



ITISE 2018

**International Conference on
Time Series and
Forecasting**

**PROCEEDINGS
OF
PAPERS**

Volumen 3

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.

All rights reserved to authors. The total or partial reproduction of this work is strictly prohibited, without the strict authorization of the copyright owners, under the sanctions established in the laws.

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 Remillard, 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-Igzuiza, Professor at Instituto Geologico y Minero de Espaa (IGME) and Prof. Francisco Javier Rodriguez-Tovar, Depart. Estratigrafia 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 University 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 Remillard	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, Faculty 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

Extended Talks	
Note on Whittle Type Estimation under Long Memory and Nonstationarity	1
<i>Uwe Hassler and Ying Lun Cheung</i>	
An overall seasonality test based on recursive feature elimination in conditional random forests	20
<i>Karsten Webel and Daniel Ollech</i>	
Simulation-based selection of prediction models	32
<i>Robert Kunst</i>	
<hr/>	
Expert systems and recent developments with Time Series- Data	
<hr/>	
Robust autocovariance estimation from the frequency domain	42
<i>Higor Henrique Aranda Cotta, Valdério Reisen, Pascal Bondon and Celine Levy-Leduc</i>	
Penalty terms for estimation of ARMA models: A Bayesian inspiration	54
<i>Helgi Tómasson</i>	
Towards an API for EEG-Based Imagined Speech classification	64
<i>Luis Alfredo Moctezuma and Marta Molinas</i>	
A simulation of a custom inspection in the airport	76
<i>Kalle Saastamoinen, Petteri Mattila and Antti Rissanen</i>	
Complex networks of scalar time series using a data compression algorithm	90
<i>Debora Correa, David Walker and Michael Small</i>	
Computation and validation of wind and solar time series based on global reanalysis	92
<i>Marta Victoria, Gorm B. Andresen and Martin Greiner</i>	
<hr/>	
Applications in Time Series (Part. I)	
<hr/>	
A Study with NDVI Time Series of the Brazilian Caatinga	95
<i>Claudionor Silva, Aracy Araujo and Sérgio Machado</i>	
Characterizing Market Behavior through Risk Forecasts: a Powerful VaR Backtesting	99
<i>Marta Malecka</i>	
The Long-term memory effects of the Baltic Dry Index	111
<i>Jose Ramon San Cristobal</i>	
Forecasting Peak Period of Travel Time	119
<i>Béla Paláncz, Jianhong Xia and Yuchen Liu</i>	
Transfer function modeling of constant work-rate tests in patients with COPD	130
<i>Joren Buekers, Hanne Cryns, Patrick De Boever, Emiel F.M. Wouters, Martijn A. Spruit, Jan Theunis and Jean-Marie Aerts</i>	

Adaptive R-peak Detection Using Empirical Mode Decomposition	134
<i>Christina Kozia, Randa Herzallah and David Lowe</i>	

Energy Forecasting

Understanding the behaviour of energy prices in Brazil	146
<i>Abdinardo Moreira Barreto de Oliveira and Anandadeep Mandal</i>	
Time series Analysis for Re-Commissioning of Building Service installations	158
<i>Wim Zeiler, Albert Jan Huls and Ben Lops</i>	
Adaptive Methods for Energy Forecasting of Production and Demand of Solar Assisted Heating Systems	170
<i>Viktor Unterberger, Thomas Nigitz, Mauro Luzzu, Daniel Muschick and Markus Gölles</i>	
Prediction of Current by Artificial Neural Networks in a Substation in order to Schedule Thermography	182
<i>Per Westerlund and Ilias Dimoukas</i>	

Real macroeconomic monitoring and forecasting (Part. I)

Permutation entropy as the measure of globalization process.....	192
<i>Janusz Miśkiewicz</i>	
Estimating macroeconomic uncertainty from surveys – a mixed frequency approach.....	197
<i>Jeffrey Sheen and Ben Wang</i>	
External Migration as a Factor of Economic Growth: Econometric Analysis for CIS Countries	227
<i>Kseniia Bondarenko</i>	
Business Cycle Synchronizaiton: The effects of Trade, Sectoral and financial linkages	239
<i>Kanya Paramaguru</i>	

Atmospheric Science Forecasting

Localized Online Weather Predictions with Overnight Adaption	250
<i>Michael Zauner, Michaela Killian and Martin Kozek</i>	
Storm characterization using a BME approach	260
<i>Manuel Cobos, Andrea Lira-Loarca, George Christakos and Asunción Baquerizo</i>	
Air Pollution Forecasting using Machine Learning Techniques	264
<i>Marijana Cosovic and Emina Junuz</i>	

Advanced econometric methods

Forward Regression with Discrete and Continuous Wavelet Time-Frequency Window -An application to the Market Line-	274
<i>Roman Mestre and Michel Terraza</i>	
Using subspace methods to model long memory processes	288
<i>Dietmar Bauer</i>	

Changepoints to Improve Forecasts.....	300
<i>Jamie-Leigh Chapman, Rebecca Killick and Idris Eckley</i>	

Health Forecasting

ProMoBed: a forecasting and simulation model for estimating future hospital bed capacity.....	302
<i>Marlies Van der Wee, Timo Latruwe, Sofie Verbrugge, Pieter Vanleenhove, Henk Vansteenkiste and Sebastiaan Vermeersch</i>	
Panel Data Unit Root Tests on the Income-Health Relationship of the Mexican States....	306
<i>Vicente German-Soto and Martha Elena Fuentes Castillo</i>	
Forecasted trends for cardiovascular disease in England and Wales to 2040 and impact of reduction in smoking prevalence: a Markov modelling study	318
<i>Sara Ahmadi-Abhari, Piotr Bandosz, Maria Guzman-Castillo, Hannah Whittaker, Martin Shipley, Mika Kivimäki, Simon Capewell, Martin O'Flaherty and Eric Brunner</i>	
Forecasting in qPCR procedure by means of hyperbolastic stochastic model	325
<i>Antonio Barrera, Patricia Román-Román and Francisco Torres-Ruiz</i>	
Effects of Electrical Stimulation on Cortical Phase Synchronization as a Measure of Excitability	327
<i>Farrokh Manzouri, Matthias Duempelmann, Christian Meisel and Andreas Schulze-Bonhage</i>	
Using time series analysis for challenging breast lesion detection and classification in DCE-MRI	331
<i>Ignacio Alvarez, Anthony Bagnall, Javier Ramirez, Juan Manuel Gorriz, Katja Pinker, Maria Adele Marino, Daly Avendaño and Anke Meyer-Baese</i>	

