, that large values ($\alpha > 1$) make the most probable bins contribute more to the entropy computation than the least probable ones. *ii*) Percentage of retained entropy $\gamma \in (0\%, 100\%]$ that rules the number of independent components to reconstruct the spatially filtered EEG channels so that the smaller the γ value - the fewer components are retained. Both free parameters are tuned using a subject-dependent grid search for the best classification accuracy on a fivefold cross-validation scheme.

Figure 1 displays the resulting tuning matrices for the two subjects with the lowest and the largest accuracy rates. Note that the asymptotic effect, at $\alpha = 1$, evenly distributes the entropy along the components so that larger γ values are required to select any reconstructing spatial component. Moreover, the best performing subjects (S14, S43) slowly increase their accuracy w.r.t. and the relevance threshold, implying that most of the components contribute to discriminate the motor imagery tasks. For these subjects, the optimal γ exceeds the 90% value. On the contrary, the percentage of retained entropy rejects more components for the subjects S17 and S38, aiming to attain the best performance. We hypothesize that this situation arises in the cases of worst-performing subjects when the criterion Eq. (4) tends to wrongly select some weak components, which may hold no discriminant information. As a result of the parameter tuning, EEG trials are reconstructed using only the most discriminant independent components for each subject.

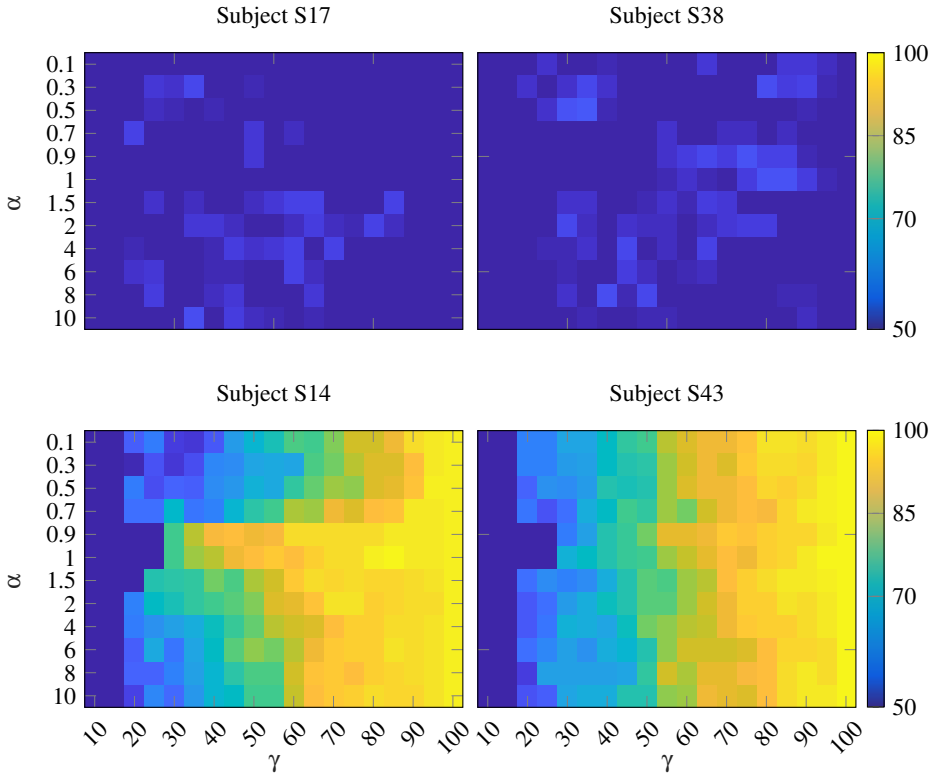
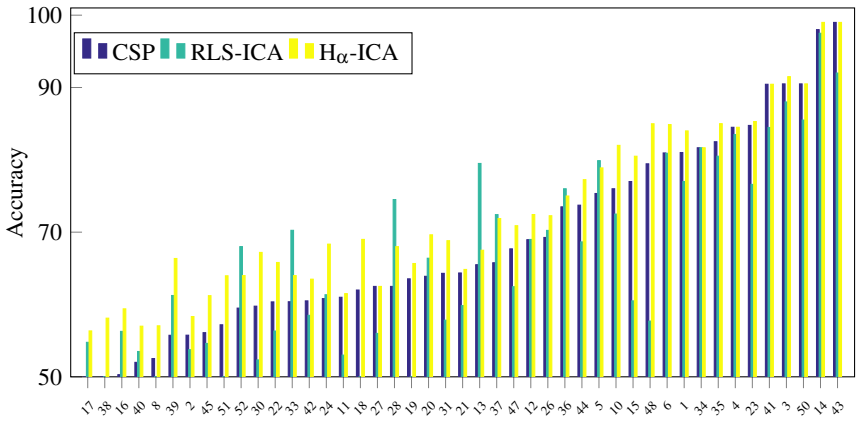


Fig. 1: Parameter tuning for the two worst performing subjects (top) and the two best performing subjects (bottom). Five-fold averaged accuracy is depicted along the percentaje of retained entropy and entropy order.

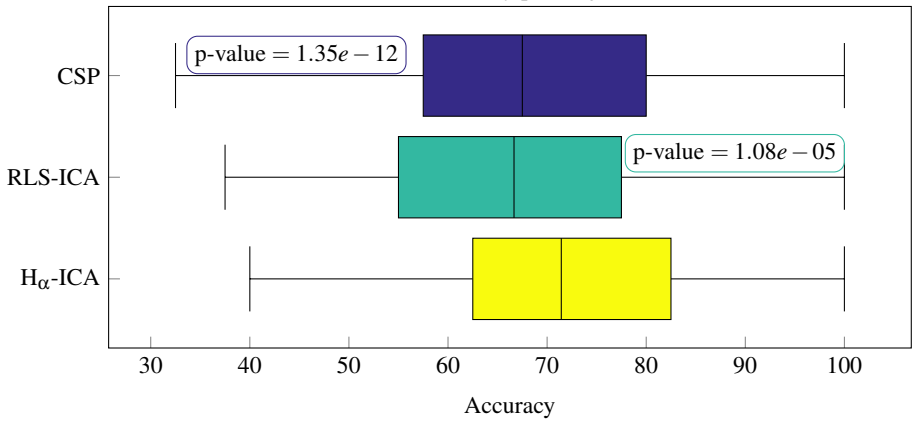
Performance assessment

The relevance analysis performance is assessed in terms of the classification accuracy for the left/right MI discrimination. For the sake of comparison, we also evaluate the performance of two baseline approaches: *i)* The standard Common Spatial Patterns (CSP), that consists in the widely accepted feature extraction approach for motor imagery tasks, that is carried out over the whole set of channels with no component filtering, i.e. γ=100%. *ii)* The component selection for removing eye movement artifacts using recursive least squares and ICA (RLS-ICA) [9]. The Figure 2a presents the classification performance attained by the considereded approaches for each subject. Aiming to gain interpretability on the performance improvement, subjects are ascending-ordered according the CSP accuracy. Results evidence that filtering independent components for each trial outperforms CSP and RLS-ICA for most of the subjects. In comparison to CSP, the proposed H_α-ICA largely improves subjects 17,38,39, for which a few components are retained; while no significant difference is attained for subjects S14 and S43, that hold most of their components. Regarding RLS-ICA, despite improving

the accuracy for some subjects, the artifact removal generally underperforms CSP and H_α-ICA. In addition, the fold-paired Student's t-test between the proposed component selection and the baseline approaches in Fig. 2b evidences a statistical difference with p-values lower than 0.001%. Therefore, the entropy-based relevance enhances the BCI performance by removing components un-related to the MI task.



(a) Performed accuracy per subject

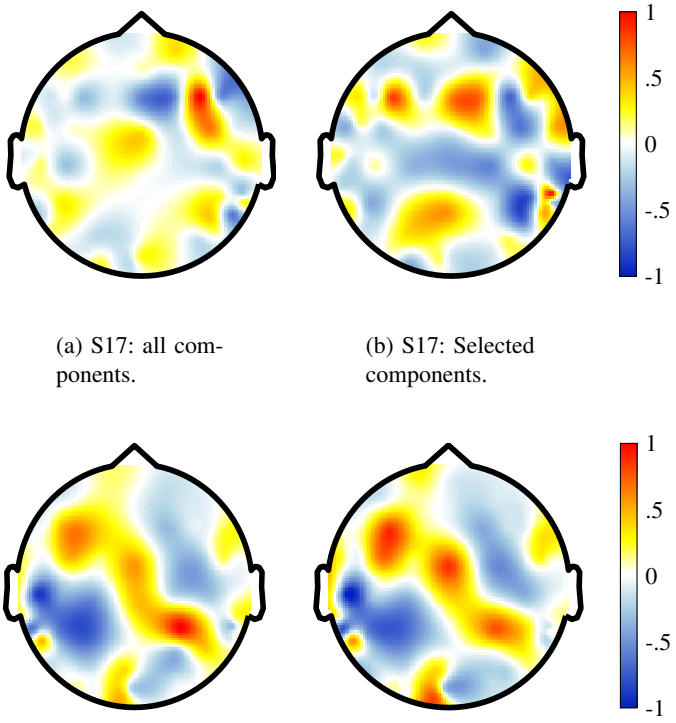


(b) Overall accuracy per approach

Fig. 2: Classification results attained by the considered approaches for each subject. Average and statistical test for five-folded cross validation are reported.

For illustrating the influence of entropy-based relevance on the CSP feature extraction, Fig. 3 depicts the spatial pattern distribution for the cases of the best and worst performing subjects, before and after performing the component selection. As seen, distribution of relevance varies from subject to subject. Thus, in the case of subject S43, the resulting topoplots show that the CSP feature set has no effect upon the pro-

posed approach of relevance selection. This result can be explained because the optimal γ reaches to 95%, retaining most of the components to reconstruct each EEG trial. In turn, for S17, the component selection redistributes all spatial patterns over the head due to the more restrictive γ value filters out the components that hinders the MI information. As a result, the optimal percentage of retained entropy allows the CSP-based feature set to better fit the discriminating patterns with no loss of relevant information.



(a) S17: all components.

(b) S17: Selected components.

(c) S43: all components.

(d) S43: Selected components.

Fig. 3: Spatial patterns for the worst (top) and best (bottom) performing subjects using all (left) and selected (right) components.

4 Discussion

The entropy order allows ranking the components following their measured probability. As seen in Fig. 1, the highest accuracy of the subjects with the worst performance (top row) is attained at low entropy orders. Hence, the least common bins contribute more to the relevance measure, implying that discriminant information concentrates on a few

time instants. On the contrary, the best performing subjects (bottom row) achieve the highest accuracy at high entropy orders. It means that the entropy depends on the most common bins, showing that the discriminant information holds within a longer time interval.

Using the entropy order as a free parameter allows to prioritize signal events according to its probability. Figure 1 evidences that the large entropy order for subjects S2 and S7 makes the most probable events contribute the most to the relevance. On the contrary, subjects S1 and S8 attain the best performance for low entropy values, so that unlikely events favor the most the activity discrimination. In another case, α lacks influence on subjects S3 and S8 since their trials are inherently more discriminable, as their accuracy proves. As a result, subject-wise adjusting the Renyi's entropy order allows taking into account the event distribution benefitting the most the motor imagery task.

Regarding the component selection criterion, the independence principle allows estimating the percentage of information that each of them holds. Therefore, one can define a subset of components to reconstruct the EEG channels by thresholding the computed entropy. Fig. 1 shows that, at least, 40% of the information is required to build a suitable reconstructing subset. Mainly, $\alpha \to 1$ uniformly distributes the component relevance. Then, a larger threshold is required to reconstruct any trial. Also, the directly proportional relationship, between the performance and the threshold for subjects S1, S3, S7, S8, and S9, implies that each of their components holds information contributing to the EEG discrimination. For the remaining subjects, the optimal performance is attained at a lower γ, meaning that there exist a large number of components hindering the task at hand. Therefore, selecting components based on the percentage of information will always outperform the classification without component selection as in Figure 2a is further illustrated.

Finally, our proposal takes advantage of the space of independent components so that the spatial filtering accounts for the discriminant information on each channel. As seen in Figure 2a, component selection outperforms channel selection since discarding channels lead to suboptimal accuracy when the discriminative information varies among them. Further, the approach will never underperform the baseline CSP, as the spatial filters in Figure 3 evidence. On the one hand, subjects with low accuracy (top) redistribute their spatial filters to improve the motor imagery performance. They show spatial patterns in frontal and prefrontal regions that are related to higher mental functions as concentration or emotional expressions associated with low MI perfomance[1]. On the other hand, subjects with high accuracy (bottom) maintain its spatial filters since all the components are relevant. MI tasks link several brain regions (i.e., the primary motor, the posterior parietal cortices, and the supplementary motor area) in charge of sensory processing, translate visual information into motor commands, generating mental movement representations, as well as, planning and coordinating motor tasks [14]. Then, using ICA as a spatial filtering stage allows CSP to find the more discriminating filters by reducing the influence of non-informative components.

5 Conclusion

This work introduces a component selection approach using a relevance analysis based on the Renyi entropy. To accomplish the task, the WASOBI algorithm exploits the temporal data structure to compute the independent components, and the entropy order and percentage of retained information are subject-wise tuned to improve MI discrimination. As an advantage, the resulting spatial filtering maintains the physiological interpretation of the EEG channels.

From the attained findings of the work, two future research directions are proposed: Optimization of free parameters using more elaborated strategies to reduce the training times without compromising the performance. Incorporation of prior information about the labels in the component selection stage to improve accuracy rates for low performing subjects.

References

1. Ahn, M., Jun, S.C.: Performance variation in motor imagery brain–computer interface: a brief review. Journal of neuroscience methods **243**, 103–110 (2015)
2. Al-Ani, A., Al-Sukker, A.: Effect of feature and channel selection on eeg classification. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2171–2174 (Aug 2006). https://doi.org/10.1109/IEMBS.2006.259833
3. Al-Ani, A., Al-Sukker, A.: Effect of feature and channel selection on eeg classification. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2171–2174 (Aug 2006)
4. Arvaneh, M., Guan, C., Ang, K.K., Quek, C.: Optimizing the channel selection and classification accuracy in eeg-based bci. IEEE Transactions on Biomedical Engineering **58**(6), 1865–1873 (June 2011). https://doi.org/10.1109/TBME.2011.2131142
5. Bridwell, D.A., Rachakonda, S., Silva, R.F., Pearlson, G.D., Calhoun, V.D.: Spatiospectral decomposition of multi-subject eeg: Evaluating blind source separation algorithms on real and realistic simulated data. Brain Topography **31**(1), 47–61 (Jan 2018). https://doi.org/10.1007/s10548-016-0479-1
6. Cho, H., Ahn, M., Ahn, S., Kwon, M., Jun, S.C.: Eeg datasets for motor imagery brain–computer interface. GigaScience **6**(7), 1–8 (2017). https://doi.org/10.1093/gigascience/gix034
7. Dai, S., Wei, Q.: Electrode channel selection based on backtracking search optimization in motor imagery brain-computer interfaces. Journal of Integrative Neuroscience **16**(3), 241–254 (feb 2017). https://doi.org/10.3233/JIN-170017
8. Franklin Alex Joseph, A., Govindaraju, C.: Channel selection using glow swarm optimization and its application in line of sight secure communication. Cluster Computing pp. 1–8 (sep 2017). https://doi.org/10.1007/s10586-017-1177-9
9. Guerrero-Mosquera, C., Navia-Vazquez, A.: Automatic removal of ocular artefacts using adaptive filtering and independent component analysis for electroencephalogram data. IET Signal Processing **6**(2), 99–106 (April 2012). https://doi.org/10.1049/iet-spr.2010.0135
10. Kee, C.Y., Ponnambalam, S.G., Loo, C.K.: Multi-objective genetic algorithm as channel selection method for P300 and motor imagery data set. Neurocomputing **161**, 120–131 (aug 2015). https://doi.org/10.1016/j.neucom.2015.02.057
11. Meinicke, P., Ritter, H.: Quantizing density estimators. In: Advances in Neural Information Processing Systems. pp. 825–832 (2002)

12. Principe, J.C.: Information theoretic learning: Renyi's entropy and kernel perspectives. Springer Science & Business Media (2010)
13. Qiu, Z., Jin, J., Lam, H.K., Zhang, Y., Wang, X., Cichocki, A.: Improved sffs method for channel selection in motor imagery based bci. Neurocomputing **207**, 519 – 527 (2016). https://doi.org/https://doi.org/10.1016/j.neucom.2016.05.035
14. Saiote, C., Tacchino, A., Brichetto, G., Roccatagliata, L., Bommarito, G., Cordano, C., Battaglia, M., Mancardi, G.L., Inglese, M.: Resting-state functional connectivity and motor imagery brain activation. Human brain mapping **37**(11), 3847–3857 (2016)
15. Wood, A.J., Blythe, R.A., Evans, M.R.: Rényi entropy of the totally asymmetric exclusion process. Journal of Physics A: Mathematical and Theoretical **50**(47), 475005 (2017)
16. Yang, H., Guan, C., Wang, C.C., Ang, K.K.: Maximum dependency and minimum redundancy-based channel selection for motor imagery of walking EEG signal detection. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 1187–1191. IEEE (may 2013). https://doi.org/10.1109/ICASSP.2013.6637838

Sub-band brain mapping based on a Multivariate Wavelet Packet Decomposition

P.A. Muñoz-Gutiérrez[1], J.D. Martinez-Vargas[2], E. Giraldo[3], G.Castellanos-Dominguez[4]

[1] Universidad del Quindío, Colombia
[2] Instituto Tecnológico Metropolitano, Medellín, Colombia,
[3] Universidad Tecnológica de Pereira, Pereira, Colombia,
egiraldos@utp.edu.co
[4] Signal Processing and Recognition Group, Universidad Nacional de Colombia, sede Manizales

Abstract. We introduce an improvement of the inverse problem solution for electroencephalographic source imaging (ESI) based on a Multi-signal Wavelet-Packets (MWP) decomposition. The method enhances the selection of relevant information of the MWP decomposition of the electroencephalographic (EEG) signals into its frequency bands using the best tree criteria. We adapt the proposed MWP method into several inverse solution algorithms including spatial and spatio-temporal prior information, namely, the Multiple Sparse Priors (MSP), and the Iterative Recursive Algorithm with l_1 norm (IRA-L1). The method is evaluated by using simulated EEG signals and real EEG signals corresponding to Evoked Response Potentials. From the obtained results, it can be concluded that the enhanced IRA-L1 method with the MWP stage improves the ESI performance in terms of the Eart Mover's Distance metric, in comparison with the other methods, for both simulated and real EEG signals.

Keywords: EEG inverse problem, Neuroimaging, Multi-signal Wavelet Packet

1 Introduction

Brain mapping has been turned into a mandatory tool for diagnosis assistance during medical examinations. EEG based brain mapping allows high temporal resolution which makes possible to include the temporal non-stationary dynamics of brain activity into the analysis. Moreover, when working with time–series with transient and non-stationary effects, relevant signal processing tools as the Short Time Fourier Transform (STFT) and Wavelet decomposition, are commonly used. In fact, time–frequency methods based on Morlet wavelet transforms have been proved to be effective when they are included in the inverse problem formulation [1]. In the other hand, since relevant neural activity associated to epilepsy seizures or Event Related Potentials (ERP) are spatially and temporal

located into the brain, the use of spatio-temporal dictionaries have been proved to enhance the quality of the estimation of neural activity, in particularly for the localization of non-stationary data such as ERPs [2].