Econometric models (Part.I)

Relationships between Shanghai, Shenzhen and Hong Kong Stock Markets considering the split-share reform	332
<i>Yang Mestre-Zhou, François Benhmad and Roman Mestre</i>	
Economic and Environmental Benefits Based on Sce-nario Analysis in Transportation Sector: A Case Study of Kuwait.....	350
<i>Sarah Alosaimi and K. J. Sreekanth</i>	
Tourism – the factor of employment sustainability in Croatian economy	362
<i>Justin Pupavac and Drago Pupavac</i>	

Computational Intelligence methods for Time Series

Enhancement of time series analysis by including label variables	373
<i>José Carlos García-García, Ricardo García-Ródenas and Francisco P. Romero</i>	
Direct and Recursive Strategies for Multi-Step Ahead Wind Speed Forecasting	385
<i>Sameer Al-Dahidi and Hisham Elmoaqet</i>	
Identification of multiregime periodic autoregressive models by genetic algorithms	396
<i>Domenico Cucina, Manuel Rizzo and Eugen Ursu</i>	

Change Detection for Streaming Data using Wavelet-based Least Squares Density Difference	408
<i>Nenad Mijatovic, Rana Haber, Mark Moyou, Anthony O. Smith and Adrian M. Peter</i>	
Fuzzy time series applications and extensions: analysis of a short term load forecasting challenge	420
<i>Guilherme Costa Silva, João Luis R. Silva, Adriano Lisboa, Douglas Vieira and Rodney Saldanha</i>	
Selection of neural network for crime time series prediction by Virtual Leave One Out tests	432
<i>Stanislaw Jankowski, Zbigniew Szymański, Zbigniew Wawrzyniak, Paweł Cichosz, Eliza Szczechla and Radosław Pytlak</i>	
Data Mining Applied for Performance Index Prediction in Highway Long Segment Maintenance Contract	444
<i>Andri Irfan, Susanti Handayani and Merry Lita</i>	
Novel order patterns recurrence plot-based quantification measures to unveil deterministic dynamics from stochastic processes	457
<i>Shuixiu Lu, Sebastian Oberst, Guoqiang Zhang and Zongwei Luo</i>	

Spatio-temporal brain dynamics in attention tasks

On Statistical Inference for Independent Colored Sources Analysis	469
<i>Young Truong and Rachel Nethery</i>	
Relevance analysis in spatio-spectral components based on Permutation Entropy supporting MI discrimination	489
<i>Juan Camilo López Montes, David Cárdenas Peña and German Castellanos Dominguez</i>	
Entropy-based relevance selection of independent components supporting motor imagery tasks	499
<i>David Felipe Luna Naranjo, David Cardenas Peña and German Castellanos Dominguez</i>	
Sub-band brain mapping based on a Multivariate Wavelet Packet Decomposition	509
<i>Pablo Andrés Muñoz Gutiérrez, Eduardo Giraldo, Juan David Martinez Vargas and German Castellanos Dominguez</i>	
Localizing the Focal Origin of Epileptic Activity using EEG Brain Mapping based on Empirical Mode Decomposition	519
<i>Pablo Andrés Muñoz Gutiérrez, Eduardo Giraldo, Marta Molinas and Maximiliano Bueno López</i>	

Forecasting performance evaluation

Performance Assessment of A short-Term Travel Forecasting Scheme for Multi-Lane Highway	529
<i>Jamal Raiyn</i>	
On the limits of probabilistic prediction in nonlinear time series analysis	550
<i>Jose Maria Amigo, Yoshito Hirata and Kazuyuki Aihara</i>	

Evaluation of regression and judgement-incorporated forecasting processes using hybrid MCDM models.....	559
<i>Yvonne Badulescu and Naoufel Cheikhrouhou</i>	
Outlier Identification in Multivariate Time Series: Boilers Case Study	571
<i>Joana Ribeiro, Mário Antunes, Diogo Gomes and Rui Aguiar</i>	
Realized volatility in the presence of structural breaks: which forecast?	583
<i>Giuseppina Albano and Davide De Gaetano</i>	

Applications in Time Series (Part.II)

Experimental Comparison and Tuning of Time Series Prediction for Telecom Analysis	586
<i>Andrè Pinho, Pedro Furtado and Helena Silva</i>	
Multivariate forecasting of extreme wave climate and storm evolution.....	598
<i>Andrea Lira-Loarca, Manuel Cobos, Asunción Baquerizo and Miguel A. Losada</i>	
Pattern similarity-based load forecasting applied to unit commitment problem	602
<i>Guilherme Costa Silva, Adriano Lisboa, Douglas Vieira and Rodney Saldanha</i>	
Modified Granger Causality in Selected Neighborhoods.....	614
<i>Martina Chvosteková</i>	
State of Charge Depended Modeling of an Equivalent Circuit of Zinc Air Batteries Using Electrochemical Impedance Spectroscopy	625
<i>Andre Loechte, Ole Gebert, Ludwig Horsthemke, Daniel Heming and Peter Gloesekoetter</i>	
Cryptanalysis of a Chaos Based Encryption Algorithm for Secure Communication	637
<i>Salih Ergun</i>	

Times series analysis in geosciences

Local fractal analysis of time series.....	645
<i>Eulogio Pardo-Igúzquiza, F. J. Rodríguez-Tovar and J. Sanchez-Morales</i>	
Discussion on Geodetic Times Series of Mixed Spectra and Levy Processes	654
<i>Jean-Philippe Montillet and Kegen Yu</i>	
Daily reference evapotranspiration forecasting for oceanic climate using autoregressive Hilbertian process.....	665
<i>Rousseau Tavegoum, Besnik Pumo and Pierre Santagostini</i>	

Forecasting Complex/Big data (Part. I)

Characterization and detection of potential fraud taxpayers in Personal Income Tax using data mining techniques.....	677
<i>María Del Camino González Vasco, Maria Jesús Delgado Rodríguez and Sonia de Lucas Santos</i>	
Detecting Anomalous Pattern-of-Life from Human Trajectory Data.....	717
<i>Yazan Qarout and David Lowe</i>	

Model-based Data Exploration	729
<i>Hans-Ulrich Kobialka, Daniel Paurat and Lisa Schrader</i>	

Nonstationarity Time Series

Identification of nonstationary processes using noncausal bidirectional lattice filtering	741
<i>Maciej Niedzwiecki and Damian Chojnacki</i>	
Likelihood based inference for an Identifiable Fractional Vector Error Correction Model ...	753
<i>Katarzyna Lasak and Federico Carlini</i>	
Identification Algorithms Based on the Associative Search of Analogs and Association Rules	783
<i>Natalia Bakhtadze, Vladimir Lototsky, Valery Pyatetsky and Alexey Lototsky</i>	

Real Macroeconomic Monitoring and Forecasting (Part.II)

The impact of the increased domestic energy prices on the Saudi Arabian economy. Insights from KGEMM.	795
<i>Fakhri Hasanov, Frederic Joutz and Jeyhun Mikayilov</i>	
Yield Curve Modeling with Macro Factors	798
<i>András Bebes, Dávid Tran and László Bebesi</i>	
Ranking multi-step system forecasts invariant to linear transformations	811
<i>Håvard Hungnes</i>	

Advanced methods in Forecasting

Conditional Heteroskedasticity in Long Memory Model FIMACH' for Return Volatilities in Equity Markets	825
<i>A.M.M. Shahiduzzaman Quoreshi and Sabur Mollah</i>	
Probabilistic forecasting and simulation of electricity prices	852
<i>Peru Muniain and Florian Ziel</i>	
Computing Environment for Forecasting based on System Dynamics Models	864
<i>Radoslaw Pytlak, Damian Suski, Tomasz Tarnawski, Zbigniew Wawrzyniak, Tomasz Zawadzki and Pawel Cichosz</i>	
The Contrast Between Management Consulting and Outsourcing Management Services: A financial perspective	876
<i>Carlos Jerónimo, Leandro Pereira, José Santos and Nelson Antonio</i>	
FPGA-based accelerator design for Echo-State networks	883
<i>Josep L Rossello, Miquel L. Alomar, Erik Sebastian Skibinsky Gitlin, Christiam F Frasser, Vicente Canals, Eugeni Isern, Fabio Galan Prado, Alejandro Morán and Miquel Roca</i>	
Stacked LSTM Snapshot Ensembles for Time Series Forecasting	895
<i>Sascha Krstanovic and Heiko Paulheim</i>	

Econometric models (Part.II)

Implications for Aggregate Inflation of Sectoral Asymmetries: an empirical application ...	907
<i>Hannu Koskinen and Jouko Vilmunen</i>	
Testing for Differences in Forecast-Error Dynamics in Path Forecasts	920
<i>Andrew Martinez</i>	
What can drive economic growth in Russia? Mid-term growth scenarios	921
<i>Svetlana Balashova, Vladimir Matyushok and Inna Lazanyuk</i>	
Determining the cointegration rank using a Residual-based Procedure	933
<i>Antonio Aznar</i>	
<hr/> Quantum Computing <hr/>	
Point Function Analysis and a Hypothesis on the Origin of Quantum Mechanics	952
<i>Bernd Burchard</i>	
<hr/> Structural Time Series Models <hr/>	
Dynamic Bayesian smooth transition autoregressive models applied to hourly electricity load in southern Brazil	966
<i>Alvaro Faria and Alexandre Santos</i>	
CP-based cloud workload annotation as a preprocessing for anomaly detection using deep neural networks	982
<i>Gilles Madi Wamba and Nicolas Beldiceanu</i>	
Time series modelling with MATLAB: the SSpace toolbox	994
<i>Diego J. Pedregal, Marco A. Villegas, Diego Villegas and Juan R. Trapero</i>	
Multivariate INAR processes - Periodic case	997
<i>Cláudia Santos, Isabel Pereira and Manuel Scotto</i>	
<hr/> Advanced in Time Series and Forecasting (Poster presentation) <hr/>	
The Impact of Feedback Trading on Option Prices	1009
<i>Thorsten Lehnert</i>	
Physical Laws Extracted from Statistical Analyses of Solar Magnetic Elements	1010
<i>Mohsen Javaherian and Hossein Safari</i>	
A robust alternative for the estimation of autocovariance from the frequency domain for multivariate processes	1011
<i>Higor Henrique Aranda Cotta, Valdério Reisen, Pascal Bondon and Céline Lévy-Leduc</i>	
Changes in rapeseed canopy spectral reflectance under different cultivars and nitrogen levels	1013
<i>Hong-Xin Cao, Wei-Tao Chen and Bao-Jun Zhang</i>	
Application of Deep-Learning Algorithm for Inflow Series Forecasting in South Korea	1015
<i>Jun-Haeng Heo, Ju-Young Shin and Taereem Kim</i>	
Evaluation of Atmospheric Particulate Matter (PM10) Time Series in Badajoz, 2010-2015	1017
<i>Selena Carretero-Peña, Conrado Miró Rodríguez and Eduardo Pinilla-Gil</i>	