The sparseness is an attractive property for studies where a small fraction of cortex is expected to be activated and the L_1-norm based regularization produce this kind of solutions for EEG/MEG inverse problems. However, most reports about L1-norm regularization for EEG/MEG inverse problems search the sparseness in the original source domain, leading over-focused solutions. To cope this issue, the recently proposed solutions have been addressed to solve EEG/MEG inverse problems using L_1-norm regularization by exploring the sparseness in transformed domains [3].

Since brain activity is highly localized in space, the property of sparse signal representation in the wavelet domain can be useful to improve the solutions to the dynamic inverse problem of ESI [4]. This can be done since the Wavelet transform (WT) is used as a multi-resolution analysis tool that can project the signal into different sub-bands with its own temporal and frequency features but guarantying perfect reconstruction [5]. For instance, the proposed approach in Ref. [6] introduces, in a unified framework, the temporal modeling using the discrete wavelet representation of both data and sources together with an spatial clustering of the brain activity in homogeneous parcels to localize time-varying oscillatory activity in MEG data. However the performance of the ESI decrease for deep or near sources since the method do not consider an adequate spatial constraint, and also it has a lower performance in terms of the ESI when considering oscillatory or spike signals since the Wavelet Transform reduced the temporal resolution for the EEG signal projected in each sub-band.

In this paper, a brain mapping improvement based on a Multi-signal Wavelet-Packet decomposition is proposed. The aforementioned method improves the selection of relevant information of the Multi-signal EEG decomposition into its frequency bands since the best tree representation is used. The resulting signal improves the brain mapping in terms of estimation of source activity. Additionally, the brain mapping can be performed for each sub-band where the activity is automatically projected or for a combination of sub-bands. We adopt the proposed improvement into two different ESI methods: Iterative Recursive Algorithm with l_1 norm (IRA-L1), and the Multiple Sparse Priors (MSP). The method is evaluated by using simulated EEG signals under two different simulated frameworks with realistic EEG activity, and real EEG signals corresponding to Evoked Response Potentials. This paper is structured as follows: In section 2 a detailed description of the mathematical methods proposed in this works are presented. In section 3 a comparison analysis is presented for simulated EEG signals for two different setups, and for real EEG signals. Finally, in section 5 the conclusions and final remarks are presented.

2 Materials and methods

2.1 Forward problem for EEG generation

Consider the forward problem for iterative EEG generation described as follows:

$$\boldsymbol{y}_k = \boldsymbol{M}\boldsymbol{x}_k + \boldsymbol{v}_k \tag{1}$$

where $\boldsymbol{y}_k$ is the EEG and $\boldsymbol{x}_k$ is the source activity at time sample k. This EEG generation can also be represented as:

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{\Upsilon} \tag{2}$$

where $\boldsymbol{Y}$ is a segment of EEG and $\boldsymbol{X}$ is a segment of neural activity for $k = 1, \ldots, N$.

2.2 Multi-signal Wavelet Packet

The EEG matrix $\boldsymbol{Y}$ can be projected in several sub-spaces $V_{(j,i)}$ by using a Multi-signal Wavelet Packet (MWP) decomposition, being $j = 0, \ldots, J$ the number of decomposition levels, and $i = 0, \ldots, 2^j - 1$ the number of sub-bands of each level, which satisfies that: $V_{(j,i)} = V_{(j+1,2^i)} \oplus V_{(j+1,2^i+1)}$. It can be seen that a reduced number of subspace can be selected in order to obtain a perfect reconstruction of the signal by using the best tree criteria. The subset of subspaces used for reconstruction can be obtained by any cost function, for example an entropy based cost function. Also an approximated reconstruction can be obtained based on the amount of retained energy, as following

$$\widehat{\boldsymbol{Y}} = \sum_{j \in E} \tilde{\boldsymbol{Y}}_j \tag{3}$$

being E the subset of subspaces which retained energy is over a threshold, and being $\tilde{\boldsymbol{Y}}_j$ the $j - th$ subspace projected in the same space of $\boldsymbol{Y}$.

Also, a subspace representation of the (2) can be defined as follows:

$$\boldsymbol{Y}_j = \boldsymbol{M}\boldsymbol{X}_j + \boldsymbol{\Upsilon}_j \tag{4}$$

which means that each j projection of the $\boldsymbol{Y}$ can be related directly to a projected activity $\boldsymbol{X}_j$. where:

$$\boldsymbol{X} = \sum_{j=0}^{J} \tilde{\boldsymbol{X}}_j \tag{5}$$

It can be seen that the subspace $\boldsymbol{X}_j$ must be projected into the original subspace $j = 0$ ($\tilde{\boldsymbol{X}}_j$), as mentioned in (3). Also, an approximated reconstruction can be obtained based on the amount of retained energy of each subspace as follows

$$\widehat{\boldsymbol{X}} = \sum_{j \in E} \tilde{\boldsymbol{X}}_j \tag{6}$$

being $\tilde{X}_j$ the projection of neural activity in the j-th subspace, and $\widehat{\boldsymbol{X}}$ the approximated reconstruction of neural activity. Two possibilities for analysis of neural activity: the first one is the brain mapping obtained in $\widehat{\boldsymbol{X}}$ where the proposed method includes a combination of the relevant information on several subspaces according to an entropy based cost function. The second one is the brain mapping obtained in $\tilde{\boldsymbol{X}}_j$ where the proposed method split the information of each sub-band. Therefore, an analysis by relevant sub-bands or by a combination of relevant sub-bands can be performed.

2.3 Inverse subspace problem

By considering (4) a inverse problem based on a subspace description of the EEG can be obtained as follows:

$$\underset{\boldsymbol{X}_j}{\text{minimize}} \quad \|\boldsymbol{Y}_j - \boldsymbol{M}\boldsymbol{X}_j\|_2^2 \tag{7a}$$

$$\text{subject to} \quad \|\boldsymbol{X}_j\|_1^1 \tag{7b}$$

$$\left\|\boldsymbol{X}_j - \widehat{\boldsymbol{X}}_{j-1}\right\|_2^2 \tag{7c}$$

where the solution is the reconstructed activity $\widehat{\boldsymbol{X}}_j$. It can be seen that the optimization problem (7) is solved for each subspace. Since this is a multi-signal multilevel decomposition, it can be seen that $\boldsymbol{Y}_j$ holds the same number of channels (rows) for each level that the original EEG $\boldsymbol{Y}$, but the number of columns (time samples) is down-sampled by a factor of 2^j according to the depth of decomposition. Therefore, the reconstructed activity $\widehat{\boldsymbol{X}}_j$ holds the same properties of $\boldsymbol{Y}_j$, which means that the number of sources is hold, but the time resolution is down-sampled.

3 Experimental set–up

In order to validate the inclusion of a MWP decomposition using the best tree criteria into brain mapping techniques, the performance of two methods IRA-L1 and MSP are compared with and without the use of MWP decomposition. By considering the underlying activity known by choosing the location of the active sources randomly into thee brain, an analysis of the subspace mapping is performed considering simulated EEG signals. An additional analysis is developed over real EEG signals where the neural activity is estimated through multi-modal solutions.

3.1 Simulated activity

For simulated signals, a realistic scenario of neural activity with a priori fixed location and/or stochastic like behaviors are used. To this end, we have into account the following considerations for underlying activity dynamics:

Setup 1 : Two sources randomly located into the brain with activity $\boldsymbol{x}_k$ in the range of 8 to 13Hz, with sampling rate of 100Hz which represent alpha oscillations [7]. In addition, 500 mutually statistically independent brain noise time series characterized by 1 / f-shaped (pink noise) power and random phase spectra are generated, and placed randomly at 500 locations sampled from the entire cortical surface. Finally, the EEG $\boldsymbol{y}_k$ is generated by multiplying the neural activity $\boldsymbol{x}_k$ by the leadfield matrix $\boldsymbol{M}$ as described in Eq. (1) and considering a SNR in the range of 0.1 to 0.9 percent.

Setup 2 : Three sources randomly located into the brain with activity in the range of 1 to 20Hz. In order to measure the performance of the method for several randomly selected source positions and robustness, 300 trials (60 for each SNR) with SNRs (0, 5, 10, 15 and 20dB) [3]. The activity in each source is generated according to the following expression: $x_i(t_k) = \exp\left(-\frac{1}{2}\left(\frac{t_k - c_i}{\sigma}\right)\right)^2 \sin\left(2\pi f_i t_k\right)$, being c_i the center of the windowed signal in seconds, and f_i the frequency of the signal, with $i = 1, \ldots, 3$. The c_i and f_i are selected in the following ranges $c_i : [0.5, 1.5]$ seconds and $f_i : [1, 20]$ Hz. The EEG is generated by multiplying $\boldsymbol{X}$ by the lead field matrix $\boldsymbol{M}$ as described in Eq. (2).

For these two approaches, the lead-field matrix is obtained from the so-called New York Head model as used in [7], which combines a highly detailed magnetic resonance (MR) image of the average adult human head with state-of-the-art finite element electrical modeling. In particular, the New York Head model holds $n = 2004$ sources and $d = 108$ electrodes.

3.2 Real EEG database

Further evaluation of the proposed algorithms, IRA-L1 and MSP with and without the MWP decompositions, is performed on real EEG signals. The analysis performed over real EEG signals where the underlying neural activity is assumed relying upon multi-modal solutions. For this case, eighteen healthy young adults (eight female) were drawn from the MRC Cognition and Brain Sciences unit Volunteer Panel. There were 300 different faces and 150 different scrambled faces. Of the faces, 150 were from famous people and 150 were from unfamiliar (previously unseen) people. Each face or scrambled face was either repeated immediately or after a lag of 5-15 intervening items. From this data-set, ERP of 15 subjects are selected for each one of the 3 stimulus [8]. It is noticeable that each patient has his own forward model (lead-field matrix).

3.3 Performance measure

The performance of the proposed algorithm is evaluated by comparison of the IRA-L1 [3] and the Multiple Sparse Priors [9] with and without the MWP decomposition. To this end, the Wasserstein metric, also known as the Earth Mover's Distance (EMD) is used. The EMD is a measurement of the amount of energy required to move the estimated brain mapping solution to the original simulated neural activity [3].

4 Results

We contrast the above-mentioned ESI scenarios, regarding their ability to localize the actual location of the simulated sources, namely, the spatial resolution throughout the EMD as the performance measure.

Fig. 1 provides an insight of the temporal (left) and frequency (right) dynamics of the simulated sources on the Setup 1. It can be seen that the source activity appears mainly in the range of 8 to 13 *Hz*. Hence, the selection of relevant information should be able to localize the main EEG generators inside such frequency range.

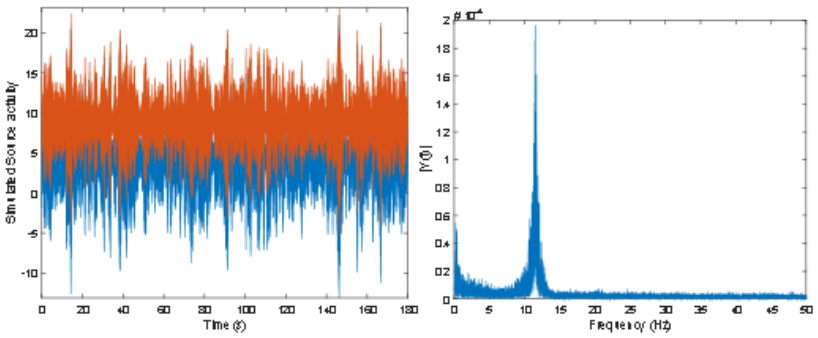


Fig. 1: Simulated activity for two sources according to the Setup 1

Fig. 2 shows the MWP decomposition of the aforementioned data, where the EEG recording $\boldsymbol{Y}$ is projected into several sub-spaces $\boldsymbol{V}(j,i)$, being $j = 0, \cdots, J$ the number of decomposition levels, and $i = 0, \cdots, 2^j - 1$ the number of sub-bands for each level. In this particular case, the algorithm sets automatically $J = 4$, according to the retained energy criterion, i.e., the retained energy at each subspace should not exceed the 50% of the total signal energy. Once the number of decomposition levels has been estimated, we calculate the best tree using the Shannon entropy of each subspace as cost function. As a result, we obtain that the EEG recording can be decomposed into the following subspaces: $V_{(0,0)} = V_{(1,1)} \oplus V_{(2,1)} \oplus V_{(3,0)} \oplus V_{(4,2)} \oplus V_{(4,3)}$.

However, almost the 90% of the EEG total energy is retained in subspaces $\boldsymbol{V}_{(4,2)}$ and $\boldsymbol{V}_{(4,3)}$, which correspond to the frequency band between 6.25 and 12.5 *Hz*, that certainly comprises most of the simulated activity. Accordingly, we corroborate that the proposed MWP decomposition has an inherent property of automatic sub-bands selection, as it is able to determine, based on energy and entropy criteria, the relevant frequency bands that contribute to the studied

neural processes. Once the best tree is estimated, the ESI is carried out employing solely the selected sub-spaces.

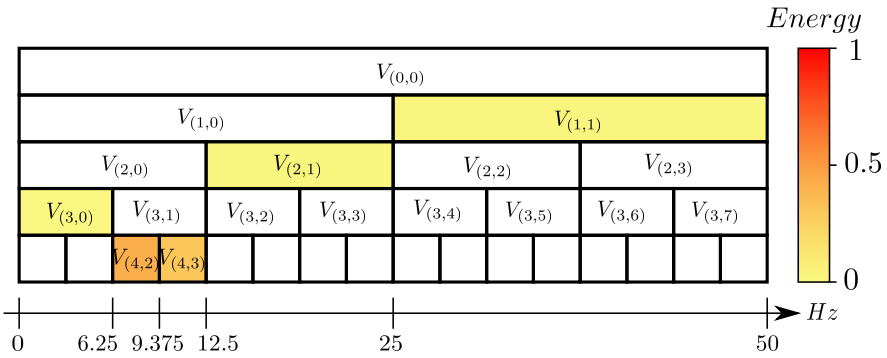


Fig. 2: MWP decomposition and best tree selection by using Shannon entropy

The second set-up is designed to test the performance of the proposed method under non-stationary brain activity conditions. Thus, Fig. 3 shows an example of three simulated non-stationary source and the obtained EEG. In the example, the sources are created with central frequencies of 4 (red), 10 (blue), and 16 (orange) central frequencies.

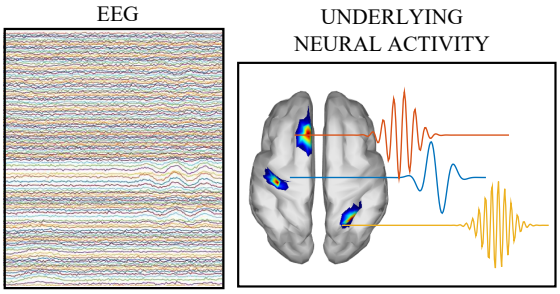


Fig. 3: EEG and neural activity of three randomly located sources with different frequencies.

Fig. 4 shows the obtained MWP decomposition, where the sub-spaces $\boldsymbol{V}_{(4,1)}$, $\boldsymbol{V}_{(4,2)}$, and $\boldsymbol{V}_{(4,6)}$ with central frequencies of 6.25, 9.375, and 18.75 *Hz*, respectively, comprise a total energy greater than the 90% of the total EEG energy. As a result, obtained source reconstruction related to each subspace clearly correspond to a simulated source. Thus, under non-stationary conditions, the proposed method is able to optimally estimate the frequency bands of interest.

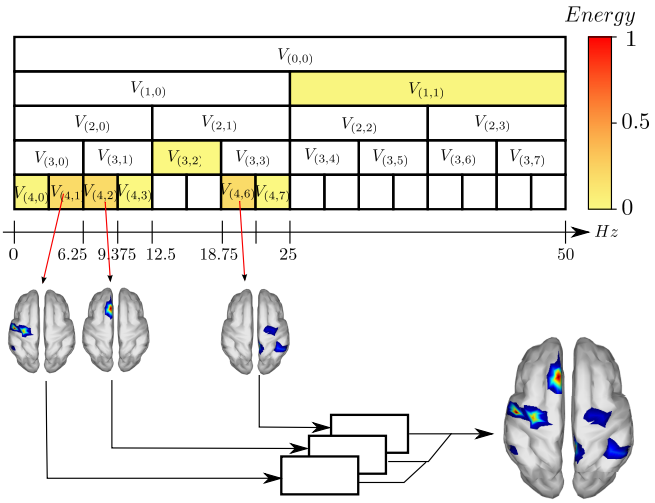


Fig. 4: Wavelet Packets decomposition for three sources

To generalize the performance of the proposed methods in the setup 2, Fig. 5 shows the EMD mean and standard deviation after 60 repetitions at each SNR level. For each repetition, source parameters, namely, location, and central frequency, are randomly selected. Once again, results show that ESI methods enhanced with the MWP decomposition step achieves the best results (lower EMD values). However, in this particular case with non-stationary sources, IRA-L1-WP overcomes the remaining comparison methods. As a result, IRA-L1-WP deals with the data non-stationarity in two different ways: i) by splitting the time-varying EEG spectrum into several sub-spaces using the MWP step, and ii) by imposing smooth temporal transitions within the estimated sources, yielding an improved non-stationary source reconstruction.

We apply all the compared ESI methods to the averaged ERP time series of each subject for each of the three stimulus conditions, namely, faces, famous faces, and scrambled faces. Moreover, we use as ground truth the source reconstruction using EEG and MEG multimodal data. Hence, we are able to calculate the EMD between the ground truth and the estimated sources with each ESI method. Fig. 6 shows the EMD mean and standard deviation for the 15 subjects under each stimulus condition. Once again, methods with the MWP step obtain the lower EMD, which means closer reconstruction to the ground truth. Furthermore, as ERPs comprise non-stationary brain activity, IRA-L1-WP achieves the best performance.

5 Conclusions

A novel inverse problem solution for electroencephalographic source imaging is presented based on a multi-signal Wavelet-Packets decomposition with best

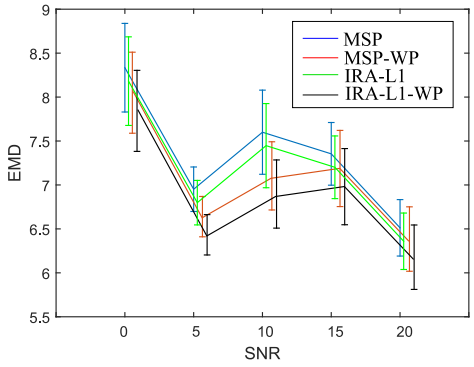


Fig. 5: Setup 2 EMD comparison for 60 repetitions of three randomly located sources at each SNR level

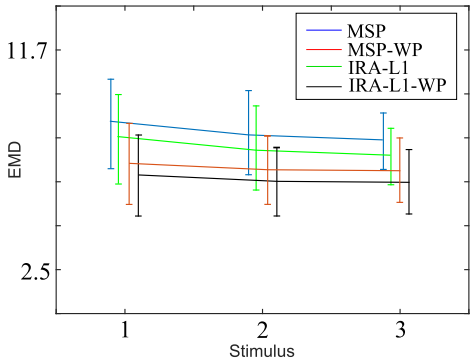


Fig. 6: EMD measure asumming multimodal solution as the ground truth.

tree selection of subspace decomposition. The method improves the selection of relevant information of the Multi-signal EEG decomposition into its frequency bands since an automatic method based on the entropy cost function is applied. It can be concluded that the proposed method has two inherent properties: automatic subspace selection based on relevant EEG sub-bands information by using an entropy based cost function, sub-band brain mapping based on a sub-band dynamic inverse problem solution. As shown in the results, the proposed method can efficiently select the relevant subspaces where the activity is spatially and temporal located, an therefore the associate neural activity for each sub-band can be mapped into the brain. Therefore, the proposed method can be used in applications where the relevant information can be found in specific sub-bands which are automatically selected by the entropy based cost function.