Long-term (2010-2015) tropospheric ozone temporal series in Badajoz (Spain). Trend and seasonal behavior	1022
<i>María Cerrato Alvarez, Conrado Miró Rodríguez and Eduardo Pinilla-Gil</i>	
Verification on winter rapeseed (<i>Brassica napus</i> L.) aboveground dry weight and yield models under waterlogging stress at anthesis	1026
<i>Hong-Xin Cao, Tai-Ming Yang and Bao-Jun Zhang</i>	
On the Impact of Shale Oil Revolution in Oil-Dollar Comovement	1028
<i>Francois Benhmad</i>	
Forecasting inflation with long-short term memory recurrent neural networks: the Colombian case	1036
<i>Andres C. Serna, Javier G. Diaz and Julio Alonso</i>	
Hybrid forecasting methods applied to the Earth's rotation and Radon time-series for anomalies detection	1038
<i>Fabrizio Ambrosino, Lenka Thinová, Miloš Briestenský and Carlo Sabbarese</i>	
Analyses of the time series based on atmospheric energy budget determination for the purpose of budget prognosis with ARMA method	1041
<i>Monika Birylo</i>	
The role of oil prices on the Russian business cycle	1051
<i>Yi Zheng and Harri Pönkä</i>	
Seasonal Variations of Sea Level in the Polish Coastal Zone from Satellite Altimetry and Tide Gauge Data	1063
<i>Katarzyna Pajak, Monika Birylo, Joanna Kuczynska-Siehn and Kamil Kowalczyk</i>	
The Performance of the Wavelet Halt-Winters Hybrid Model in Forecasting the Groundwater Level Time Series (Case Study: Urmeih Coastal Aquifer, Iran).....	1073
<i>Hamid Reza Nassery, Ali Mirarabi, Mohammad Nakhaei and Farshad Alijani</i>	
Tipping point analysis and its applications in geophysics, environmental sciences, and smart sensor systems.....	1089
<i>Valerie Livina</i>	
Combination of neural network and wavelet to predict suspended sediment load in river by using data clustering.....	1091
<i>Samir Bengherifa, Abd El Wahab Lefkir and Abd El Malek Bermad</i>	
Investigation and forecasting of hydrological time series	1096
<i>Svetlana Polukoshko</i>	
Using a naive Bayes classifier to explore the factors driving the harmful dinoflagellate <i>Alexandrium minutum</i> dynamics	1108
<i>Wafa Feki, Asma Hamza, Hasna Njah, Nouha Barraç, Mabrouka Mahfoudi, Ahmed Rebai and Malika Bel Hassen</i>	
Modeling Global Radiation in Kuwait	1110
<i>Shafiqah Alawadhi</i>	

The predictability of heat-related mortality in Prague, Czech Republic during summer 2015 – A comparison of selected thermal indices	1111
<i>Aleš Urban, David M. Hondula, Hana Hanzlíková and Jan Kysely</i>	
Power laws in stock market and fractal complexity of S&P500 and DAX.....	1113
<i>Anna Krakovská</i>	
Selection of Geographical Factors Using the Random Forest Analysis Method for Developing Site Index of <i>Pinus densiflora</i> stands in Republic of Korea	1125
<i>Hee-Jung Park, Se-Ik Park, Hyun-Soo Kim, Eun-Seong Lee, Hyun-Jun Kim and Sang-Hyun Lee</i>	
The Non-Stationary Unconstrained BINAR(1) Process with Geometric Marginals.....	1135
<i>Yuvraj Sunecher, Vandna Jowaheer, Naushad Mamode Khan, Isven Veerasawmy and Azmi Muslun</i>	
Characterising Dependency in Computer Networks Using Spectral Coherence.....	1147
<i>Alexander Gibberd, Jordan Noble and Edward Cohen</i>	
Time Series Analysis as a Powerful Tool in Space Weather Event Studies.....	1158
<i>Agnieszka Gil-Swidarska</i>	
The Utility of POI Data for Crime Prediction	1166
<i>Pawel Cichosz, Zbigniew Wawrzyniak, Radoslaw Pytlak, Grzegorz Borowik, Eliza Szczechla, Pawel Michalak, Dobieslaw Ircha, Wojciech Olszewski and Emilian Perkowski</i>	
Hawkes processes for credit indices time series analysis: How random are trades arrival times?	1178
<i>Achraf Bahamou, Maud Doumergue and Philippe Donnat</i>	
Tests for Segmented Cointegration: An Application to US Governments Budgets	1193
<i>Paulo Rodrigues and Luis Martins</i>	
One-pass incremental-Learning of temporal patterns with a bounded memory constraint ..	1253
<i>Koki Ando and Koichiro Yamauchi</i>	
Nonlinear relationship detection using pseudocorrelation.....	1265
<i>Jozef Jakubík</i>	
Automatic detection of sleep disorders: Multi-class automatic classification algorithms based on Support Vector Machines	1270
<i>David López-García, María Ruz, Javier Ramírez Pérez de Inestrosa and Juan Manuel Górriz Sáez</i>	
Relevance of Filter-Banked Features using Multiple Kernel Learning for Brain Computer Interfaces	1281
<i>Daniel Guillermo García-Murillo, David Cárdenas-Peña and German Castellanos-Domínguez</i>	
Multiple Instance Learning Selecting Time-Frequency Features for Brain Computing Interfaces	1291
<i>Julian Camilo Caicedo Acosta, Luisa Fernanda Velasquez-Martinez, David Cardenas-Peña and German Castellanos-Domínguez</i>	