An additional property associated to the sub-bands decomposition is the reduction of columns (time samples) for each level of decomposition according to the down-sampling property, where the reconstructed activity is also down-sampled in time. Even when the time resolution is diminished by applying the

successive projections on the adequate subspaces, the time resolution can be perfectly recovered. Therefore, the proposed method allow a sub-band analysis that can be useful for connectivity based methods where the information of interest is only in some frequency bands.

6 Acknowledgments

This work was carried out under the funding of the Departamento Administrativo Nacional de Ciencia, Tecnología e Innovación (Colciencias). Research project: 111077757982 Sistema de identificación de fuentes epileptogénicas basado en medidas de conectividad funcional usando registros electroencefalográficos e imágenes de resonancia magnética en pacientes con epilepsia refractaria: apoyo a la cirugía resectiva.
This work is also part of the research project "Solución del problema inverso dinámico considerando restricciones espacio-temporales no homogéneas aplicado a la reconstrucción de la actividad cerebral" funded by the Universidad Tecnológica de Pereira under the code E6-17-2.

References

1. Gramfort, A., Strohmeier, D., Haueisen, J., Hmlinen, M., Kowalski, M.: Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. NeuroImage **70**(0) (2013) 410 – 422
2. Castaño-Candamil, S., Höhne, J., Martínez-Vargas, J.D., An, X.W., Castellanos-Domínguez, G., Haufe, S.: Solving the eeg inverse problem based on space–time–frequency structured sparsity constraints. NeuroImage **118** (2015) 598–612
3. Giraldo, E., Martinez-Vargas, J., Castellanos-Dominguez, G.: Reconstruction of neural activity from EEG data using dynamic spatio-temporal constraints. International Journal of Neural Systems **26**(7) (2016) 16500261–165002615
4. Behjat, H., Leonardi, N., Srnmo, L., Ville, D.V.D.: Anatomically-adapted graph wavelets for improved group-level fmri activation mapping. NeuroImage **123**(Supplement C) (2015) 185 – 199
5. Liao, K., Zhu, M., Ding, L.: A new wavelet transform to sparsely represent cortical current densities for eeg/meg inverse problems. Computer Methods and Programs in Biomedicine **111**(2) (2013) 376 – 388
6. Lina, J., Chowdhury, R., Lemay, E., Kobayashi, E., Grova, C.: Wavelet-based localization of oscillatory sources from magnetoencephalography data. IEEE Trans. Biomed. Engineering **61**(8) (2014) 2350–2364
7. Haufe, S., Ewald, A.: A simulation framework for benchmarking eeg-based brain connectivity estimation methodologies. Brain Topography (2016) 1–18
8. Henson, R., Wakeman, D.G., Litvak, V., Friston, K.J.: A parametric empirical bayesian framework for the eeg/meg inverse problem: Generative models for multi-subject and multi-modal integration. Frontiers in Human Neuroscience **5**(76) (2011) 1–16
9. Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J.: Multiple sparse priors for the m/eeg inverse problem. NeuroImage **39**(3) (2008) 1104–1120

Localizing the Focal Origin of Epileptic Activity using EEG Brain Mapping based on Empirical Mode Decomposition

M. Bueno-López[1], P.A. Muñoz-Gutiérrez[2], E. Giraldo[3], Marta Molinas[4]

[1] Department of Electrical Engineering, Universidad de la Salle, Bogotá, Colombia, maxbueno@unisalle.edu.co
[2] Electronic Instrumentation Technology, Universidad del Quindío, Armenia, Colombia, pamunoz@uniquindio.edu.co
[3] Department of Electrical Engineering, Universidad Tecnológica de Pereira, Colombia,egiraldos@utp.edu.co
[4] Department of Engineering Cybernetics
Norwegian University of Science and Technology, Norway marta.molinas@ntnu.no

Abstract. Empirical Mode Decomposition (EMD) is an adaptive time-frequency analysis method, which is very useful for extracting information from noisy nonlinear or nonstationary data. The applications of this technique in Biomedical Signal analysis has increased and is now common to find publications that use EMD to identify behaviors in the brain or heart. In this paper, a novel identification method of relevant Intrinsic Mode Functions (IMFs), obtained from EEG signals, using an entropy analysis is proposed. The idea is to reduce the number of IMFs that are necessary for the reconstruction of neural activity. The entropy cost function is applied on the IMFs generated by the EMD in order to automatically select the IMFs with relevant information. A relative error measure has been used to validate our proposal.

Keywords: Brain mapping, Empirical Mode Decomposition, Epilepsy, Signal Analysis.

1 Introduction

The EEG is used to measure the electrical activity of the brain characterizing its various normal and pathological states. Due to their non-linear and nonstationary nature, these signals are very difficult to analyze in the time-frequency frame. However, important features can be extracted for assisting in the analysis of Alzheimer's disease, attention-deficit/hyperactivity disorder (ADHD), autism, autistic spectrum disorder, alcoholism, epilepsy, depth of anesthesia, etc. [1], [2], [3]. EEG is usually pre-processed by pass-band and stop-band filters that can modify some features of the EEG signals. The nature of EEG signals has allowed to use EMD and Hilbert Huang Transform (HHT) to obtain a better signal representation [4]. In [5], the authors show an evolution of EMD called Multivariate Empirical Mode Decomposition (MEMD), which

is very useful in direct multichannel data analysis; in that case used for a full data-driven analysis to decompose resting-state fMRI (functional Magnetic Resonance Imaging) data into different sub-bands looking for connectivity functions. The use of fMRI implies higher costs due to the equipment required for acquisition and processing of information. Different strategies have been used for the process of reconstruction of Neural Activity from EEG data, but to the best knowledge of the authors, EMD has not been used for this purpose. For neural activity reconstruction, an iterative regularized method that explicitly includes space (grounded in a physiological model) and time constraints within the dynamic solution of the EEG inverse problem, is presented in [6].
One of the main applications of the signal analysis in Bioenineering is the support in the detection of diseases. One of theses fields correspond with epilepsy. The treatment of focal epilepsy when dopaminergic drugs are not effective, consist in a surgery where a part of the brain is removed. The surgery is performed after carefully locating the sources or brain zones that initiate the epileptic seizures. Highly precise brain mapping techniques are required to performed this task. However, when precise mapping is not available, an additional surgery is performed in order to obtain additional estimation of the zone that has to be removed by using intra-cranial electrodes [7]. In this paper, an improvement of brain mapping techniques is performed by using a pre-processing stage of the EEG. The main goal is to develop a highly precise brain mapping method that classifies the information in frequency bands from the IMFs. In this way, it will be possible to more precisely locate the area in the brain where the epileptic seizures occur. To this end, an optimal selection of IMFs is performed based on an entropy cost function. Some previous works have considered the use of entropy to detection of epileptic seizure [8], [9]. We applied EMD in simulated EEG signals containing normal and epileptic activity based on a nonlinear complex model. The simulated EEG signals are generated for one active source under several noise conditions. In order to detect the sources associated to epileptic seizures a brain mapping stage is performed using a reconstructed EEG signals obtained from the selected IMFs according to the entropy cost function.
This paper is organized as follows: Section 2 gives an introduction to the essential concepts about EMD and EEG signals. The experimental setup is presented in Section 3 and the results obtained with the EEG signals are shown in Section 4. The discussion of the results is presented in Section 5. Finally, some conclusions are given in Section 6.

2 Methods

2.1 Brain Mapping: The Inverse Problem

The forward problem of EEG generation can be formulated

$$\boldsymbol{y}(t_k) = \boldsymbol{M}\boldsymbol{x}(t_k) + \boldsymbol{\epsilon}(t_k) \tag{1}$$

being $\boldsymbol{y}(t_k) \in \mathbb{R}^d$ the EEG, and $\boldsymbol{x}(t_k) \in \mathbb{R}^n$ the neural activity, with $t_k = kh$ the time at sample k being $k = 1, \ldots, T$ the total number of samples, h the sample

time and $\boldsymbol{M} \in \mathbb{R}^{d \times n}$ the lead-field matrix that relates the neural activity with the EEG. The evolution of $\boldsymbol{x}(t_k)$ in time can be used to model several behaviors of the EEG. It is possible to formulate an iterative inverse problem [6] in order to estimate the neural activity $\widehat{\boldsymbol{x}}(t_k)$ for each measurement $\boldsymbol{y}(t_k)$, as described in:

$$\begin{aligned}\widehat{\boldsymbol{x}}(t_k) = \arg\min_{\boldsymbol{x}(t_k)} \|\boldsymbol{y}(t_k) - \boldsymbol{M}\boldsymbol{x}(t_k)\|_2^2 + \lambda_k \|\boldsymbol{x}(t_k) - \widehat{\boldsymbol{x}}(t_{k-1})\|_2^2 \\ + \alpha_k \|\boldsymbol{x}(t_k)\|_1\end{aligned} \tag{2}$$

being λ_k and α_k the regularization parameters computed by generalized cross validation [6].

2.2 Empirical Mode Decomposition

The Empirical Mode Decomposition (EMD) is an adaptive, and data-dependent method. The aim of the EMD method is to decompose the nonlinear and non-stationary signal $\boldsymbol{y}(t_k)$ into a sum of intrinsic mode functions (IMFs) that satisfies two conditions [10]:

1. The number of extrema and the number of zero crossings must be the same or differ at most by one.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Empirical Mode Decomposition is applied over $\boldsymbol{y}(t_k)$ to obtain $\boldsymbol{\gamma}_i(t_k)$ being i the intrinsic mode function (IMF), and

$$\boldsymbol{y}(t_k) = \sum_{i=1}^{N} \boldsymbol{\gamma}_i(t_k) + \boldsymbol{r}(t_k) \tag{3}$$

where N is the number of IMFs and $\boldsymbol{r}(t_k)$ a residual. Recently, some optimization techniques have been proposed to improve the performance of the EMD [11], [12].

Having obtained the intrinsic mode function components, we can apply the Hilbert transform to each component, and compute the instantaneous frequency according to equation (4).

$$f_i(t) \triangleq \frac{1}{2\pi} \cdot \frac{d\theta_i(t)}{dt}, \tag{4}$$

where $\theta_i(t)$ is the function phase of each IMF calculated from the analytical signal associated [13]. Finally, the instantaneous frequency can be observed in the Hilbert Spectrum.

2.3 Automatic IMF selection: Entropy Function

An entropy based cost function is applied over each IMF $\boldsymbol{\gamma}_i(t_k)$ as follows:

$$e_i = -\sum_{k} \|\boldsymbol{\gamma}_i(t_k)\|_2^2 \log(\|\boldsymbol{\gamma}_i(t_k)\|_2^2) \tag{5}$$

being e_i the entropy of each IMF, and $\boldsymbol{e} = [e_1 \ldots e_N]$. In order to reconstruct the EEG signal $\tilde{\boldsymbol{y}}(t_k)$ the automatic selection of the IMFs (IMFs with highest entropy) is applied according to the measured entropy e_i.

$$\tilde{\boldsymbol{y}}(t_k) = \sum_{i \in O} \boldsymbol{\gamma}_i(t_k) \tag{6}$$

being O the subset of of IMFs whose entropy e_i is over a threshold τ_e computed as follows

$$\tau_e = \frac{\max \boldsymbol{e} - \min \boldsymbol{e}}{2} + \min \boldsymbol{e} \tag{7}$$

3 Experimental setup

The performance of the aforementioned method is evaluated by using simulated signals with epileptic activity. The experimental setup is depicted in Fig. 1.

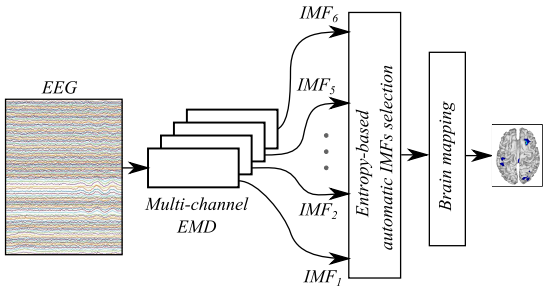


Fig. 1. Experimental setup for entropy-based automatic selection of IMFs

Two methods are considered for brain mapping in order to evaluate the performance of the algorithm:

1. Brain mapping ($\widehat{\boldsymbol{x}}(t_k)$) using the EEG database $\boldsymbol{y}(t_k)$ without EMD.
2. Brain mapping ($\tilde{\boldsymbol{x}}(t_k)$) using the reconstructed EEG $\tilde{\boldsymbol{y}}(t_k)$ obtained from EMD standard decomposition and an entropy based IMF selection.

A common procedure to evaluate the performance of brain mapping techniques is by using simulated EEG signals where the underlying brain activity is known. In this case, a measure of the brain mapping quality can be evaluated with the relative error measure [14] as follows:

$$e_s = \sum_k \frac{\|\widehat{\boldsymbol{x}}(t_k) - \boldsymbol{x}(t_k)\|_2^2}{\|\boldsymbol{x}(t_k)\|_2^2} \tag{8}$$

$$e_r = \sum_k \frac{\|\tilde{\boldsymbol{x}}(t_k) - \boldsymbol{x}(t_k)\|_2^2}{\|\boldsymbol{x}(t_k)\|_2^2} \tag{9}$$

being e_s the reconstruction error of the brain mapping estimation $\widehat{\boldsymbol{x}}(t_k)$ resulting from $\boldsymbol{y}(t_k)$, e_r the reconstruction error of the brain mapping estimation $\tilde{\boldsymbol{x}}(t_k)$ resulting from $\tilde{\boldsymbol{y}}(t_k)$.

For the first simulated database (SD-1) a complex nonlinear model of neural activity is used for EEG generation during an epileptic seizure based on [15] as follows

$$\begin{aligned}\boldsymbol{x}(t_k) &= \boldsymbol{A}_1\boldsymbol{x}(t_{k-1}) + \boldsymbol{A}_2\boldsymbol{x}(t_{k-2}) \\ &+ \boldsymbol{A}_3\boldsymbol{x}(t_{k-\tau}) + \boldsymbol{A}_4\boldsymbol{x}(t_{k-1})^{\circ 2} + \boldsymbol{A}_5\boldsymbol{x}(t_{k-1})^{\circ 3} + \boldsymbol{\eta}(t_k)\end{aligned} \tag{10}$$

being $\boldsymbol{A}_1 = a_1\boldsymbol{I}_n$, $\boldsymbol{A}_2 = a_2\boldsymbol{I}_n$, $\boldsymbol{A}_3 = a_3\boldsymbol{I}_n$, $\boldsymbol{A}_4 = a_4\boldsymbol{I}_n$ and $\boldsymbol{A}_5 = a_5\boldsymbol{I}_n$, where $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix and $a_i \in \mathbb{R}$ are the model parameters which describe the dynamics of the brain activity, where $\boldsymbol{c}_{k-1}^{\circ 2}$ denotes the Hadamard Power. The model parameter are set to $\tau = 20$, $a_1 = 1.0628$, $a_2 = -0.42857$, $a_3 = 0.008$, $a_4 = 0.000143$, $a_5 = -0.000286$, and $\|\boldsymbol{\eta}(t_k)\| \leq 0.05$. The epileptic seizure is simulated at time $t_k = 0.5\,s$ by modifying the values of a_1 from 1.0628 to 1.3, while a_2 from -0.428 to -1 over the entire diagonal. The simulated EEG $\boldsymbol{y}(t_k)$ is obtained from $\boldsymbol{x}(t_k)$ using (1) where $\boldsymbol{\epsilon}(t_k)$ is set to achieve the Signal-to-Noise Ratios (SNRs) of 0, 5, 10, 15 and 20 dB, the sample rate is 250Hz, and a number of $d = 128$ electrodes and $n = 8196$ sources are considered.

For the second simulated database (SD-2), the epileptic seizure is simulated at time $t_k = 0.5\,s$ by using a sinusoidal signal with frequencies linearly varying in the range of 8 to 12Hz [6]. The simulated EEG $\boldsymbol{y}(t_k)$ is obtained from $\boldsymbol{x}(t_k)$ using (1) where $\boldsymbol{\epsilon}(t_k)$ is set to achieve the Signal-to-Noise Ratios (SNRs) of 0, 5, 10, 15 and 20 dB. The sample rate is 500Hz, and a number of $d = 128$ electrodes and $n = 8196$ sources are considered.

4 Results

The analyzed signals are nonlinear and non-stationary, therefore the EMD is a good alternative to obtain information. An example of the EEG signal for SD-1 for a SNR of 0 dB is shown in Fig. 2.

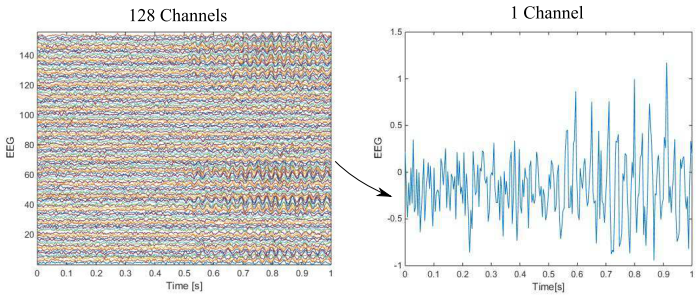


Fig. 2. Simulated EEG for SD-1

After analyzing the database with the EMD, we obtained 6 IMFs per channel. In the IMF 2 in Fig. 3, it is possible to observe two areas in red that show how different frequencies (different oscillations) appear in the same IMF. In these IMFs the mode mixing problem is evident.

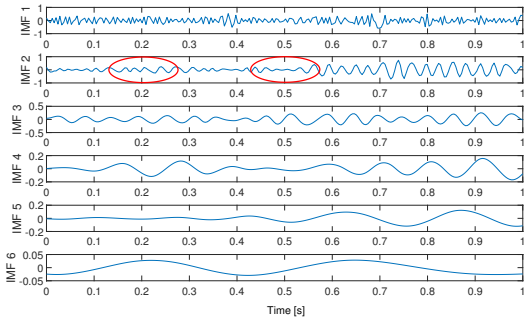


Fig. 3. IMFs of y_s for SD-1 using standard EMD

An example of the Hilbert spectrum is presented in Fig. 4, it is possible to see how the instantaneous frequency is changing. As expected, it is observed that the highest frequency is in IMF 1. A comparison of the original $\boldsymbol{y}(t_k)$ and reconstructed $\tilde{\boldsymbol{y}}(t_k)$ signals is presented in Fig. 5.

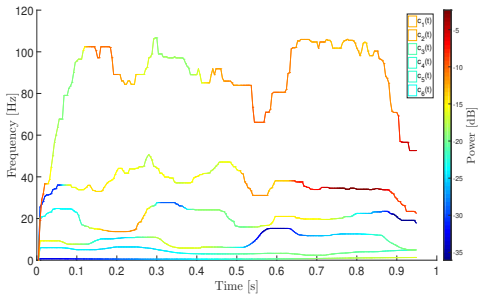


Fig. 4. Hilbert spectrum of $\boldsymbol{y}(t_k)$ for SD-1 using standard EMD

An example of the retained energy and entropy for each IMF using SD-1 is presented in Fig. 6a. In this example, the threshold is $\tau_e = 1930.9$ and then the EEG is reconstructed by using the IMF_1 and IMF_2. Relative error measure is used for evaluation. For the above example the relative errors based on (8) are as follows: $e_s = 1.3284$ and $e_r = 1.2942$, showing that the best result is obtained for the brain mapping computed from the reconstructed neural activity using

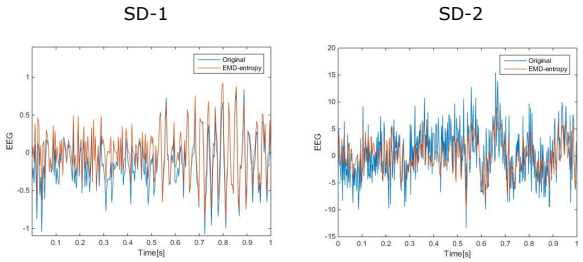


Fig. 5. Comparison of simulated $\boldsymbol{y}(t_k)$ and optimally reconstructed $\tilde{\boldsymbol{y}}(t_k)$ signals for SD-1 and SD-2 by using standard EMD for one channel

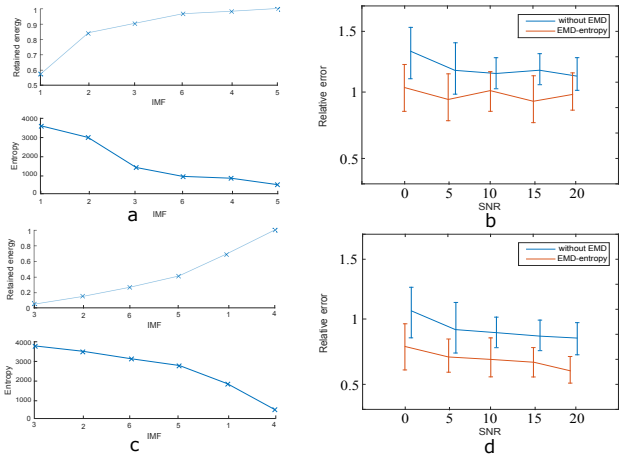


Fig. 6. a) Retained energy and entropy of $\boldsymbol{y}(t_k)$ for SD-1 using standard EMD. b)Relative error comparison of SD-1 under several noise conditions. c) Retained energy and entropy of $\boldsymbol{y}(t_k)$ for SD-2 using standard EMD. d)Relative error comparison of SD-2 under several noise conditions.

entropy-based selection of IMFs. An analysis based on 30 trials for each noise condition is shown in Fig. 6b. The retained energy and entropy for each IMF using SD-2 are presented in Fig. 6c. In this example, the threshold is $\tau_e = -1.4239 \times 10^6$ and then the EEG $\tilde{\boldsymbol{y}}(t_k)$ is reconstructed by using the IMF_3, IMF_2, IMF_6 and IMF_5. Relative error measure is used for evaluation. For the above example, the relative errors are as follows: $e_s = 1.0839$ and $e_r = 0.8994$. A comparison in terms of the relative error for 30 trials of the SD-2 is shown in Fig. 6d.

From the above, it can be seen an improvement of the source localization in terms of the relative error. That allows an improvement of epilepsy treatment when a smaller part of the brain needs to be removed.

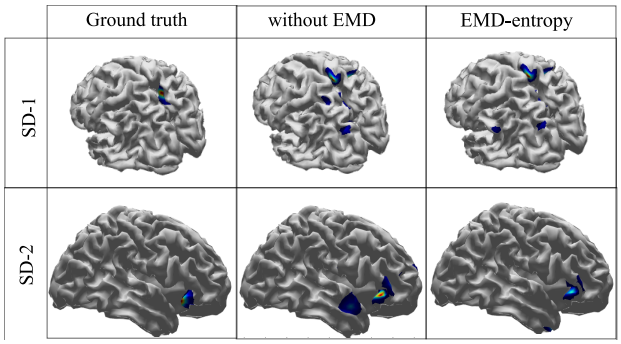


Fig. 7. Comparison of brain mapping obtained for simulated $\boldsymbol{x}(t_k)$, estimated without EMD $\widehat{\boldsymbol{x}}(t_k)$ and using the entropy-based automatic selection of IMFs $\tilde{\boldsymbol{x}}(t_k)$ neural activity for SD-1 and SD-2

5 Discussion

In this section we highlight some aspects that allow us to show the usefulness of the methodology proposed. The decomposition with IMFs allows to reconstruct the neuronal activity using only the information that is considered relevant for this application. According to some previous works [4], the EMD does not have the good performance in decomposing and reconstructing the signals with low frequency because of the problem of mode mixing. In the Figure 3 is possible to observe this phenomena. Some methodologies such as the masking signal [16] or Ensamble Empirical Mode Decomposition (EEMD) have been proposed to avoid this problem [17]. The mode mixing does not disappear completely, however the technique is very interesting when it is compared with other strategies quite common for this type of application, for example with Discrete Wavelet Transform (DWT) is necessary to consider four factors affecting the performance in epileptic focus localization: the mother wavelet, the level of decomposition, frequency bands, and features [18], [19]. Based on the above, we have proposed a new and simple methodology based on an entropy function that allows to select the IMFs regardless of the mode mixing problem. The threshold value proposed in (7) was obtained after several tests with the values of entropy and retained energy in each IMF. The first validation using simulated databases allowed us to calculate the relative error and to affirm that the technique presented provides an accurate detection of sources associated to epileptic seizures.

In Figure 7 is possible to observe the desired mapping (ground-truth) and the values obtained without EMD and with our method (EMD-entropy) for SD-1 and SD-2. In both cases, the relative error is lower with our method. In the first case, the EEG is reconstructed by using the IMF_1 and IMF_2 and in the second case we used the IMF_3, IMF_2, IMF_6 and IMF_5. The IMFs are selected automatically, and depending of the EEG the number of IMFs could change, but in either case the sources are located exactly. In both cases, the epileptic seizure appears at time $t_k = 0.5s$ and although time localization was

not one of the purposes of this paper, it can be observed in the Hilbert spectrum (Fig. 4) that the instantaneous frequencies associated with each IMF have a change in their behavior at exactly this time, therefore an additional analysis of the instantaneous frequency could be performed in order to automatically detect the beginning of an epileptic activity.

6 Conclusions

In this work, we presented an automatic detection of sources associated to epileptic seizures based on EMD and an entropy function. This strategy avoid to use the common methods that include expert clinicians with visual observation of electroencephalography (EEG) signals, which tends to be time consuming and sensitive to bias. The tests carried out with the simulated databases and the calculation of the relative error measure show an excellent performance of the proposed methodology. The new method is an alternative to solve the problem of mode mixing that usually deals with Ensamble-EMD (EEMD) or masking signal.

Acknowledgment

This work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme and also under the funding of the Departamento Administrativo Nacional de Ciencia, Tecnología e Innovación (Colciencias). Research project: 111077757982 Sistema de identificación de fuentes epileptogénicas basado en medidas de conectividad funcional usando registros electroencefalográficos e imágenes de resonancia magnética en pacientes con epilepsia refractaria: apoyo a la cirugía resectiva.
This work is also part of the research project "Solución del problema inverso dinámico considerando restricciones espacio-temporales no homogéneas aplicado a la reconstrucción de la actividad cerebral" funded by the Universidad Tecnológica de Pereira under the code E6-17-2.

References

1. Subha, D.P., Joseph, P.K., Acharya U, R., Lim, C.M.: EEG signal analysis: A survey. Journal of Medical Systems **34**(2) (April 2010) 195–212
2. Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transactions on Affective Computing **7**(1) (January 2016) 17–28
3. Lin, K.Y., Chen, D.Y., Tsai, W.J.: Face-based heart rate signal decomposition and evaluation using multiple linear regression. IEEE Sensors Journal **16**(5) (March 2016) 1351–1360
4. Bueno-Lopez, M., Giraldo, E., Molinas, M.: Analysis of neural activity from EEG data based on EMD frequency bands. In: 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS). Volume 1., Batumi, Georgia, IEEE (December 2017) 1–5

5. Zhang, T., Xu, P., Guo, L., Chen, R., Zhang, R., He, H., Xie, Q., Liu, T., Luo, C., Yao, D.: Multivariate empirical mode decomposition based sub-frequency bands analysis of the default mode network: a resting-state fmri data study. Applied Informatics **2**(1) (January 2015) 2
6. Giraldo-Suarez, E., Martinez-Vargas, J., Castellanos-Dominguez, G.: Reconstruction of neural activity from eeg data using dynamic spatiotemporal constraints. International Journal of Neural Systems **26**(07) (2016) 1–15
7. Plummer, C., Harvey, A.S., Cook, M.: EEG source localization in focal epilepsy: Where are we now? Epilepsia **49**(2) (2008) 201–218
8. Xiang, J., Li, C., Li, H., Cao, R., Wang, B., Han, X., Chen, J.: The detection of epileptic seizure signals based on fuzzy entropy. Journal of Neuroscience Methods **243**(Supplement C) (2015) 18 – 25
9. Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., Huang, C.: Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis. Entropy **19**(6) (2017) 3–17
10. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **454**(1971) (1998) 903–995
11. Xu, D., Xia, Y., Mandic, D.P.: Optimization in quaternion dynamic systems: Gradient, hessian, and learning algorithms. IEEE Transactions on Neural Networks and Learning Systems **27**(2) (February 2016) 249–261
12. Hou, T.Y., Shi, Z.: Data-driven time-frequency analysis. Applied and Computational Harmonic Analysis **35**(2) (September 2013) 284–308
13. Boashash, B.: Estimating and interpreting the instantaneous frequency of a signal. Part 1: Fundamentals. Proceedings of the IEEE **80**(4) (April 1992) 520–538
14. Grech, R., Cassar, T., Muscat, J., Camilleri, K.P., Fabri, S.G., Zervakis, M., Xanthopoulos, P., Sakkalis, V., Vanrumste, B.: Review on solving the inverse problem in EEG source analysis. Journal of NeuroEngineering and Rehabilitation **5**(1) (Nov 2008) 25
15. Munoz, P., Giraldo, E.: Time-course reconstruction of neural activity for multiples simultaneous source. In: IFMBE Proceedings CLAIB 2016. Volume 60., Bucaramanga, Colombia, Springer (October 2016) iv/485–iv/488
16. Deering, R., Kaiser, J.F.: The use of a masking signal to improve empirical mode decomposition. In: Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Volume 4. (March 2005) iv/485–iv/488
17. Wu, Z., Huang, N.E.: Ensemble Empirical Mode Decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis **1**(1) (2009) 1–41
18. Chen, D., Wan, S., Bao, F.S.: Epileptic focus localization using discrete wavelet transform based on interictal intracranial EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(5) (May 2017) 413–425
19. Bueno-Lopez, M., Molinas, M., Kulia, G.: Understanding instantaneous frequency detection: A discussion of Hilbert-Huang Transform versus Wavelet Transform. In: International Work-Conference on Time Series Analysis-ITISE. Volume 1., Granada, Spain, University of Granada (September 2017) 474–486

Performance Assessment of A short–Term Travel Forecasting Scheme for Multi-Lane Highway

Jamal Raiyn
Computer Science Department
Al Qasemi Academic College, Israel
raiyn@qsm.ac.il

Abstract

Traffic congestions and road accidents continues to increase in industry countries. There are three basic strategies to relieve congestion. The first strategy is to increase the transportation infrastructure. However, this strategy is very expensive and can only be accomplished in the long term. The second strategy is to limit the traffic demand or make traveling more expensive that will be strongly opposed by travelers. The third strategy is to focus on efficient and intelligent utilization of the existing transportation infrastructures. This strategy is gaining more and more attention because its well. Currently, the Intelligent Transportation System (ITS) is the most promising approach to implement the third strategy. Various forecast schemes have been proposed to manage the traffic data. Many studies showed that the moving average schemes are offer meaningful results compared to different forecast schemes. The performance of this proposed model has been further analyzed and compared with the performances of other methods. This paper considered the moving average schemes, namely, simple moving average, weighted moving average, and exponential moving average. Furthermore, the performance analysis of the short-term forecast schemes will be discussed. Moreover the real-time forecast model will consider the abnormal condition detection.

Keywords: Forecast scheme, EMA, statistical error measurements

1. INTRODUCTION

This paper introduces a modern forecast strategy. Conceptually, traffic information [6-7] may fall into one of the three categories as follows; historical information, real-time information, and predictive information. The historical data [10] is a collection of past observations of the system. Real-time information is most up-to-date and can be calculated, e.g., by on-line simulations. The real-time information achieved to update the historical adaptive information, special in the case that the real-time information does not matching the historical information. To optimize the forecast algorithm we have collected travel data by the mobile phone. For a successful forecast of traffic flow, it ought to apperceive the variety of environment and can adjust the parameters automatically. Furthermore it is important that the forecast model takes into consideration the abnormal conditions that occurred in real-time [9]. The paper is organized as fellow. Section 2 describes the problem in transportation engineering. Section 3 introduces the information collection based on cellular phone services. Section 4 introduces the short-term forecast scheme based on historical and real-time information. Section 5 discusses the performance analysis of the proposed short-term forecast scheme based on exponential moving average.

2. METHODOLOGY

The purpose of this paper is to assess the strengths and limitations of available of the traffic data collection based on the cellular mobile services and their corresponding processing algorithms. The performance of an incident detection system is determined on two levels: data collection technologies and data processing algorithms. Variations in cellular mobile services [32] and algorithm schemes result in a variety of solutions for incident detection. Various short-term traffic forecasting scheme have been proposed [18-27]. In this section we introduce the forecast model based on the moving average. There are three types of moving average, that is, simple moving average (SMA), weight moving

adfa, p. 2, 2011.

average (WMA), and exponential moving average (EMA). In this study, an exponential moving average is considered. An exponential moving average uses a weighting or a smoothing factor which decreases exponentially. The weighting for each older data point decreases exponentially, giving much more importance to recent observations while not discarding the older observations entirely. Figure 1 illustrated the proposed forecast model. The forecast model is divided into two phases, namely, detection phase, and forecast phase. The detection phase focused on the collected data analysis. To increase the accuracy of the forecast model we need to detect the abnormal events in the collected data. The forecast scheme is based on the exponential moving average. The robustness and accuracy of the exponential smoothing forecast is high and impressive. The accuracy of the exponential smoothing technique depends on the weight smoothed factor alpha value of the current demand. To determine the optimal alpha factor value we use the fitting curve.

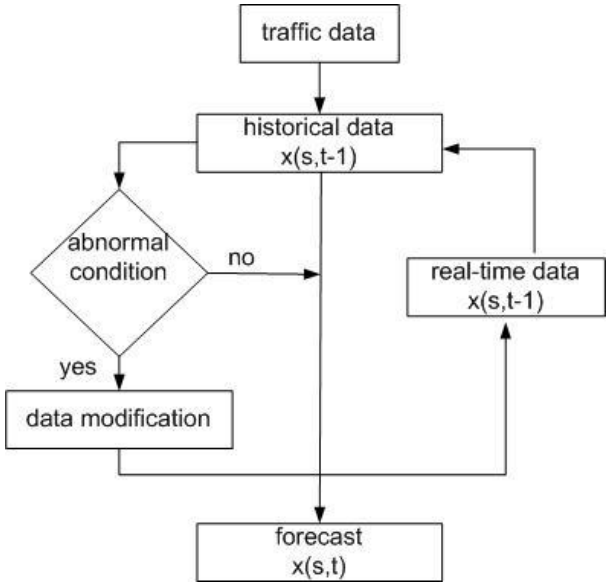


Figure 1: Algorithm Process

There are two kinds of exponential moving average forecasting (EMA) that is exponential moving average based historical information (EMA-H) and exponential moving average based real-time information

(EMA-R). The EMA-R consists of two main phases, namely detection phase and forecast phase.

4.1 Forecast based Historic Observations

The historical database is a collection of past travel observations of the system. Exponential smoothing is forecasting method that gives weight to the observed time series unequally. The unequal weight is accomplished by using one or more smoothing parameters, which determine how much weight is given to each observation. The major advantage of exponential smoothing methods is that gives good forecasts in a wide variety of applications. In addition, data storage and computing requirements are minimal, which makes exponential smoothing suitable for real-time application.

$$tt(t+1,k) = \alpha * tt^{M}(t,k) + (1-\alpha) * tt^{H}(t,k) \quad (3)$$

Where $0 < \alpha \leq 1$, $tt^{M}(t, k)$ the actual travel time in section *k* at the time *t*. $tt^{H}(t, k)$ the historical travel time in section *k* at time *t*.

4.1.1 Smoothed parameter alpha

To achieve short-term traffic flow forecasting with high accuracy, the proposed forecast scheme required to optimize the smoothed parameter alpha. Alpha determines how responsive a forecast is to sudden jumps and drops. It is the percentage weight given to the prior, and the remainder is distributed to the other historical periods. Alpha is used in all exponential smoothing methods. The lower the value of alpha, the less responsive the forecast is to sudden change. The smoothing parameter "alpha" lies between 0 and 1. To determine the optimized smoothing factor, a sum of the square errors between the observed and the forecasted alpha dose rates was analyzed by increasing the smoothing filter factor from 0.1. Sum of the square errors is decreased as the smoothing filter factor is increased as showed in Figure 2.

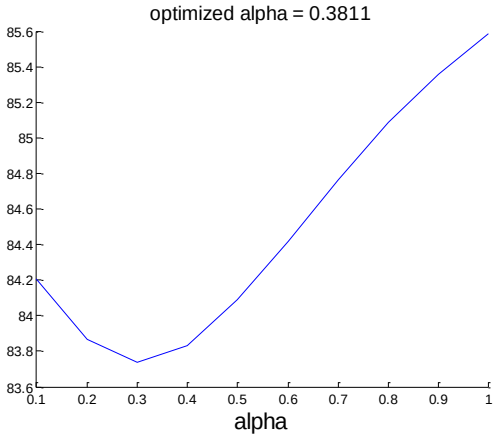


Figure 2: Smoothed parameter alpha

4.2 Real-Time Forecasting

Occurrence of Abnormal conditions in flow travel information decrease the accuracy of the forecasting based historical information and may increase the complexity of the forecasting of unusual incidence. The forecast model in real-time gives a small weight to the history information and a big weight to the real-time observation.

$$tt(t+1,k) = tt^{H}(t+1,k) + \gamma * (tt^{M}(t,k) - tt^{H}(t,k)) \qquad (4)$$

Where $0 < \gamma < 1$.

Figure 4 illustrates the real-time forecast model in abnormal conditions.

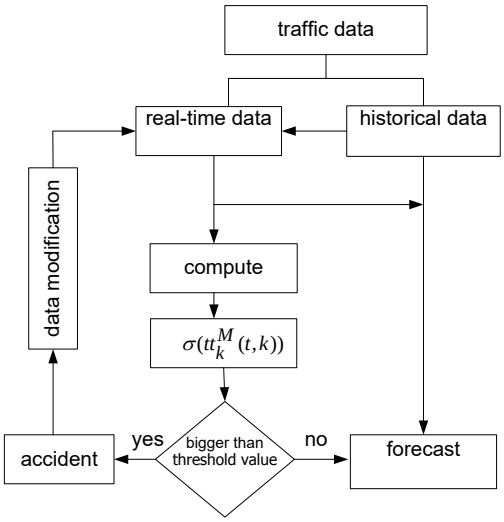


Figure 3: Real-time forecast

4.2.1 Smoothed parameter -gamma

Figure 4 illustrated that the value of gamma for real-time forecasting is closed to 0.9885.

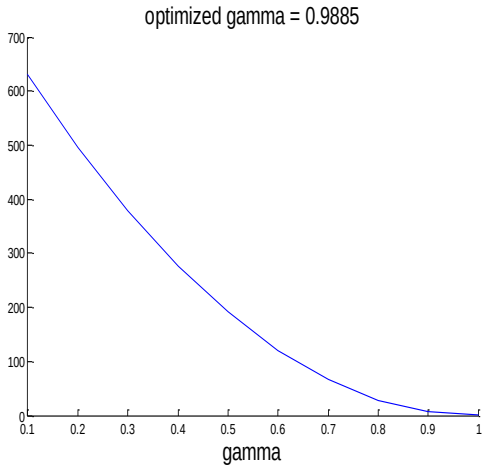


Figure 4: Smoothed parameter gamma

4.2.2 Section Mutual Influence

In the real-time forecasting we take into consideration the effect of the upstream(UP) and downstream(DS).

$$tt(t+1,k) = tt^H(t+1,k) + \gamma_1 * desired + \gamma_2 * UP + \gamma_3 * DS \quad (5)$$

Where

$$desired = \left| tt^{M}(t,k) - tt^{H}(t,k) \right|$$

$$upstearm = \left| tt^{M}(t,k-1) - tt^{H}(t,k-1) \right|$$

$$downstream = \left| tt^{M}(t,k+1) - tt^{H}(t,k+1) \right|$$

k is the desired section, *(k-1)* is the upstream section, *(k+1)* is the downstream section.