Event Study in Tehran Stock Exchange: Central Bank Intervention and Market Impact Reaction	1300
<i>Gholamreza Keshavarz Haddad and Hadi Heidari</i>	
Influence of time-series extraction on binge drinking interpretability using functional connectivity analysis	1308
<i>Jorge Ivan Padilla Buritica and Cesar German Castellanos Dominguez</i>	
MoCap multichannel time series representation and relevance analysis by kernel adaptive filtering and multikernel learning oriented to action recognition tasks	1316
<i>Juan Diego Pulgarin-Giraldo, Andres Marino Alvarez-Meza, Steven Van Vaerenbergh, Ignacio Santamaría and German Castellanos</i>	
Forecast Model for Current, Wave and Wind Climate at the Danish Test Site for Wave Energy, DanWEC	1328
<i>Amélie Têtu</i>	
Density forecast comparison for disaggregated macroeconomic random variables using bayesian VAR models, bayesian global VAR models and large bayesian VAR models with stochastic volatility	1340
<i>Roberto Arsenal and Miguel Ángel Gómez Villegas</i>	
Simple estimators for higher-order stochastic volatility models and forecasting	1342
<i>Md Nazmul Ahsan and Jean-Marie Dufour</i>	
Entropy-based Channel Selection using Supervised Temporal Patterns in MI Tasks	1344
<i>Luisa Velasquez, Frank Zapata, David Cardenas and German Castellanos</i>	
<hr/> Applications of time series for hydro-climatic data. Complex/Big Data. <hr/>	
Maximum Entropy Methodologies in Large-Scale Data	1355
<i>Maria Da Conceição Costa and Pedro Macedo</i>	
Forecasting time series using topological data analysis	1367
<i>Nailia Gabdrakhmanova</i>	
Forecasting Subtidal Water Levels and Currents in Estuaries. Assessment of Management Scenarios	1374
<i>Miguel Ángel Reyes Merlo, María De Los Reyes Siles Ajamil and Manuel Díez Minguito</i>	
Nonstationary time series forecasting of wind and waves, combining hindcast, measured and satellite data	1385
<i>Christos Stefanakos</i>	
Spatial distribution of climatic cycles in Andalusia (southern Spain)	1406
<i>José Sánchez-Morales, Eulogio Pardo-Igúzquiza and Francisco Javier Rodríguez-Tovar</i>	
<hr/> Applications in Time Series (Part. III) <hr/>	
Real time anomaly detection in network traffic time series	1417
<i>Sergio Martinez Tagliafico, Gastón Garcia González, Alicia Fernández, Gabriel Gómez Sena and José Acuña</i>	

Spacecraft Mission Control Center Resource State Estimation and Contingency Forecasting	1429
<i>Natalia Bakhtadze, Denis Elpashev, Alexey Lototsky, Vladimir Lototsky and Eddy Zakharov</i>	
Towards Hybrid Prediction over Time Series with Non-Periodic External Factors	1431
<i>Xavier Fontes and Daniel Silva</i>	
A Forecasting Methodology based on growth models, for assessing performance: Application on the Moroccan Railway.....	1443
<i>Karima Selmani Bouayoune</i>	
Pereira Market Scan	1451
<i>Leandro Pereira, Carlos Jerónimo and José Santos</i>	
Forecasting health of complex IT systems using system log data	1460
<i>Shivshanker Singh Patel</i>	
<hr/> Forecasting Complex/Big data (Part.II) <hr/>	
Comparing linear and non-linear dynamic factor models for large macroeconomic datasets	1468
<i>Alessandro Giovannelli and Marina Khoroshiltseva</i>	
Simultaneous Multi-Response Multi-Covariate Best Subset Selection- with application to fault modelling.....	1469
<i>Aaron Lowther, Matt Nunes, Paul Fearnhead and Kjeld Jensen</i>	
A comparison of statistical methods for estimating individual location densities from smartphone data.....	1471
<i>Francesco Finazzi and Lucia Paci</i>	
<hr/> Financial Forecasting and Risk Analysis <hr/>	
Forecasting of Multiple Yield Curves Based on Machine Learning.....	1483
<i>Eva Lütkebohmert, Christoph Gerhart and Marc Weber</i>	
Empirical evaluation of advanced oversampling methods for improving bankruptcy prediction.....	1495
<i>Wedyan Alswiti, Hossam Faris, Huthaifa Aljawazneh, Salah Al-Deen Safi, Pedro Castillo Valdivieso, Antonio Mora García, Ruba Abukhurma and Hamad Alsawalqah</i>	
The changing shape of sovereign default intensities	1507
<i>Yusho Kagraoka and Zakaria Moussa</i>	
<hr/> Vector processes in Time Series <hr/>	
PoARX models for count time series	1519
<i>Jamie Halliday and Georgi Boshnakov</i>	
Gaussian Variational Bayes Kalman Filtering for Dynamic Sparse Bayesian Learning	1531
<i>Christo Kurisummoottil Thomas and Dirk Slock</i>	
<hr/> Nonparametric and Functional Methods in Time Series <hr/>	

A geometric proxy of economic uncertainty based on the disagreement in survey expectations	1543
<i>Oscar Claveria, Enrique Monte-Moreno and Salvador Torra Porras</i>	
Prediction of crime from time series data-driven model	1554
<i>Grzegorz Borowik, Zbigniew Wawrzyniak, Pawel Cichosz, Radoslaw Pytlak, Eliza Szczechla, Pawel Michalak, Dobieslaw Ircha and Wojciech Olszewski</i>	
Measurement and Modelling of Business Cycles using Linear and Nonlinear methods	1565
<i>Nomeda Bratčikovié</i>	