3. Accident detection strategy

The performance of an incident detection system is determined on two levels: data collection and data processing. Data collection refers to the detection/sense/surveillance technologies that are used to obtain traffic flow data. Data processing refers to the algorithms used for detecting and classifying incidents through analyzing the traffic parameters from detectors or sensors for the purpose of alerting observers of the occurrence, severity, and location of an incident. The hybrid of data collection strategies and data processing methodologies results in a variety of solutions for incident detection. The main task of the proposed accident detection (AD) algorithm is to identify and distinguish different traffic modes. It depends on an upstream occupation increase and a downstream occupation decrease at the level of loop detector where an incident happened. This algorithm compares a value of a traffic flow parameter with a known value. The algorithm trusts that an upstream occupation will increase and downstream occupation will decrease where an incident happened. In traffic incident detection, a time sequence is used to describe a traffic state. When a current measured value is deviated from the output of the algorithm seriously, the algorithm will think that an incident has occurred. The time sequence analytic algorithms include a moving average algorithm, an exponential smoothing algorithm.

- The accident characterized by temporal variation of speed at fixed road section (location) that expressed as the coefficient of variation in speed.

- The spatial variation of speed along road sections expressed as the difference in speed between upstream and downstream location (Q).

$$Q = \left| \tilde{tt}(t, s1) - \tilde{tt}(t, s2) \right| \quad (6)$$

Where $\tilde{tt}(t, s1)$, $\tilde{tt}(t, s2)$ average speeds computed over period of *t* upstream and downstream of a road sections, respectively (km/h).

5.1 Incident-influence traffic data

An incident occurring on section *i* within time interval *t* is considered to have a significant impact on traffic when traffic measurements from the upstream and downstream stations satisfy the following conditions:

i. the difference between upstream speed *si,t* and downstream speed *si+1,t* is greater than the threshold value;
ii. the ratio of the difference between the upstream and downstream speeds to the upstream speed (*si ,t - si+1,t*)/*si,t* , is greater than the threshold value; and
iii. the ratio of the difference between the upstream and downstream speeds to the downstream speed (*si,t - si+1,t*)/ *si+1,t* is greater than the threshold value.

The abnormal record shows that at least 30 km/h lower traffic speed than the average speed of all records at the same time on the same day of the week. The threshold of 30km/h is a symbolic value of the smallest speed change that people would consider "abnormal".The vehicle speed starts to decrease in upstream however the speed in downstream starts to increase.

$$\frac{tt(k,t) - tt(k+1,t)}{tt(k,t)} > threshold \quad (7)$$

$$\frac{tt(k,t) - tt(k+1,t)}{tt(k+1,t)} > threshold \quad (8)$$

5.2 Real-Time Accident Detection

The travel time forecast model considers the incident and non-incident conditions. We make different between

- Accident during peak time (morning/afternoon)
- Accident during regular time
- Heavy accident
- Light accident

The accident is cleared at current time *t* in section s, the duration is known and the speed is considered to be 30 km reduced of the average speed.

$$tt(t+1,k) = tt^{H}(t+1,k) + \gamma * (P_t) * (tt_t^{M} - tt_t^{H}) \qquad (9)$$

$$P_t = P(accident)_t = \frac{1}{1+e^{-\upsilon_t}},\ \upsilon_t = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$x_1 = \frac{\left(\sigma_t - \sigma_t^H\right)}{\sigma_t^H}, x_2 = \frac{\left(tt_t - tt_t^H\right)}{\sigma_t^H},\ x_3 = \frac{\left(tt_t - tt_t^H\right)}{\sigma_t^H} - \frac{\left(tt_{t-1} - tt_{t-1}^H\right)}{\sigma_{t-1}^H},$$

$$x_4 = \frac{\left(\sigma_t - \sigma_t^H\right)}{\sigma_t^H} - \frac{\left(\sigma_{t-1} - \sigma_{t-1}^H\right)}{\sigma_{t-1}^H}$$

Where X denotes the vector of predictor variables. β is the vector of coefficient associated with the predictor variables. and can be computed according to the binary logit model. νt is the logit link function (which is a linear combination of the predictor variables).

5.3 Smoothed Parameter Optimization

To increase the exponential moving average forecast accuracy in real-time, the smoothed parameter alpha and gamma in equation 4 should be optimized. Figure 5 illustrated the value of the optimized smoothed parameter gamma in real-time accident conditions.

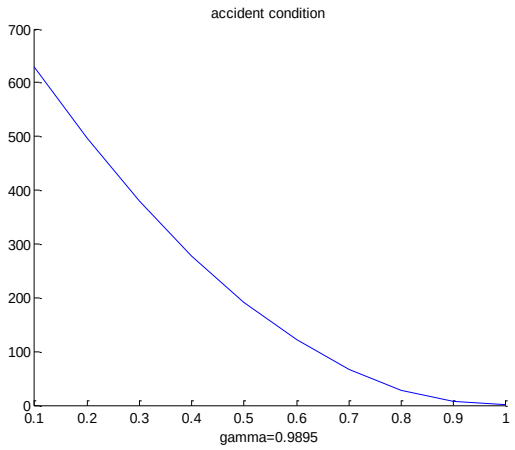


Figure 5: gamma value AD

Figure 6 illustrated value of the optimized smoothed **parameter** gamma in real-time non-accident conditions.

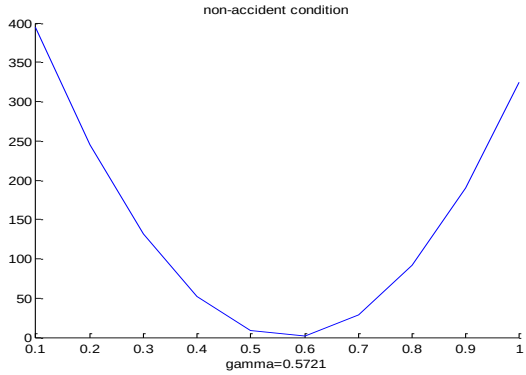


Figure 6: Gamma value in NAD

Based on equation 9, the optimized parameters in real-time accident conditions and real-time non-accident conditions are summarized in Table 1.

	AD	no AD
γ	0.9993	0.5346
β_1	1.0004	0.2215

β_2	1.0018	0.1138
β_3	0.9998	0.2315
β_4	0.9993	0.4643

Table 1: optimized parameters in AD/NAD

4. PERFORMANCE ANALYSIS

There are various measures of forecasting accuracy techniques proposed in the literature [29-36]. The aim of this study is to evaluate forecast accuracy travel observations. The forecasting accuracy techniques are used to be able to select the most accurate forecast scheme. Furthermore we aim to analyze the moving average schemes, namely simple moving average, weighted moving average, and exponential moving average. The forecasting performance of the various models and the measures of the predictive effectiveness was evaluated using various summary statistics. The comparing experiments are carried out under normal traffic condition and abnormal traffic condition to evaluate the performance of four main branches of forecasting models on direct travel time data obtained by license plate matching (LPM). The mean absolute error (MAE) is a measure of overall accuracy that gives an indication of the degree of spread, where all errors are assigned equal weights. The MSE is also a measure of overall accuracy that gives an indication of the degree of spread, but here large errors are given additional weight. It is the most common measure of forecasting accuracy. Often the square root of the MSE, RMSE, is considered, since the seriousness of the forecast error is then denoted in the same dimensions as the actual and forecast values themselves. Mean square percentage error (MSPE) is the relative measure that corresponds to the MSE. The more commonly used measure is the root mean square percentage error (RMSPE). Theil's Coefficient is another statistical measure of forecast accuracy. One specification of theil's compares the accuracy of a forecast model to that of a naïve model. A theil's U greater than 1.0 indicates that the forecast model is worse than the naïve model; a values

less than 1.0 indicates that it is better. The closer U is to 0, the better the model.

6.1 Modern vs. traditional traffic data

In this section we illustrate the simulation results and analysis of the implementation of the measured traffic speeds and travel time. The information of the dual magnet loop detectors will be compared to the information that is provided from cellular phone service. Based on the WEKA platform we have carry out analysis and comparison of different Prediction schemes. WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules and visualization [13]. We have used the WEKA to make comparison between the following schemes:

i. Smoothed Linear Models (LM)
ii. Tree Decision (TD)
iii. Nearest- Neighbor Classifier (NN)

The comparison is focused on various statistical measurements error, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), root relative squared error (RRSE), and Theil's coefficient. Table 2-4 illustrates general comparison between cellular travel speed and sensor travel speed. The results of the quality measurements are summarized in Table 2-4. Furthermore Table 2-4 illustrates that the Nearest Neighbor Scheme offers a clear and the best results compared to the linear model and tree decision schemes. Table 5 illustrates a comparison between the SMA, WMA, EMA.

Statistical Measur.	Linear model-LM	
	Cellular	Sensor
MAE	7.973	7.1967
RMSE	11.6976	11.5308
RAE	41.4674%	69.6342%

RRSE	49.1034%	68.9102%

Table 2: Cellular vs. sensor based on LM

Statistical Measur.	Tree decision-TD	
	Cellular	Sensor
MAE	9.5974	7.57
RMSE	14.0847	11.6718
RAE	49.916%	73.2468%
RRSE	59.1237%	69.753%

Table 3: Cellular vs. sensor based on TD

Statistical Measur.	Nearest Neighbour	
	Cellular	Sensor
MAE	6.4734	6.6224
RMSE	10.1594	11.0445
RAE	33.6678%	64.0777%
RRSE	42.6465%	66.0042%

Table 4: Cellular vs. sensor based on NN

statistical Measurements	SMA	WMA	EMA
MAE	6.22	8.11	5.17
RMSE	12.33	14.04	9.57
RAE	11.84	16.54	11.54
Theil's Coefficient	7.21	9.55	5.61

Table 5: SMA vs. WMA vs. EMA

4.2 Simulation Results

Results indicate that all three moving average methods, SMA, WMA and EMA, have more or less similar performance in forecasting short-

term travel times. However, as one would expect the method using optimized weights produced slightly better forecasts at a higher computational cost. Quality of forecasts is diminished as the time for which forecasts are made is farther in the future. Moving average methods overestimate travel speeds in slow-downs and underestimate them when the congestion is clearing up and speeds are increasing. Figure 7-9 described the comparison between SMA, WMA and EMA based on the various statistical measurements error. Figure 10 compared the EMA to optimized EMA based on historical observations. Figure 11 and Figure 12 showed the actual observations compared to EMA based Information and to EMA based on real-time information. Results indicate that all three moving average methods have more or less similar performance in forecasting short-term travel times. However, as one would expect the method using optimized weights produced slightly better forecasts at a higher computational cost. Quality of forecast is diminished as the time for which forecasts are made is farther in the future. Moving average methods overestimate travel speeds in slow-downs and underestimate them when the congestion is clearing up and speeds are increasing. Figure 13 illustrated the comparison between the exponential moving average based historical information (EMA-H) and the exponential moving average based real-time information (EMA-R) compared to actual observations. EMA-H detects the abnormal conditions in travel flow traffic based pervious information that are collected in same location and at the same time. The advantage of the EMA-H is an identification of incident in flow traffic. However a repeated incident with the same characteristics in the future is not certain. Furthermore Figure 13 illustrated that EMA-R identify the incident in the flow traffic and provides incident clearness. Table 6 and Table 7 illustrated the comparison between EMA based historical information and EMA based real- time in accident and in non accident conditions.

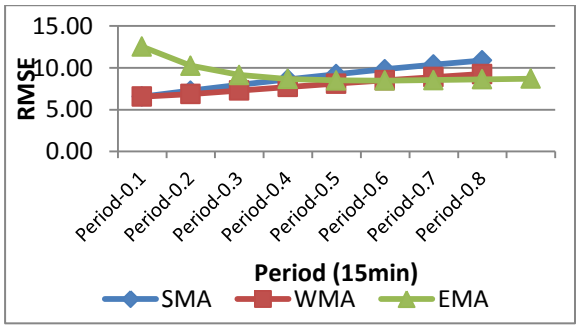


Figure 7: SMA,WMA, EMA in Comparison

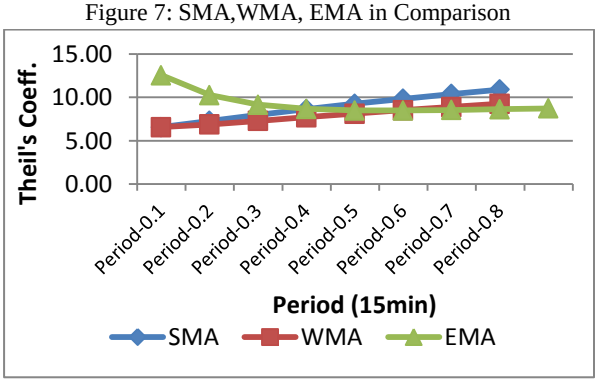


Figure 8: Theil's Coefficient

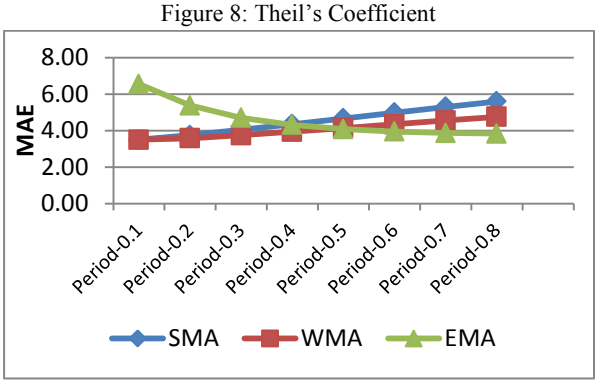


Figure 9: MAE

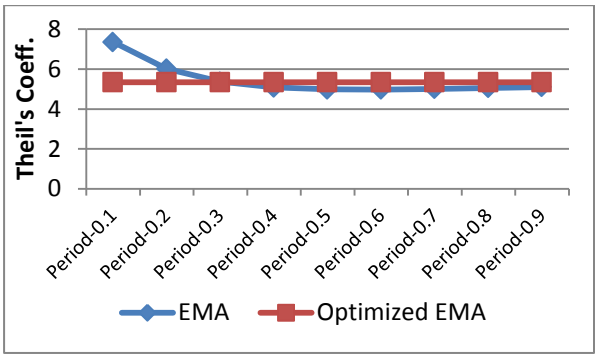


Figure 10: Comparison between EMA and Opt-EMA

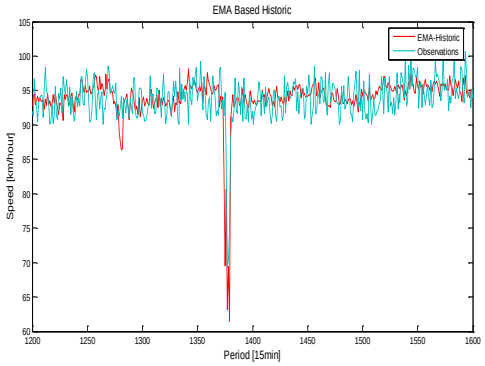


Figure 11: EMA-H vs. Actual Observations

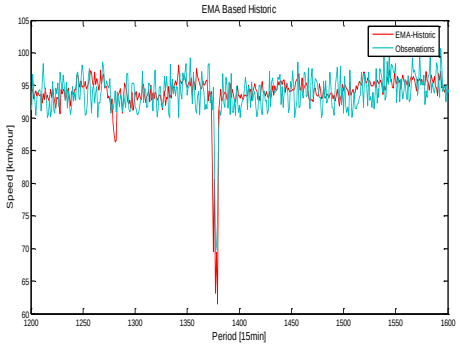


Figure 12: EMA-R vs. Actual Observations

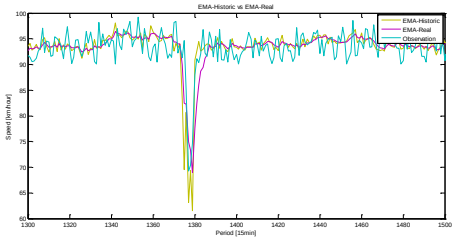


Figure 13: EMA-H, EMA-R in comparison

	Hist	Real
mean data	79.234	79.234
mean prediction	75.324	78.981
std data	17.737	17.737
std prediction	22.993	16.673
Observations with error over 5 km/hr	42.206	40.281
Observations with error over 10 km/hr	26.006	22.356
max abs. error	93.492	81.288
max. relative error	1181.7	4538.8
mean error	3.9104	0.25324
mean abs. error	10.588	7.4191
mean relative error	16.743	12.656
root mean squared error	20.505	12.118
root mean squared per-cent error (1)	39.798	32.049
root mean squared per-cent error (2)	25.88	15.294
Theil's coefficient	12.82	7.4844
bias proportion	3.6367	0.04367
variance proportion	6.57	0.77046
co-variance proportion	89.793	99.186

Table 6: Historical vs. Real-Time in AD

	Hist	Real
mean data	67.805	67.805
mean prediction	65.622	66.798
std data	17.809	17.809
std prediction	18.682	16.968
Observations with error	33.086	31.293

over 5 km/hr		
Observations with error over 10 km/hr	17.385	15.735
max abs. error	73.39	73.264
max. relative error	587.12	586.11
mean error	2.183	1.0076
mean abs. error	6.6768	5.472
mean relative error	12.238	10.562
root mean squared error	12.452	9.2418
root mean squared percent error (1)	26.42	23.514
root mean squared percent error (2)	18.364	13.63
Theil's coefficient	9.0011	6.6476
bias proportion	3.0737	1.1886
variance proportion	0.49122	0.82716
co-variance proportion	96.435	97.984

Table 7: Historical vs. Real-Time in NAD

CONCLUSION

Various forecast schemes have been proposed to manage the travel flow. In order to select the fit forecast scheme we have carried out analysis and comparison between different forecast schemes. In this paper we have introduced various forecast schemes based on the historical data and real-time observations. Furthermore in this paper we discuss and summarize some prediction methods based on their performance analysis. We conclude that the optimized exponential moving average is the most accurate method. Moreover the proposed algorithm has been given best solution for traffic travel forecast. However the road accidents increased rapidly. To reduce the incidents should be developed a new detection scheme that takes driver's behaviors into consideration.

ACKNOWLEDGMENTS

The author wishes to thank the Center for Innovation in Transportation at the University of Tel Aviv for funding this research.

Reference

[1] Ahmed, M.S. and Cook, A. R. (1997), "Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques", *Transportation Research Record*, 722, pp. 1-9.

[2] Alger, M. (2004), "Real-time traffic monitoring using mobile phone data", Proceedings on 49th European Study.

[3] Andrada-Felix, J. and Fernandez-Rodriguez, F. (2008) "Improving Moving Average Trading Rules with Boosting and Statistical Learning Methods", *Journal of Forecasting*, 27, pp.433-449.

[4] Andrawis, R. R. and Atiya, F.A. (2009), "A New Bayesian Formulation for Holt's Exponential Smoothing", *Journal of Forecasting*, 28, pp.218-234.

[5] Raiyn, J. (2014). Development of Resource Allocation Strategies Based on Cognitive Radio, LAMBERT Academic Publishing, Germany.

[6] Bickel, J. P., Chen, C., Kwon, J., Rice J., Zwet, V. E. and Varaiya, P. (2007) "Measuring Traffic", *Statistical Science*, Vol. 22, No. 4, pp. 581-597.

[7] Borzacchielo, T.M. (2010), "the use of data from mobile phone networks for transportation applications", *TRB Annual Meeting*.

[8] Brag A.W. and Chou W. (1991), "Traffic Analysis for High Networks", *GLOBECOM*.

[9] Chrobok, R., Kaumann, O., Wahle, J. and Schreckenberg, M. (2004), "Different methods of traffic forecast based on real data", *European Journal of Operational Research* 15, pp. 558-568.

[10] Chrobok, R., Kaumann, O., Wahle, J. and Scheckenberg, M. (2000), "Three Categories of Traffic Data: Historical, Current, and Predictive", *9th IFAC Symposium*, Braunschweig, Germany, 13-15, Vol. 1, June. pp. 221-226.

[11] Elvik, R.(2000), "How much do road accidents cost the national economy?", *Accident Analysis and Prevention* 32, pp.849-851.

[12] Guin, A.(2006), "Travel Time Prediction using a Seasonal Autoregressive Integrated Moving Average Time Series Model", *Proceedings of the IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, September 17-20, pp. 493-498.

[13] Han, J. and Kamber, M. (2006), "Data Mining: Concepts and techniques", *2nd edn. Morgan Kaufmann*, San Francisco.

[14] Högberg P.(1975), "Estimation of Parameters in Models for Traffic Prediction: A Non-Linear Regression Approach", *Transportation Research*, Vol. 10, 1975. pp. 263-265.

[15] Jo, H., Lee B., Na, Y-C, Lee H. and Oh, B. (2007), "Prioritized Traffic Information Delivery Based on Historical Data Analysis", *Proceedings of the 2007 IEEE Intelligence Transportation Systems Conference*, Seattle, WA, USA, Sept. 30-Oct.3.

[16] Karim, A. and Adeli, H. (2003), "Fast Automatic Incident Detection on Urban and Rural Freeways Using Wavelet Energy Algorithm", Journal of Transportation Engineering, ASCE, Vol. 129, No.1, pp.57-68.

[17] Lee, H., Chowdhury, K, N. and Chang J. (2008), "A new Travel Time Prediction Method for Intelligent Transportation Systems", *Springer-Verlag Berlin*. pp. 473-483.

[18] Lv, Y. and Tang, S. (2010), "Real-time Highway Traffic Accident Prediction Based on the *k*-Nearest Neighbor Method", *TRB Annual Meeting*.

[19] Lee,Y. and Wei, C.H. (2010), "A Computerized Feature Selection Using Genetic Algorithms to Forecast Freeway Accident Duration Times," *Computer-Aided Civil and Infrastructure Engineering*, 25:2, 132-148.

[20] Martínez-Zarzuela, M. (2010), "Wavelet-Based Denoising for Traffic Volume Time Series Forecasting with Self-Organizing Neural Networks," *Computer-Aided Civil and Infrastructure Engineering*, 25:7.

[21] Nicholson, H. and Swann, C.D.(1974), "The Prediction of Traffic Flow Volumes Based on Spectral Analysis", *Transportation Research*, Vol. 8, , pp. 533-538.

[22] Okutani, I. and Stephanedes, J. Y. (1984), "Dynamic Prediction of Traffic Volume through Kalman Filtering Theory", *Transportation Research*, Vol. 18B, No.1, pp.1-11.

[23] Ronen, B., Coman, A. and Schragenheim, E. (2011), "Peak management", *Int. J. Prod. Res.*, Vol.39, No. 14, pp. 3183-3193.

[24] Sabry M., Abd-El-Latif H., Yousef, S., and Badra, N.(2007a), "A Time-Series Forecasting of Average Daily Traffic Volume", *Australian Journal of Basic and Applied Sciences*, 1(4): 386-394.

[25] Sabry, M., Abd-El-Latif, H., Badra, N.(2007b) "Comparison Between Regression and Arima Models in Forecasting Traffic volume", *Australian Journal of Basic and Applied Sciences*, 1(2): 126-136.

[26] Stathopoulos, A., Dimitriou, L., and Tsekeris, T. (2008), "Fuzzy Modeling Approach for Combined Forecasting of Urban Traffic Flow," *Computer-Aided Civil and Infrastructure Engineering*, 23:7, pp. 521-535.

[27] Raiyn, J. (2016). Speed Adaptation in Urban Road Network Management, *Transport and Telecommunication*, *17*, 2. pp. 11-121.

[28] SooBeom, L. (2009), "Analysis of Design Elements for Urban Highway Safety", *4th IRTAD Conference* 16-17 September, pp.251-262.

[29] Tu, H., Van Lint, H. and Van Zuylen, H. (2008), "The Effects of Traffic Accidents on Travel Time Reliability", *IEEE Conference on Intelligent Transportation Systems*, Beijing, China, October 12-15.

[30] Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2008), "Temporal Evolution of Short-Term Urban Traffic Flow: A Non-Linear Dynamics Approach," *Computer-Aided Civil and Infrastructure Engineering*, 23:7 pp. 536-548.

[31] Wang, Z. and Murray-Tuite, P. (2010), "Modeling Incident-Related Traffic and Estimating Travel Time with a Cellular Automaton Model", *TRB 2010 Annual Meeting.*

[32] Raiyn, J. (2014). Real-Time Road Traffic Anomaly Detection, *Journal of Transportation Technologies*, *4*, pp. 256-266.

[33] Xia, J (2010), "Predicting Freeway Travel Time Under Incident Condition", *TRB 2010 Annual Meeting.*

[34] Xiaoqiang, Z., Ruimin, L. and Xinxin, Y. (2010), "Incident Duration Model on Urban Freeways Based on Classification and Regression Tree", *second International Conference on Intelligent Computation Technology and Automation, TRB 2010 Annual Meeting.*

[35] Xie, Y., Zhang, Y., and Ye, Z. (2007), "Short-term Traffic Volume Forecasting Using Kalman Filter with Discrete Wavelet Decomposition," Computer-Aided Civil and Infrastructure Engineering, 22:5, pp. 326-334.

[36] Zheng, X. and Liu, M. (2009), "An overview of accident forecasting methodologies", *Journal of Loss Prevention in the Process Industries* 22, pp.484-491.

[37] Raiyn, J. (2012). Handoff Self-Management Based on SNR in Mobile Communication Networks. *International Journal of Wireless and Mobile Computing*. *6*, 1, pp. 39-48.

On the limits of probabilistic prediction in nonlinear time series analysis

José M. Amigó[1], Yoshito Hirata[2], and Kazuyuki Aihara[3]

[1]Centro de Investigación Operativa, Universidad Miguel Hernández, 03202 Elche, Spain
[2]Mathematics and Informatics Center, The University of Tokyo, Tokyo 113-8656, Japan
[3]Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan

Abstract. This communication reviews the main results of Refs. [1] and [2] on the limits of prediction in nonlinear time series analysis. In particular, it describes the properties of the ignorance score in the nonlinear perfect model scenario and applications to forecasting performance evaluation.

Keywords: Probabilistic prediction, nonlinear time series analysis, perfect model scenario, ignorance score, imperfect model scenario.

1 Introduction

Probabilistic time series forecasting is instrumental in several fields of science and technology, including meteorology, seismology, climatology and, more recently, renewable energies. Therefore, it is useful, both theoretically and practically, to know its limits. To this end, we assume a *perfect model scenario*, that is to say, the system producing the data is supposed to be perfectly known, while the uncertainty results solely from the finite precision of the observations. Two further qualifications are added to this general setting to connect with nonlinear time series analysis. First, the 'system' is a discrete-time dynamical system. Second, the uncertainty is due to the quantization of the state space rather than being introduced via additive noise (as in the conventional statistical approach). In this 'nonlinear' perfect model scenario, the quality of probabilistic predictions n time steps ahead will be measured by the so-called *ignorance score*, a concept borrowed from meteorology [3] that is a conditional entropy in our nonlinear framework. It turns out that the ignorance score is bounded as a function of the prediction horizon n, the boundaries of its admissible domain depending on the entropies of the underlying dynamics and the state space quantization [1].

Other properties of the ignorance score will be derived in two further scenarios: (i) the *imperfect model scenario* [1], corresponding to the real situation in which the underlying system is imperfectly known, and (ii) the *differential perfect and imperfect model scenarios* [2], where the probability mass functions are replaced by probability densities, and the discrete entropies by differential

entropies (as usually in practice). It follows that the slope of the ignorance score in an imperfect scenario can be used to compare the performance of probabilistic forecasting algorithms, as we will show with numerical examples.

2 The ignorance score in the nonlinear perfect scenario

The **mathematical framework** is set by the so-called *nonlinear perfect model scenario* (PMS). According to this ideal scenario:

(H1) Observations are output by a discrete-time, measure-preserving dynamical system $(\Omega, \mathcal{B}, \mu, f)$, where (i) Ω is the state space, (ii) $\mathcal{B}$ is a σ-algebra of subsets of Ω, (iii) μ is a probability measure on the measurable space (Ω, B), and (iv) $f : \Omega \to \Omega$ is a μ-*invariant* map, i.e., it is measurable ($f^{-1}B \in \mathcal{B}$ for all $B \in \mathcal{B}$) and $\mu(f^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$.

(H2) Uncertainty is due to the finite precision of the observations.

In a nonlinear PMS, finite precision is modeled via finite **partitions**

$$\alpha = \{A_1, A_2, ..., A_{|\alpha|}\} \tag{1}$$

where the sets (or 'bins') $A_i \in \mathcal{B}$ $(1 \leq i \leq |\alpha|)$ are pairwise disjoint and $\Omega = A_1 \cup A_2 \cup ... \cup A_{|\alpha|}$.

A partition (or 'coarse-graining') of the state space introduces randomness in the dynamics. A formal approach to this fact goes by the name of *symbolic dynamics*, which is a stationary random process with alphabet $\{1, 2, ..., |\alpha|\}$.

Definition 1. *Given* $\alpha = \{A_1, A_2, ..., A_{|\alpha|}\}$, *let* $X_n^\alpha : \Omega \to \{1, 2, ..., |\alpha|\}$ *be the random variables*

$$X_n^\alpha(x) = i_n \;\; \textit{iff} \;\; f^n(x) = A_{i_n} \;\; (\textit{i.e., } x \in f^{-n}(A_{i_n})) \tag{2}$$

Then, $\mathbf{X}^\alpha = (X_0^\alpha, ..., X_n^\alpha, ...)$ is a *stationary* random process called the *symbolic dynamics* of f with respect to α. The probability mass functions are given by

$$p(i_0, i_1, ..., i_q) = \mu(\{A_{i_0} \cap f^{-1}(A_{i_1}) ... \cap f^{-q}(A_{i_q}\}) \tag{3}$$

Let us remind at this point the basic concepts of entropy:

Definition 2. **(i)** *Entropy of the partition* α*:*

$$H(\alpha) = -\sum_{i=1}^{|\alpha|} \mu(A_i) \log \mu(A_i). \tag{4}$$

(ii) *Entropy of* f *with respect to the partition* α*:*

$$h_\mu(f, \alpha) = -\lim_{n \to \infty} \frac{1}{n} \sum_{i_0, ..., i_{n-1}=1}^{|\alpha|} p(i_0, ..., i_{n-1}) \log p(i_0, ..., i_{n-1}). \tag{5}$$

(iii) *Kolmogorov-Sinai (KS) entropy of f:*

$$h_\mu(f) = \sup_\alpha h_\mu(f, \alpha). \tag{6}$$

If $h_\mu(f) = h_\mu(f, \alpha)$ we say that α is a *generating partition*. For easy of notation we suppose henceforth that f is *invertible*.

In a nonlinear PMS, the quality of a prediction can be measured by the **ignorance score**, which can be defined via the symbolic dynamics $\mathbf{X}^\alpha$ as follows:

$$\begin{aligned}\left\langle -\log p\left(i_n \middle| i_0^{-\infty}\right)\right\rangle &\equiv \lim_{N\to\infty} \left\langle -\log p\left(i_n \middle| i_0, i_{-1}, ..., i_{-N}\right)\right\rangle \\ &= H\left(X_n^\alpha \middle| X_0^\alpha, ..., X_{n-1}^\alpha, ...\right) \\ &= \inf_{N\in\mathbb{N}} H\left(X_n^\alpha \middle| X_0^\alpha, ..., X_N^\alpha\right)\end{aligned}$$

That is, $\left\langle -\log p\left(i_n \middle| i_0^{-\infty}\right)\right\rangle$ is the uncertainty about the outcome of X_n^α, given the outcomes of $X_0^\alpha, ..., X_{n-1}^\alpha, ...$ (0 is present time).

In [1] we proved the basic properties of the ignorance score in the nonlinear PMS.

Theorem 1. **(i)** $\left\langle -\log p(i_1 \left| i_0^{-\infty}\right.)\right\rangle = h_\mu(f, \alpha)$.
(ii) *Monotonicity:*

$$\left\langle -\log p(i_n \left| i_0^{-\infty}\right.)\right\rangle \leq \left\langle -\log p(i_{n+1} \left| i_0^{-\infty}\right.)\right\rangle. \tag{7}$$

(iii) *Upper bound:*

$$\left\langle -\log p(i_n \left| i_0^{-\infty}\right.)\right\rangle \leq H(\alpha). \tag{8}$$

The bound $H(\alpha)$ is reachable if f is strong-mixing.
(iv) *Boundedness of the slope:*

$$\left\langle -\log p(i_n \left| i_0^{-\infty}\right.)\right\rangle \leq n h_\mu(f, \alpha), \tag{9}$$

i.e.,

$$\frac{1}{n-1}\left(\left\langle -\log p(i_n \left| i_0^{-\infty}\right.)\right\rangle - \left\langle -\log p(i_1 \left| i_0^{-\infty}\right.)\right\rangle\right) \leq h_\mu(f, \alpha). \tag{10}$$

Properties (i)-(iv) are depicted in Fig. 1 as an admissible domain for the ignorance score [1]. Note that property "boundedness of the slope" (also in Theorem 6 below) refers actually to the slope of the linear interpolation between the points $(1, \left\langle -\log p(i_1 \left| i_0^{-\infty}\right.)\right\rangle)$ and $(\left\langle -\log p(i_n \left| i_0^{-\infty}\right.)\right\rangle)$.

Remark 1. **(i)** The slope of the left boundary is $h_\mu(f, \alpha)$ $(= h_\mu(f)$ if α is a generating partition).
(ii) *The upper vertex occurs at $n = H(\alpha)/h_\mu(f, \alpha) \geq 1$.*
(iii) *The boundaries are optimal. For example, if*

$$\Omega = [0, 1),\ f(x) = 2x \bmod 1, \alpha = \{[(i-1)2^{-k}, i2^{-k}) : 1 \leq i \leq 2^k\} \tag{11}$$

then $h_\mu(f, \alpha) = h_\mu(f) = 1$ bit and $H(\alpha) = k$ bit.

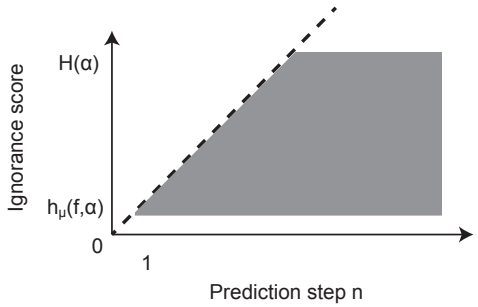


Fig. 1. Admissible domain of the ignorance score (the boundaries have been linearly interpolated for a better visualization). The lower boundary comprises the points $(n, h_\mu(f,\alpha))$ for $n \geq 1$. The left boundary comprises the points $(n, h_\mu(f,\alpha)n)$ for $1 \leq n \leq n_{\text{ramp}} := \lfloor H(\alpha)/h_\mu(f,\alpha) \rfloor$. The upper boundary comprises the points $(n, H(\alpha))$ for $n \geq n_{\text{ramp}} + 1$. Figure borrowed from Ref. 1.

In an **imperfect model scenario** (e.g., time series analysis), the probabilities

$$p\left(i_n|\, i_0^{-N}\right) \equiv p(i_n\,|i_0, ..., i_{-N})\,,\ N \geq 0 \tag{12}$$

have to be estimated.

Theorem 2. *[1] Let* $\tilde{p}\left(i_1|\, i_0^{-N}\right)$ *be estimated probabilities. Then*

$$\left\langle -\log \tilde{p}(i_n\left|i_0^{-\infty}\right)\right\rangle \geq \left\langle -\log p(i_n\left|i_0^{-\infty}\right)\right\rangle. \tag{13}$$

Therefore, the admissible domain may be violated in an imperfect model scenario. This will be exploited in Sect. 4 for gauging the performance of prediction algorithms.

3 The ignorance score in the differential perfect model scenario

Sometimes precision is high enough to approximate probability mass functions by densities. In this case we speak of a *differential perfect model scenario* [2]. Hereafter we tailor this scenario to the praxis in nonlinear time series analysis. Thus, the underlying dynamical system is $(I, \mathcal{B}, \mu, f)$, where (i) I is a bounded interval of $\mathbb{R}^d$, (ii) $\mathcal{B}$ is the Borel σ-algebra of I, (iii) μ is an *absolutely continuous* measure, $\mu(B) = \int_B \nu(x)dx$ for all $B \in \mathcal{B}$, and (iv) $f : I \to I$ is a μ-*invariant* map, i.e.,

$$\int_B \nu(y)dy = \int_{f^{-1}(B)} \nu(x)dx. \tag{14}$$

The differential PMS can be dealt with similarly to the 'discrete' one with the following formal changes:

- The general partition $\alpha = \{A_1, A_2, ..., A_{|\alpha|}\}$ is replaced by a uniform partition with $\mathrm{Vol}(A_i) = \Delta$ for all i.
- The labels of the random variables $X_n^\alpha(x)$ are now arbitrary points of the bins A_i, i.e., $X_n^\alpha(x) = x_{i_n} \in A_{i_n}$.
- The probability mass functions $p(i_0, ..., i_n)$ are replaced by the densities $\rho^\Delta(x_0, ..., x_n)$, where $\rho^\Delta(x_0, ..., x_n)$ is defined on the rectangles $A_{i_0} \times A_{i_1} \times \cdots \times A_{i_n} \in \alpha^n$ as

$$\begin{aligned}\rho^\Delta(x_0, ..., x_n)\big|_{A_{i_0}\times A_{i_1}\times\cdots\times A_{i_n}} &= \chi_{A_{i_0}\cap f^{-1}A_{i_1}\cap\ldots\cap f^{-n}A_{i_n}}(x_0)\nu(x_0)\\ &\quad\times\frac{\nu(x_1)}{\mu(A_{i_1})}\cdots\frac{\nu(x_n)}{\mu(A_{i_n})}.\end{aligned}$$

- The Shannon (or 'discrete') entropy $H(p(i_0, ..., i_n))$ is replaced by the differential entropy

$$h(\rho^\Delta(x_0, ..., x_n)) = -\int \rho^\Delta \log \rho^\Delta d^{n+1}x. \tag{15}$$

– The discrete ignorance score $\left\langle -\log p\left(i_n|\, i_0^{-\infty}\right)\right\rangle$ is replaced by the 'differential' ignorance score

$$\left\langle -\log \rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)\right\rangle = h(\rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)), \tag{16}$$

where

$$h(\rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)) = \lim_{N\to\infty} h(\rho^{\Delta}\left(x_n|\, x_0, x_{-1}, ..., x_{-N}\right)). \tag{17}$$

The basic *properties of the differential ignorance score* are summarized in the following theorem.