Advanced in Time Series and Forecasting. Virtual Presentation

Examination of forecasting in education field	1574
<i>Wafa Terouzi, Fatima Zahra Mahjoubi and Abdel Khalek Oussama</i>	
Time Series Versus Causal Forecasting: An Application of Artificial Neural Networks	1576
<i>Prithviraj Lakkakula</i>	
A value-based evaluation methodology for renewable energy supply prediction	1589
<i>Robert Ulbricht, Bijay Neupane, Martin Hahmann and Wolfgang Lehner</i>	
Analysis of Terrestrial Water Storage Variations on the Terrain of Vistula and Odra Basins in Poland	1601
<i>Zofia Rzepecka</i>	
Fourier Analysis of Cerebral Metabolism of Glucose: Gender Differences in Mechanisms of Colour Processing in the Ventral and Dorsal Streams in Mice	1612
<i>Philip Njemanze, Mathias Kranz and Peter Brust</i>	
NIST tests versus bifurcation diagrams and Lyapunov exponents when evaluating chaos-based pRNGs	1640
<i>Octaviana Datcu and Radu Hobincu</i>	
Risk Assessment Approach to Support IT Collaboration Network	1650
<i>Dikra Chikhaoui, Mohammed Salim Benqatla and Bouchaib Bounabat</i>	
Enhancing Stock Index Forecasting With Ensemble-based Techniques	1660
<i>Dhanya Jothimani and Surendra S. Yadav</i>	
GARCH-VMD Based Forecasting for Volatile Time Series of Indian Small Car Sales	1670
<i>Rajeev Pandey</i>	
Solar Irradiance forecasting of Ahmedabad based on Ant Colony Optimization and Neural Network	1680
<i>Md. Janibul Alam Soeb, Md. Irfanul Hasan and Md. Shahid Iqbal</i>	
Determination of energy losses in distribution transformers using a compensation algorithm in energy meters	1693
<i>Marco Toledo, Carlos Alvarez Bel, Paul Cando, Juan Maldonado, Pablo Méndez and Diego Morales</i>	
Oil Flow Rate Forecasting For Wells Drilled in Unconventional Reservoirs	1703
<i>Umer Farooq, Randy Hazlett and Krishna Babu</i>	

Predictive model of the techno-environmental performance of novel multi-function window combined ventilation system and solar photovoltaic blind using finite element method	1720
<i>Taecheon Hong, Jongbaek An, Jeongyoon Oh and Minhyun Lee</i>	
Establishment of operational strategy of the ventilation system in a building by considering the indoor and outdoor concentration of fine dust	1724
<i>Taecheon Hong, Jeongyoon Oh, Woojin Jung and Hakpyeong Kim</i>	
Analysis of interchannel phase connectivity for EEG event-related potentials using auditory oddball paradigm in attention tasks	1728
<i>Juana Valeria Hurtado, Juan David Martinez, Germán Castellanos, Francia Restrepo and Jorge Iván Padilla</i>	
Big-Learn 2+: Integrating Apache Spark with Solr Framework to improve the online search in Big Data environment	1738
<i>Karim Aoulad Abdelouarit, Boubker Sbihi and Noura Aknin</i>	

Time Series Analysis as a Powerful Tool in Space Weather Event Studies

Agnieszka Gil-Swidarska¹

Siedlce University, Faculty of Sciences, Institute of Mathematics and Physics, Siedlce, Poland

gila@uph.edu.pl

Abstract. Sun is central star of Solar System. Its activity grows nearly every 11 years. When Sun reaches activity maximum then can distribute into interplanetary space huge amount of energy and matter. These may impact our planet. We can observe it as beautiful Northern Lights (in Spain and Poland, and other sites) during the Halloween Storm in October-November 2003. However, they also escalate likelihood of geomagnetically induced currents occurrence and radio signals disturbances, such as the event on November 4, 2015 when Swedish airports were paralyzed. Scientists try to anticipate these events analyzing solar and geomagnetic activity parameters. Geomagnetic storms predictions are in the interest of many branches, among them: airlines, power systems operators, etc. Here we analyze various time series: solar and heliospheric parameters combined with geomagnetic indexes, around periods of extraordinary solar activity using wavelet-coherence. Our calculations show that the coherence between data changes its character around those intervals.

Keywords: wavelet coherence, correlation, space weather

1 Introduction

Curiosity is imprinted in the Humankind. We all want to know more and more. We intend to understand and foresee the world and mechanisms which regulate it. Nowadays, time series analysis methods are growing their impact and importance, because they help to recognize and, what is more, to predict some features of fast changing world.

One of the issues, which scientist need to forecast are space weather events. As a space weather we understand the whole set of changes occurring in the interplanetary space and in geosphere having a solar origin (e.g [1, 2]). Those changes are brought in to the heliosphere by the solar wind with frozen-in magnetic field [3]. Solar wind is a supersonic flow instantly emitted by the Sun.

One of a key parameter illustrating changes in the geomagnetic field is planetary magnetospheric Kp index. Planetary Kp index is a measure, in a quasi-logarithmic scale, geomagnetospheric conditions (globally averaged) recorded by thirteen magnetometers located at mid-latitudes (from 44 to 60 degrees). It

ranges from 0 to 9 (e.g., [4]). It is considered that geomagnetic storm starts from Kp=5. But the most interesting are strong (Kp=7), severe (Kp=8) and extreme (Kp=9) storms. The other important for space weather studies geomagnetic index is Dst index (disturbance storm time) measuring degree of a quasi-uniform magnetic field disturbances near the Earth (e.g., [5]). The next, used in this paper, geomagnetic activity parameter is the auroral electrojets (AE) index resulting from the north-south components of twelve magnetic northern-hemisphere stations, located around 70 degrees latitude (e.g., [6]).

Here we use a wavelet coherence in order to study properties of time intervals with strong disturbances occurring in the nearest Earth's vicinity caused by the solar activity during the solar cycle 23.

2 Data and Methods

We consider here hourly data covering the whole solar cycle 23, i.e. 108864 hours, from years 1996-2009. We analyse twelve various time series describing solar wind and heliospheric conditions: solar wind speed, density, pressure, the strength of the heliospheric magnetic field and its Bx, By, Bz components, electric field, and geomagnetic indexes: Kp, DST, AE. All of the considered data are available in the OMNI-2 database. Figure 1 displays four samples of the analysed data.