Theorem 3. *[2] For any $\varepsilon > 0$ and sufficiently small Δ, it holds:*

(i) *Monotonicity:*

$$\left\langle -\log \rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)\right\rangle \leq \left\langle -\log \rho^{\Delta}\left(x_{n+1}|\, x_0^{-\infty}\right)\right\rangle + \varepsilon. \tag{18}$$

(ii) *Upper boundedness:*

$$\left\langle -\log \rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)\right\rangle \leq h(\nu). \tag{19}$$

(iii) *Boundedness of the slope:*

$$\left\langle -\log \rho^{\Delta}\left(x_n|\, x_0^{-\infty}\right)\right\rangle - \left\langle -\log \rho^{\Delta}\left(x_1|\, x_0^{-\infty}\right)\right\rangle \leq (n-1)h_{\mu}(f) + \varepsilon. \tag{20}$$

(iv) *Imperfect model scenario:*

$$\langle -\log \rho^{\Delta}(x_n\left|x_0^{-\infty}\right)\rangle < \langle -\log \tilde{\rho}^{\Delta}(x_n\left|x_0^{-\infty}\right)\rangle \leq h(\tilde{\rho}^{\Delta}). \tag{21}$$

Therefore, the differential ignorance score fulfills similar properties to those of the discrete ignorance score for sufficiently fine partitions (or small size Δ). There are minor though important changes. In Property (ii) $H(\alpha)$ is replaced by $h(\nu)$ (the differential entropy of the probability density of the invariant measure μ) and, in Property (iii), $h(f, \alpha)$ (the entropy of f with respect to the partition α) is replaced by the KS entropy of f.

4 Applications

As said above, the discrepancy of the theoretical properties with the actual properties of the differential ignorance score can be used as a metric for prediction quality. The handier property in this regard is perhaps the boundedness of the slope since it involves the KS entropy. We are going to illustrate these claims via numerical simulations with the following maps.

(M1) *Logistic map*:

$$f(x) = 4x(1-x), \;\; 0 \leq x \leq 1. \tag{22}$$

(M2) *Hénon map*:

$$f(x, y) = (1 - 1.4x^2 + 0.3y,\; x). \tag{23}$$

(M3) *Ikeda map*:

$$f(x,y) = (1 + 0.9(x\cos\phi - y\sin\phi),\ 0.9(x\sin\phi + y\cos\phi)), \tag{24}$$

where

$$\phi = 0.4 - \frac{6}{1+x^2+y^2} \tag{25}$$

Given the (finite) time series $(x_t)_{1\le t\le T}$, we estimate the density of the forecast for x_{t+n} (based on the 20 *nearest neighbors* $x_{\tau_i(t)}$ of x_t for $t > T/2$) by:

$$\tilde{\rho}_n^{\Delta}(z|\,x_t) = \sum_{i=1}^{20} \frac{1}{20\Delta}\Theta(0.5\Delta - |z - x_{\tau_i(t)+n}|) \tag{26}$$

$$+h\frac{\Theta(M-m-|z-0.5(M+m)|)}{2(M-m)}, \tag{27}$$

where

$$M = \max_{t\le\lfloor T/2\rfloor} y_t,\ \ m = \min_{t\le\lfloor T/2\rfloor} y_t, \tag{28}$$

and

$$\Theta(w) = \begin{cases} 1, & \text{if } w \ge 0, \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

The second term of Eq. (26) hedges the risk of a forecast probability being 0. Furthermore, set

$$\langle -\log_2 \tilde{\rho}_n^{\Delta}(z\,|x_0^{-\infty})\rangle = -\frac{1}{T-\lfloor T/2\rfloor}\sum_{t=\lfloor T/2\rfloor+1}^{T} \log_2 \tilde{\rho}_n^{\Delta}(z|\,x_t). \tag{30}$$

In our numerical simulations.

$$\Delta = 0.025,\ T = 10^5,\ \text{and}\ \ n = 1,2,...,10 \tag{31}$$

We used double precision so that $\tilde{\rho}_n^{\Delta}(z|\,x_t) \simeq \tilde{\rho}_n^{\Delta}(z|\,x_t, x_{t-1}, ..., x_1)$. The numerical results are shown in Fig. 2.

It turns out that the boundedness of the slope

$$\frac{1}{n-1}\left(\left\langle -\log_2 \rho^{\Delta}(x_n|\,x_0^{-\infty})\right\rangle - \left\langle -\log_2 \rho^{\Delta}(x_1|\,x_0^{-\infty})\right\rangle\right) \lesssim h_\mu(f) \tag{32}$$

is violated. For example, for $n = 7$:

Map	**Slope**	**KS entropy**	**Violation**
Logistic	1.4918	1	0.4918
Hénon	0.8235	0.6223	0.2012
Ikeda	1.1878	0.7450	0.4428

(33)

In view of these results, we conclude that the dynamic generated by the logistic map is the most difficult to predict with the algorithm (26), followed by the dynamic generated by the Ikeda map.

Alternatively, we can gauge the performance of different prediction algorithms when applied to the same dynamic.

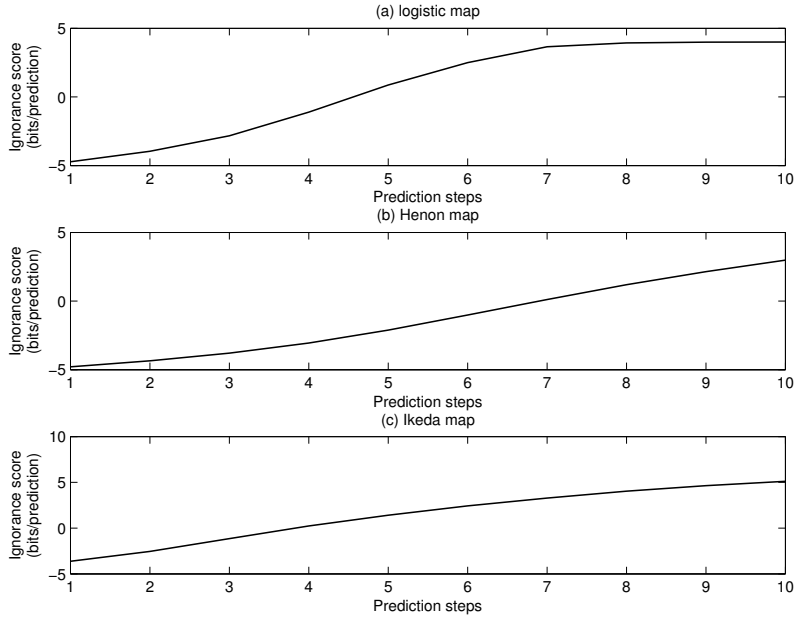


Fig. 2. The differential ignorance score $\langle -\log_2 \hat{\rho}^{\Delta}(y_n \,|y_0^-)\rangle$ for $1 \leq n \leq 10$ obtained with the prediction method of Sect. 4. The underlying dyamics is generated by the logistic map (a), the Henon map (b), and the Ikeda map (c). The points $(n, \langle -\log_2 \hat{\rho}^{\Delta}(y_n \,|y_0^-)\rangle)$ have been linearly interpolated for a better visualization. (Figure borrowed from Ref. 2).

5 Conclusion

The results presented above can be summarized in the following points.

- Prediction in nonlinear time series analysis has some fundamental limitations.
- These limitations translate into an admissible domain for the ignorance score in a perfect model scenario.
- Violations of these limitations can be used to quantify prediction quality.

References

1. J.M. Amigó, Y. Hirata, and K. Aihara, On the limits of probabilistic forecasting in nonlinear time series analysis, *Chaos* **26**, 123114 (2016).
2. J.M. Amigó, Y. Hirata, and K. Aihara, On the limits of probabilistic forecasting in nonlinear time series analysis II: Differential entropy, *Chaos* **27**, 083125 (2017).
3. H. Du and L. Smith, Pseudo-orbit data assimilation. Part I: The perfect model scenario, *J. Atmos. Sci.* **71**, 469-482 (2014).

Evaluation of forecasting approaches using hybrid multi-criteria decision-making models

Yvonne Badulescu[1,2]✉ and Naoufel Cheikhrouhou[1]✉

[1]Geneva School of Business Administration, University of Applied Sciences Western Switzerland (HES-SO), 1227 Carouge, Switzerland
[2]Faculty of Business and Economics (HEC), University of Lausanne (UNIL), Switzerland
✉Yvonne.Badulescu@hesge.ch
Yvonne.Badulescu@unil.ch
✉Naoufel.Cheikhrouhou@hesge.ch

Abstract. The demand forecast is the most influential input into enterprise activities planning thus creating a challenging issue for Demand Planning experts in model selection. Models including quantitative, qualitative and hybrid forecasting methods have been developed and are widely used. The literature reveals the use of several case-dependent error measures to evaluate forecasting accuracy, however, these performance indicators may at times differentiate in results making it more difficult in determining the most appropriate forecasting model for the users' needs. This paper presents the development of two hybrid multi-criteria decision making approaches, AHP-TOPSIS and ANP-TOPSIS, to evaluate and rank the relative performances based on error measures of alternate forecasting models. Validation is provided through an industrial application using empirical data from a plastic bag manufacturer based on five models; three regression forecast models and two hybrid demand forecast models using expert judgement. Results illustrate that subjective adjustment by experts of mathematical forecasts consistently gives a higher ranking due to proximity to the ideal solution, and that collaborative adjustment limits the risk of outliers due to forecasting errors that could be done by a single decision maker.

1 Introduction

There has been much research on the development of forecasting models to accurately determine the real future demand. Quantitative methods rely on mathematical approaches that can lead to reliable forecasts by extrapolating regular patterns in time-series. In addition to regression models, key actors in enterprises frequently obtain knowledge and insights about future non-periodic events that are expected to strongly influence the demand. Consequently, hybrid forecasting methods are employed, using single and collaborative judgmental approaches. Within this context, researchers and practitioners alike are confronted to the issue of selecting the most appropriate forecasting model and conventionally use goodness-of-fit performance criteria in determining the model which produces a forecast with the smallest error versus the actual demand

[1]. These error measures should meet some criteria identified in [2] and show the importance of variety, reliability, ease of interpretation, clarity of presentation, and support of statistical evaluation. However, a single error measure is insufficient in evaluating different forecasting techniques as each measurement may provide diverging results as well as only illustrate specific aspects of the forecast, such as, the impact of an outlier. Moreover, the hybrid approach used may provide a focus on improving only a particular pattern, which leads to substantial changes of some error measures and negligible improvements of some others. Thus, there is a need for a method to allow decision makers to select the most appropriate forecasting model within a group of alternatives based on the varying weights of importance they place on a range of error measures.

Therefore, the question arises: which method is the most appropriate to select in order to have the most reliable forecast for future demand? In an attempt to find an answer to this question, this work proposes the development of a multi-criteria decision making (MCDM) approach to evaluate different forecasting models based on a selected number of widely used error measures in forecasting, their interdependencies and their influence on model selection.

This paper is divided into 6 sections: The next section provides an overview of the state of the art in MCDM application in forecast selection. Section 3 describes the proposed framework to answer our research question followed by an empirical analysis using industrial data from a plastic bag manufacturer in Section 4. Section 5 presents the results and discussion, followed by the conclusion and suggestions for future work in Section 6.

2 State of the Art

Hybrid forecasts, combining both quantitative results and qualitative information, have showed substantial improvements in forecasting accuracy [3]. A systematic approach is developed by [4] to structure and integrate human knowledge of contextual factors into demand forecasting. The statistical forecast based on ARIMA method is considered as a basis and the structured knowledge of the experts is provided to adjust the initial forecasts based on four factors that represent most types of events: transient, transfer, trend change, and quantum jump factors. Further, as sequel to improve the forecasting accuracy, these factors corresponding to the human judgments forecast adjustment are grouped and structured to determine the forecast adjustment due to collaborative human judgement and are subsequently evaluated through a fuzzy inference system which show substantial benefits in the improvement of forecasting accuracy [3]. The collaborative process consists of integrating the judgements of several forecasters and structuring the information using complementarities of the different perceptions. The results of both [3, 4] illustrate an improvement in forecast accuracy based on the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE). However, the choice of a forecasting model cannot be solely selected based on any one specific error measure [5]. The MAE is shown to be weak to outliers as a strong error at one point of the data series will skew the mean [6]. In addition, the MAPE is not a symmetrical error measurement in that equal errors above the actual value result in a

greater average percentage error than those below the actual value. The MAPE is therefore best used when dealing with positive actual observations [7]. The literature highlights the importance of considering several error criteria in evaluating the performance of a forecasting model.

In [1], the authors propose a framework using MCDM methods in selecting the most appropriate regression forecasting models based on a tradeoff between several error measures as their criteria, and apply ELECTRE III, PROMETHEE I, and PROMETHEE II to evaluate the most appropriate forecast for oil prices. They find that the best performing models, a linear regression and exponential smoothing model, are not sensitive to the criteria weighting or the different MCDM methods used. The authors in [8] investigate the use of the PROMETHEE approach for selection of the most appropriate classification forecast algorithms. Although easier to use than ELECTRE, the PROMETHEE method lacks clarity in weight determination for the criteria and ignores potential interdependencies between them [9].

Often hybrid MCDM methods are employed, such as AHP and ANP paired with TOPSIS, which allows for easier ranking between alternatives [10]. A fuzzy AHP-TOPSIS method, including expert weighting, is used in [11] to determine the best alternative between a selection of collaborative software available on the market. Raut et al. [12] integrates the Balanced Scorecard (BSC) with a fuzzy AHP and fuzzy TOPSIS framework to determine the degree of sustainability in the banking industry that was implemented in the six largest commercial banks in India. Wang [13] proposed the hybrid method of decision making trial and evaluation laboratory technique (DEMATEL) with ANP (DANP) to decide upon interactive trade strategy to be adopted for Taiwan. Both [13, 14] employ DEMATEL in order to build a network relations map by investigating the interrelation among aspects and criteria. Chiu et al. [14] assess e-store strategies, with respect to Marketing and Customer Service, by also employing DANP with VIKOR methods to rank the alternatives from a set of often conflicting criteria. The hybrid DANP-VIKOR model aided in determining the effect of e-store management on sales and resulted in recommendations for strategy improvement. The main difference between VIKOR and TOPSIS methods is the aggregating function for determining the ranking of the alternatives. TOPSIS ranks the alternative based on the additive combination of the best and worst distances from the ideal solution whereas the VIKOR method takes into account the relative importance of the distances from the ideal solution by balancing the total and individual satisfaction. The use of n-dimensional Euclidean distance in the TOPSIS method accounts for this balance between the individual and total satisfaction [15].

Contrary to other MCDM methods such as the Multi-Attribute Utility Theory (MAUT) and Multi-Attribute Value Theory (MAVT), AHP and ANP apply pairwise comparison to compare alternatives as well as estimate weighting to the criteria and priority scales [16], and are therefore easy to use [9]. The MAUT/MAVT methods require high precision on the specific criteria weights which prove difficult in real-life circumstances [9]. On the other hand, the AHP and ANP methods are susceptible to rank reversal at the end of the process which could result in the final ranking to be reversed in order. Using TOPSIS addresses the issue of rank reversal when a non-optimal alternative is introduced [17], however the TOPSIS method alone does not consider criteria interrelationships nor provide an easy method of determining criteria weights, often being paired with the AHP and ANP methods [18]. The authors in [19] use a fuzzy AHP alongside

the TOPSIS method to rank fifteen Turkish cement firms based on their financial performance as well as subjective judgmental information. In addition, [20] also propose a hybrid approach using AHP and Fuzzy TOPSIS to rank banking performance in Iran. The hybrid methods AHP-TOPSIS and ANP-TOPSIS have also been applied to support the decision making process for personnel selection in a manufacturing firm [21]. The research concludes that the hybrid approaches of AHP-TOPSIS and ANP-TOPSIS are robust MCDM techniques to evaluate performance based on a set of defined criteria. These hybrid techniques have not yet been used in the evaluation of demand forecasting models based on error measurement criteria. This paper aims to bridge this gap in the literature by proposing the AHP-TOPSIS and ANP-TOPSIS models, to evaluate and rank forecasting models based on multiple error measures as the decision criteria.

3 Proposed Framework

The objective of this paper is to develop an MCDM approach to evaluate and rank quantitative and hybrid (quantitative and qualitative) demand forecasting models with a particular focus on performance measures. Since different forecasting methods usually lead to different error measurements, the criteria selected in the MCDM method are:

- The Mean Error (ME), which indicates whether a forecast is biased, however it is possible that a negative error on one data point would counterbalance a positive error on another data point. [22] suggests that to counter this last effect, the Median can be used over the mean;
- The Mean Absolute Error (MAE), which measures the absolute error however but may skew the mean when confronted with large outliers [6];
- The Mean Percentage Error (MPE) which is based on the actual values rather than the absolute values and therefore is a good measure of the relative size and direction of the bias.
- The Root Mean Squared Percentage Error (RMSPE) takes only positive values due to the squaring function and therefore provides an average relative size of the error. On the other hand, large outliers will dramatically impact the measurement [23].
- The Root Mean Squared Error (RMSE), which is representative of the size of a "typical" error. RMSE gives extra weight to large errors due to the squaring function and is sensitive to scale [6]. In addition, it has been observed that the results frequently differ when applied to various sets of data [23].
- The Mean Absolute Percentage Error (MAPE) is one of the most common error measurements in demand forecasting. MAPE is pulled upward by asymmetrical distributions and outliers. It has a minimum of 0 but no upper boundary. Additionally, MAPE is unit-free and the result is given as a percentage and is best used when dealing with positive actual observations [7].
- The R-squared (R^2) is the coefficient of determination that represents the proportion of variability in a data set that is accounted for by the forecasting model. It provides a measure of how well future outcomes are likely to be predicted by the model. An R^2 measure of 1.0 can be understood in that the regression line perfectly fits the data. Negative values of R^2 may occur when fitting non-linear trends to

data. In seasonal time-series with non-linear trends, we can expect to see negative values of R^2.

These criteria have several key aspects to take into account. First of all, the criteria are for most cases incomparable, in the sense that they are not in the same units (for example, RMSPE and MAPE are in percentages, and R^2 is a ratio and therefore has no unit). Secondly, a common point is that the overall goal is to minimize the set of criteria, as a higher value means a larger error in general. R^2 differs from the others as it does not represent a mean computed using errors however it measures how well future outcomes are likely to be predicted by the model. The weights and importance corresponding to each criteria depends upon the chosen MCDM method.

In this paper, two hybrid MCDM techniques are proposed and developed: AHP-TOPSIS and ANP-TOPSIS, selected due to their robustness (based on the constant number of calculation steps for TOPSIS regardless of the number of attributes), scalability and possibility to integrate interdependencies (in the case of ANP) [9], as well as ability to consider both quantitative and qualitative information [11]. They are used to evaluate and compare five different forecasting techniques. Recognition of the dependence among criteria is also considered to calculate the criteria weights in using ANP as the same data is used to calculate the error measures. In AHP-TOPSIS, the interdependence is omitted. These alternatives are then evaluated with respect to criteria and then ranked.

4 Case study using industrial data

The case study is based on data and additional information collected from Company X, a plastic bag manufacturer in the south of Spain. The forecast experts are asked to analyze the polyethylene bag market to provide a demand forecast. The time-series used are composed of the aggregated monthly demand collected over a period of 36 months (2004 to 2006). The objective is to plan the demand for the year 2007. The three main customers of Company X are all supermarkets. Each expert is invited to analyze the historical data and the influencing factors in the plastic bag demand and forecast the specific events relying on their knowledge.

4.1 Forecasting using empirical case study

The historical data is plotted as a time-series and shows a strong linear trend and seasonality (with peaks in summer and winter - due to demand increases in plastic bags during the months before summer and Christmas holidays, cf. [4]). The forecasts (alternative solutions) are determined using five different techniques: the Holt-Winter decomposition method with multiplicative seasonality [24], the ARIMA and SARIMA Methods, and two Hybrid Forecasting processes that include human judgement: ARIMA integrating single judgmental adjustment, and ARIMA integrating collaborative judgmental adjustment using a team of three experts.