Most of the time series analysis methods indicate the changes' character of studied time series, but they do not show the relationships between the considered data in time-frequency space. As one of the useful measure of interdependence between data sets is their covariance, e.g. [7]. We study the crosswavelet spectrum of analyzed time series, geomagnetic indexes, as well as, heliospheric and solar wind parameters in the form [8]:

$$W_{x,y} = W_x W_y^*, \quad (1)$$

where x, y denote geomagnetic activity parameters and solar wind features, respectively, and * is a complex conjugation. The crosswavelet power, being a measure of common power, is $|W_{x,y}|$ and the phase angle designates the relationship of the phases between x and y, in time-frequency space [9]. The computations were carried out using the tools presented in [10, 9].

Subsequently, we consider the wavelet coherence [10], being a measure of the covariance intensity in time-frequency space (e.g. [11]). The coherence is defined as [12]:

$$R^2 = \frac{|S(W_{x,y}/s)|^2}{|S(W_x/s)|^2 \cdot |S(W_y/s)|^2}, \quad (2)$$

where s is the wavelet scale and an operator S is responsible for smoothing. The wavelet coherence (WTC) is treated as a kind of local correlation measure, between the two continuous wavelet transforms, allowing to find a substantial coherence even when the ordinary power is low [10]. Computations were performed by the tools introduced in [10].

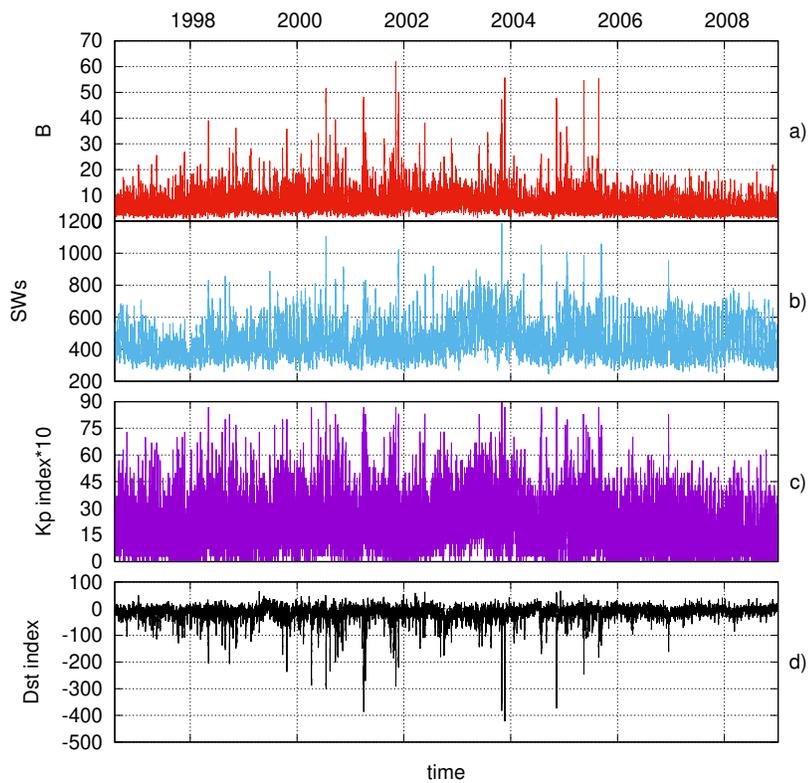


Fig. 1. Hourly data of a) strength of the heliospheric magnetic field (B [nT]), b) solar wind speed (SWs [km/s]), c) geomagnetic Kp index, d) geomagnetic Dst index [nT], in 1996-2009

3 Results and Discussion

During the solar activity cycle 23 there were noticed almost 200 events when 3-hour Kp index was not less than 7 (among them four 3-hour intervals of extreme geomagnetic storms and fifty four 3-hour intervals of severe geomagnetic storms). Figure 2 presents the same set of data as Fig. 1, but during only one week covering the end of October 2003 and the beginning of November 2003 (i.e. 168 hours), when the Halloween Storm started. One can observe that it was a period with drastic changes visible in all of the data. For this particular week we present the results of our calculation. Unfortunately, during around two days it was impossible to measure solar wind speed, pressure, density and temperature. If we consider the whole solar cycle 23 (108864 hours) the highest correlation coefficient is between solar wind speed and Kp index: 0.57 ± 0.01 , and also between the heliospheric magnetic field strength and Kp index: 0.55 ± 0.01 . When we consider only this particular week, it grows for the heliospheric magnetic field strength and Kp index up to: 0.75 ± 0.01 . Taking into account each day of this week separately, for the heliospheric magnetic field strength and Kp index maximum value of correlation coefficient was on 30 X 2003: 0.83 ± 0.02 and for the heliospheric magnetic field strength and Dst index the anti-correlation was even stronger on 31 X 2003: -0.97 ± 0.02 .

Figure 3a shows that during almost the whole week Kp index and heliospheric magnetic field were in-phase at the periodicities around one day with the strong common power being statistically significant (95 % significance marked with thin black curve in the Figure 3 a-d). At the time of the highest level of disturbances Dst index and heliospheric magnetic field were nearly perfectly in anti-phase (arrows are directed to the left, Figure 3b), with strong and statistically significant mutual power. Figure 3 shows that the interconnections between geomagnetic indexes and heliospheric magnetic field, in the range of short-term variabilities, are much more robust during the extremely active time interval (II) than during more quiet time (I, III). This is because the solar wind and heliospheric magnetic field were much more changeable in the second interval (II) comparing to the first (I) and the third (III).