With comparison to the actual data, Holt-Winters gives results that follow the demand seasonality, but has a lower amplitude than the actual data. The Holt-Winters method provides a forecast with few outliers shown by the close values of the MAE and RMSE, 75.9 and 92.5 respectively, in Table 1. Figure 1 illustrates that this forecast has values

that are relatively close (both above and below) to the actual results. The error measures in Table 1 quantify these discrepancies and show the ME and MPE (both non-absolute measures) are relatively low at 36.9 (from a demand of >1200) and 3.27%, respectively. In addition, the R^2 measure of 0.61 shows us that the fit to the regression is not too low for a time-series presenting strong seasonality.

The second forecasting method analyzed is the ARIMA (5,0,4) method (based on [4]). Figure 1 shows that the ARIMA method follows the actual demand relatively well, including the seasonality peaks (albeit with a slight delay on the second peak). The fit is better at the beginning of the time-series than at the end where there is a more noticeable positive error. The visual analysis is supported by the error measures in Table 1, where the ME and MPE are very low (1.33 and 0.51% respectively) suggesting an equal distribution of positive and negative errors. The RMSE is marginally larger than the MAE, which can explain the larger positive error in the later months. The R^2 measure of 0.24 is very low attributing to the high variance of error along the time-series.

The third model is the SARIMA forecast which introduces seasonality to the ARIMA (5,0,4) model used as the basis. The three parameters representing the orders of the seasonal autoregressive and moving average parts of the model are determined by simulating various configurations, using the R programming language and free software, to yield the lowest ME and RMSE, which result in (1,1,1). The results for the SARIMA (5,0,4)(1,1,1) are illustrated in Figure 1 showing a strong adherence to the real data and its seasonality. As the time-series in this case study explicitly shows a seasonal character, it was expected that SARIMA provides a better forecast than ARIMA. The forecast initially has some difficulty following the curve at the beginning of the 12-month period, followed by a very good fit in the later months (June to October) and then another deviation in the final two months of the forecast.

Analysis of the error measures of the SARIMA forecast in Table 1 shows a relatively low ME and MPE of 17.50 and 1.70%, respectively, showing a near-equal positive and negative error distribution. In addition, the closeness in the values of MAE and RMSE indicates the absence of outliers.

Table 1. Error Measures per forecasting model

Error Measures	Holt-Winter	ARIMA	SARIMA	Single Adjustment + ARIMA	Collab Adjustment + ARIMA
ME	36.94	1.33	17.5	-52.42	5.58
MAE	75.99	112.17	92	97.42	33.25
MSE	8562.4	16821	11473.33	29564.92	1836.08
RMSE	92.53	129.7	107.11	171.94	42.85
MPE	3.27 %	0.51 %	1.70 %	-3.29 %	0.47 %
MSPE	0.54 %	0.95 %	0.63 %	1.22 %	0.09 %
RMSPE	7.35 %	9.74 %	7.97 %	11.06 %	3.04 %
MAPE	5.78 %	8.19 %	6.68 %	6.66 %	2.36 %
R2	0.61	0.24	0.48	-0.34	0.92

The final two hybrid forecasting methods utilize both time-series and qualitative information to determine the forecast. The first uses ARIMA(5,0,4) single judgement adjustment in which the expert judgement is integrated in a structured manner as complementary information to the ARIMA(5,0,4) forecast. It can be seen from Figure 2 that this forecast has a very good fit with the actual data, including month-over-month trend, seasonality, and peaks. Conversely, this forecast has a very strong negative outlier error (-549) due to human error when the decision maker was adjusting the mathematical forecast. In addition, there are very large differences observed between the MAE and RMSE of 97.42 and 171.94, respectively (Table 1) which shows the strong impact of the outlier. However, as the remaining fit is very good, other error measures mitigate the outlier's impact. The R^2 measure is negative at -0.34, which is expected due to the extremely high variation in errors due to the outlier and thus non-linear error trend.
The second hybrid forecasting method utilizes the collaborative expert judgement of several people [3] and is based on ARIMA(5,0,4). Figure 2 shows a very strong overall fit between the forecast and the actual data. The only minor differences observed are slight positive errors at the beginning of each seasonal rising slope (March and July). Both the ME and MPE are small showing quasi-equal distribution of positive and negative errors in Table 1, and the MAE and RMSE are also very low due to the impact of the two small peak errors. The R^2 measure of 0.92 shows that there is an extremely good fit between forecast and data: there is very little difference in error trend along the time-series.

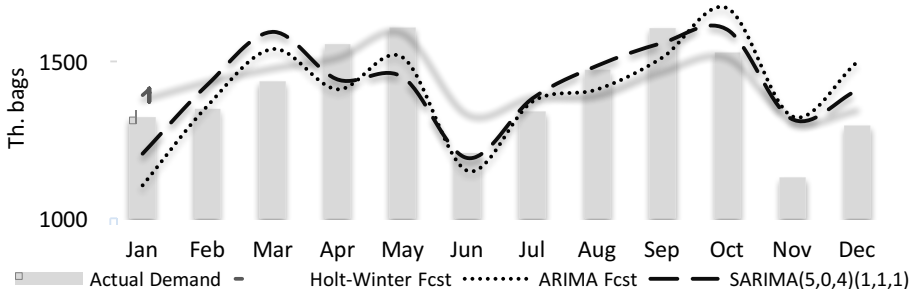


Fig. 1. Demand vs Quantitative Forecasts for 12 months horizon

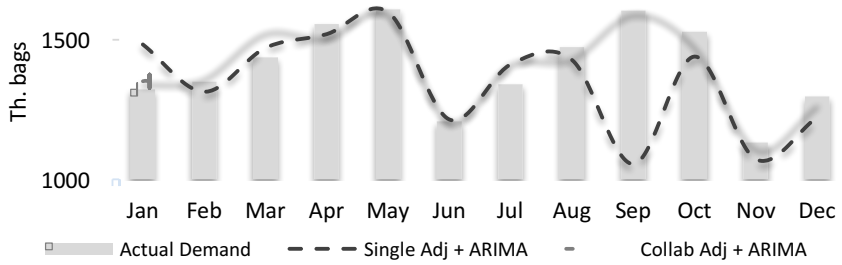


Fig. 2. Demand vs Hybrid Forecasts for 12 months horizon

4.2 Solving the MCDM Problem

We develop two hybrid decision-making models: AHP-TOPSIS and ANP-TOPSIS and apply them to compare the forecasting models.
For both AHP and ANP, a pairwise comparison matrix is formed by experts using Saaty's 1-9 scale [25] on the identified error measures criteria. Initially, no interdependencies are considered. Their weights in terms of priority are determined and presented in Table 2. The consistency ratio (CR) for this case is 0.0935, therefore considered acceptable as it is < 0.1 [26]. Priorities are selected based on the amount of useful information conveyed by the criteria. The RMSE has a little higher overall priority to MAE due to its use in identifying the presence of outliers (square power). Following the pairwise comparisons, a normalized weight vector *w* is determined by calculating the eigenvectors of the priority matrix [21].

Table 2. Criteria priorities & weights using AHP

Criteria	ME	MAE	RMSE	MPE	RMSPE	MAPE	R^2	Weights (w)
ME	1	1/5	1/4	1/3	1/5	1/5	1	0.0464
MAE	5	1	1/3	1	1/3	1/3	1	0.0975
RMSE	4	3	1	3	1	1/3	1	0.1666
MPE	3	1	1/3	1	1/3	1/5	1	0.0802
RMSPE	5	3	1	3	1	1/3	1	0.1726
MAPE	5	3	3	5	3	1	1	0.3075
R^2	1	1	1	1	1	1	1	0.1292

In order to consider the interdependencies between criteria for the ANP, another pairwise comparison matrix is created where the decision makers examine the impact of all the criteria on each other. In this case, interdependencies are chosen based on the way the criteria are calculated. For example, R^2 is not calculated in the same way as any other criteria, but the MAE and MPE both show components of how MAPE is calculated. The links are shown in the Figure 3.

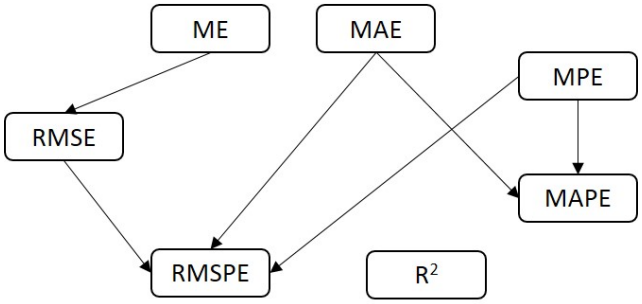


Fig. 3. Interdependencies between MCDM criteria

Table 3. Normalized interdependences and weights (ANP-TOPSIS)

Criteria	ME	MAE	RMSE	MPE	RMSPE	MAPE	R^2	Weight (w_n)
ME	1	0	0.106	0	0	0	0	0.0642
MAE	0	1	0.149	0	0	0.200	0	0.1838
RMSE	0	0	0.745	0	0.138	0	0	0.1479
MPE	0	0	0	1	0.172	0.200	0	0.1714
RMSPE	0	0	0	0	0.690	0	0	0.1190
MAPE	0	0	0	0	0	0.600	0	0.1845
R^2	0	0	0	0	0	0	1	0.1292

Using the hybrid methodology in [21], a decision matrix for the ranking of alternatives is created on the basis of the criteria per alternative. The decision table is then normalized, to allow for comparison between values, shown in Table 3.

The final part of both AHP-TOPSIS and ANP-TOPSIS hybrid models consists of determining the positive-ideal and negative-ideal solutions, then calculating each alternative's distance to them and establishing a ranking based on these distances. To establish the positive-ideal solution vector (noted V+) and the negative-ideal solution vector (noted V-), the minimum value is determined from the alternatives for each criteria. In the same way, we determine which value is maximum for V- from the alternatives for each criteria (Table 5).

Subsequently the separation measures are calculated using the Euclidian distance [21]. The separation of each alternative to the positive-ideal V+ is noted D+. Similarly, the separation of each alternative from the negative-ideal solution V- is noted D-. The relative closeness to the ideal solution is calculated as $C_i = \frac{D_i^-}{D_i^- + D_i^+}$. and the performance order is ranked (Table 6 for AHP-TOPSIS and 7 for ANP-TOPSIS). A larger index value means the better the performance of the alternative.

Table 5. Positive-ideal and negative-ideal solutions for AHP-TOPSIS and ANP-TOPSIS

Error Measures		ME	MAE	RMSE	MPE	RMSPE	MAPE	R^2
AHP-TOPSIS	V+	-0.0364	0.01676	0.00828	0.01315	-0.1139	0.0161	0.07702
	V-	0.02571	0.05653	0.13331	0.05278	0.11303	0.21353	0.02115
ANP-TOPSIS	V+	-0.0504	0.03159	0.02425	-0.1131	0.01949	0.03124	0.09315
	V-	0.03553	0.10658	0.09733	0.11227	0.07098	0.10819	-0.035

Table 6. Distances and final ranking (AHP-TOPSIS)

	Holt-Winter	ARIMA	Single Adj	Collab Adj	SARIMA
D+	0.252601085	0.21921914	0.23921742	0.14738874	0.15769996
D-	0.15807352	0.13286061	0.24201139	0.25998246	0.148354036
C	0.38491185	0.37735942	0.50290295	0.63819548	0.48473157
Rank	**4**	**5**	**2**	**1**	**3**

Table 7. Distances and final ranking (ANP-TOPSIS)

	Holt-Winter	ARIMA	Single Adj	Collab Adj	SARIMA
D+	0.253813597	0.200892915	0.177048763	0.140907385	0.211287706
D-	0.11918427	0.11944873	0.24253111	0.21469460	0.11289793
C	0.31953071	0.37287917	0.57803324	0.60374973	0.34825088
Rank	**5**	**3**	**2**	**1**	**4**

5 Results and Discussion

In both AHP-TOPSIS (Table 6) and ANP-TOPSIS (Table 7), a recurring aspect is that the hybrid forecasts combining mathematical forecasts with expert adjustments are ranked higher than pure mathematical forecasts. In addition, the collaborative expert adjustment method is ranked highest using both MCDM hybrid models. In fact, its relative closeness C to the ideal solution is much higher than the other methods, especially using AHP-TOPSIS. Single adjustment comes closer to collaborative adjustment in ANP-TOPSIS (C=0.60 to C=0.58) due to the interdependence between error measures. The results indicate that the judgmentally adjusted forecasts, whether done by a single expert or as a collaboration between a number of experts, is better than the purely statistical methods.

On the other hand, the purely mathematical models ARIMA, SARIMA and Holt-Winter models, all changed in ranking between the AHP-TOPSIS and ANP-TOPSIS methods. In particular, the models that integrate seasonality, SARIMA and Holt-Winter, indicate a stronger sensitivity to the weights of the criteria as well as their interdependencies. Conversely, ARIMA seemed to be least sensitive to the change in weights in the criteria between the two MCDM methods, even though it had the largest change in ranking relative to the other four forecasting models. Surprisingly, the Holt-Winter method is consistently ranked in the bottom two, as two of its resulting weighted criteria are chosen for the negative-ideal solution and none for the positive (particularly due to the high ME). Based on the comparison with SARIMA, it could be argued that SARIMA provides a better mathematical basis for the judgmental adjustment methods than ARIMA, due to a much higher value of C. Though SARIMA is ranked 4th in ANP-TOPSIS, after ARIMA, their respective values of C are very close (0.37 to 0.35 respectively).

The AHP-TOPSIS and ANP-TOPSIS methods, which produce a final ranking in terms of relative distance to ideals, the actual error measures in respect to weights for each alternative are less decisive than their proximity to the positive-ideal/negative-ideal solutions.

6 Conclusion and Future Work

This paper develops two hybrid MCDM approaches, namely AHP-TOPSIS and ANP-TOPSIS, to support the selection of the most appropriate demand forecasting method applied to a plastic bag manufacturer case. The alternative forecasting methods include typical quantitative regression and hybrid demand forecasting models. The criteria on

which the forecasts are evaluated are their goodness-to-fit, measured as the error between the forecasts and actual sales. Several error measures are considered for the selection in a first step as independent for which AHP-TOPSIS is developed, and in a second step, as interdependent, for which ANP-TOPSIS is developed. The results show that in both cases, the judgmental adjusted forecasts are ranked highest, and that collaborative adjustment provides better results than single adjustment since it takes into account complementary judgments from several experts. However, the rankings are strongly influenced by the chosen criteria weights and nature (both priority and interdependence).
Although it shows high sensitivity to the interdependencies between criteria, it is recommended to replace the ARIMA model with SARIMA as the mathematical basis in the judgmentally adjusted forecasts for the industrial case study. Even though SARIMA changed rank between the two MCDM methods, the relative closeness C is still very close to the result for ARIMA (in the case of ANP-TOPSIS).
As it is demonstrated that collaborative judgment adjustments improve the forecasts and reduce the risk of significant outliers, future research will investigate the impact of the number of experts taking part to the collaborative consensus process, taking into account SARIMA as the mathematical basis for the approach.

7 Acknowledgement

The research leading to these results was funded by the Swiss National Science Foundation under project n° [176349], "Inventory management of short life cycle products with demand forecasts using Big data and judgmental information"

8 References

1. Xu, B., Ouenniche, J.: Performance evaluation of competing forecasting models: A multidimensional framework based on MCDA. Expert Systems with Applications. 39, 8312–8324 (2012).
2. Tayman, J., Swanson, D.A.: On the validity of MAPE as a measure of population forecast accuracy. Population Research and Policy Review. 18, 299–322 (1999).
3. Cheikhrouhou, N., Marmier, F., Ayadi, O., Wieser, P.: A collaborative demand forecasting process with event-based fuzzy judgements. Computers & Industrial Engineering. 61, 409–421 (2011).
4. Marmier, F., Cheikhrouhou, N.: Structuring and integrating human knowledge in demand forecasting: a judgemental adjustment approach. Production Planning & Control. 21, 399–412 (2010).
5. Mahmoud, E.: Accuracy in forecasting: A survey. Journal of Forecasting. 3, 139–159.
6. Chatfield, C.: Apples, oranges and mean square error. International Journal of Forecasting. 4, 515–518 (1988).
7. Ren, L., Glasure, Y.: Applicability of the Revised Mean Absolute Percentage Errors (MAPE) Approach to Some Popular Normal and Non-normal Independent Time Series. International Advances in Economic Research. 15, 409–420 (2009).
8. Mehdiyev, N., Enke, D., Fettke, P., Loos, P.: Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Computer Science. 95, 264–271 (2016).

9. Velasquez, M., Hester, P.T.: An Analysis of Multi-Criteria Decision Making Methods. International Journal of Operations Research. 10, 11 (2013).
10. Sipahi, S., Timor, M.: The analytic hierarchy process and analytic network process: an overview of applications. Management Decision. 48, 775–808 (2010).
11. Kara, S.S., Cheikhrouhou, N.: A multi criteria group decision making approach for collaborative software selection problem. Journal of Intelligent & Fuzzy Systems. 26, 37–47 (2014).
12. Raut, R., Cheikhrouhou, N., Kharat, M.: Sustainability in The Banking Industry: A Strategic Multi-Criterion Analysis. Business Strategy and the Environment. 26, 550–568.
13. Wang, T.-C.: The interactive trade decision-making research: An application case of novel hybrid MCDM model. Economic Modelling. 29, 926–935 (2012).
14. Chiu, W.-Y., Tzeng, G.-H., Li, H.-L.: A new hybrid MCDM model combining DANP with VIKOR to improve e-store business. Knowledge-Based Systems. 37, 48–61 (2013).
15. Opricovic, S., Tzeng, G.-H.: Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. European Journal of Operational Research. 156, 445–455 (2004).
16. Saaty, T.L.: Decision Making with Dependence and Feedback: The Analytic Network Process : the Organization and Prioritization of Complexity. Rws Publications (2001).
17. Zanakis, S.H., Solomon, A., Wishart, N., Dublish, S.: Multi-attribute decision making: A simulation comparison of select methods. European Journal of Operational Research. 107, 507–529 (1998).
18. Tao, L., Chen, Y., Liu, X., Wang, X.: An integrated multiple criteria decision making model applying axiomatic fuzzy set theory. Applied Mathematical Modelling. 36, 5046–5058 (2012).
19. Ertuğrul, İ., Karakaşoğlu, N.: Performance evaluation of Turkish cement firms with fuzzy analytic hierarchy process and TOPSIS methods. Expert Systems with Applications. 36, 702–715 (2009).
20. Beheshtinia, M.A., Omidi, S.: A hybrid MCDM approach for performance evaluation in the banking industry. Kybernetes. 46, 1386–1407 (2017).
21. Dağdeviren, M.: A hybrid multi-criteria decision-making model for personnel selection in manufacturing systems. Journal of Intelligent Manufacturing. 21, 451–460 (2010).
22. Franses, P.H.: Averaging Model Forecasts and Expert Forecasts: Why Does It Work? Interfaces. 41, 177–181 (2011).
23. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting. 8, 69–80 (1992).
24. Holt, C.C., Modigliani, F., Muth, J.F., Simon, H.A., Bonini, C.P., Winters, P.R.: Planning Production, Inventories, and Work Force. Prentice Hall (1960).
25. Saaty, T.L.: How to make a decision: The analytic hierarchy process. European Journal of Operational Research. 48, 9–26 (1990).
26. Chu, P., Liu, J.K.-H.: Note on Consistency Ratio. Mathematical and Computer Modelling. 35, 1077–1080 (2002).

Keyword Index

.

Author Index