For comparison we consider one week at the beginning of the solar cycle 23, 1–7 VIII 1996 (Fig. 4), during a quiet time, when $Kp < 4$. Correlations during this week were much weaker. The highest one was between Dst index and solar wind proton density having value only 0.51 ± 0.02 . During a quite week the common power is less extensive with a slightly less ordered phase difference (Fig. 5).

4 Summary

The first results presented in the Sect. 3 of this work show that both, wavelet coherence and cross-wavelet transform are useful tools for tracking relationships between parameters of heliospheric and geomagnetic conditions occurring during the time intervals of abrupt changes in the heliosphere. Though it is a beginning of our work and more studies need to be done.

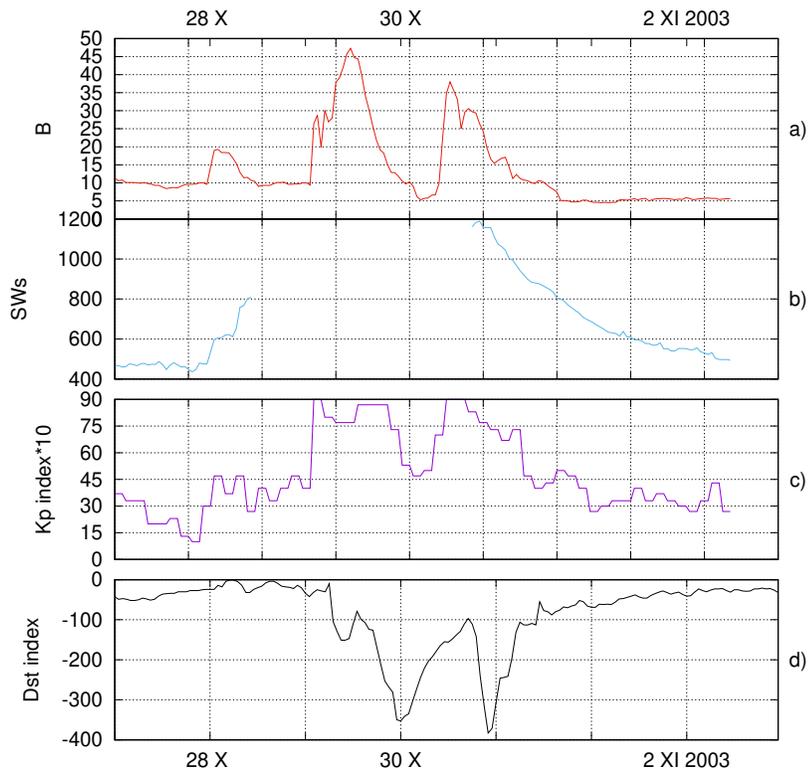


Fig. 2. Hourly data of a) strength of the heliospheric magnetic field (B [nT]), b) solar wind speed (SWs [km/s]), c) geomagnetic Kp index, d) geomagnetic Dst index [nT], during one week: 27 X–2 XI 2003

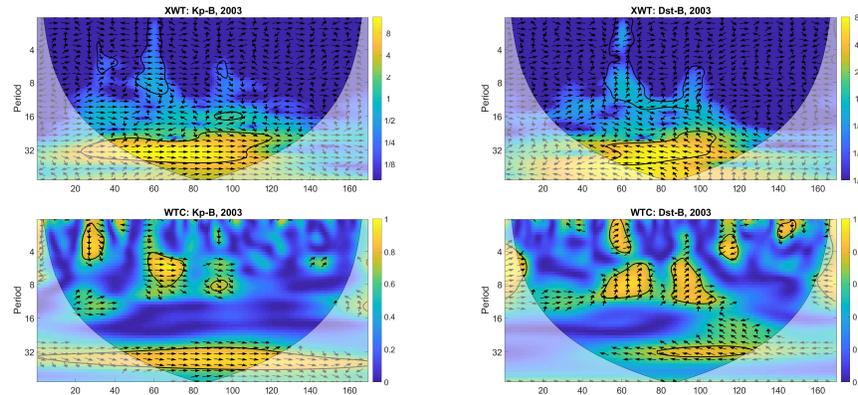


Fig. 3. Cross wavelet transform (a, b) and wavelet coherence (c, d) for geomagnetic indexes (Kp and Dst) and heliospheric magnetic field strength B. On the x-axis is time in hours during one week: 27 X–2 XI 2003. Arrows directed to the right mean that time series agree in phase and arrows directed to the left mean that series are in anti-phase

Acknowledgments

Measurements of heliospheric and solar variability parameters, as well as geomagnetic indexes are from <http://omniweb.gsfc.nasa.gov>. Cross-wavelet analysis tools: XWT & WTC by Grinsted et al. are available at <http://www.pol.ac.uk/home/research/waveletcoherence/>. We acknowledge the financial support by the Polish National Science Centre, decision number DEC-2016/22/E/HS5/00406.

References

1. Marubashi, K.: The Space Weather Forecast Program. *Space Science Reviews* 51, 197–214 (1989)
2. Riley, P., Baker, D., Liu, Y. D., Verronen, P., Singer, H., Gudel, M.: Extreme Space Weather Events: From Cradle to Grave. *Space Science Reviews* 214, article id. 21 (24 pp.) (2018)
3. Parker, E. N.: Dynamics of the Interplanetary Gas and Magnetic Fields. *Astrophysical Journal* 128, 664–676 (1958)
4. Maynard, N. C., Chen, A. J.: Isolated cold plasma regions: Observations and their relation to possible production mechanisms. *Journal of Geophysical Research* 80, 1009–1013 (1975)
5. Mayaud, P. N.: Derivation, Meaning, and Use of Geomagnetic Indices. AGU, *Geophysical Monograph* 22 (1980)
6. Pallochia, G., Amata, E., Consolini, G., Marcucci, M. F., Bertello, I.: AE index forecast at different time scales through an ANN algorithm based on L1 IMF and plasma measurements. *Journal of Atmospheric Solar Terrestrial Physics* 70, 663–668 (2008)

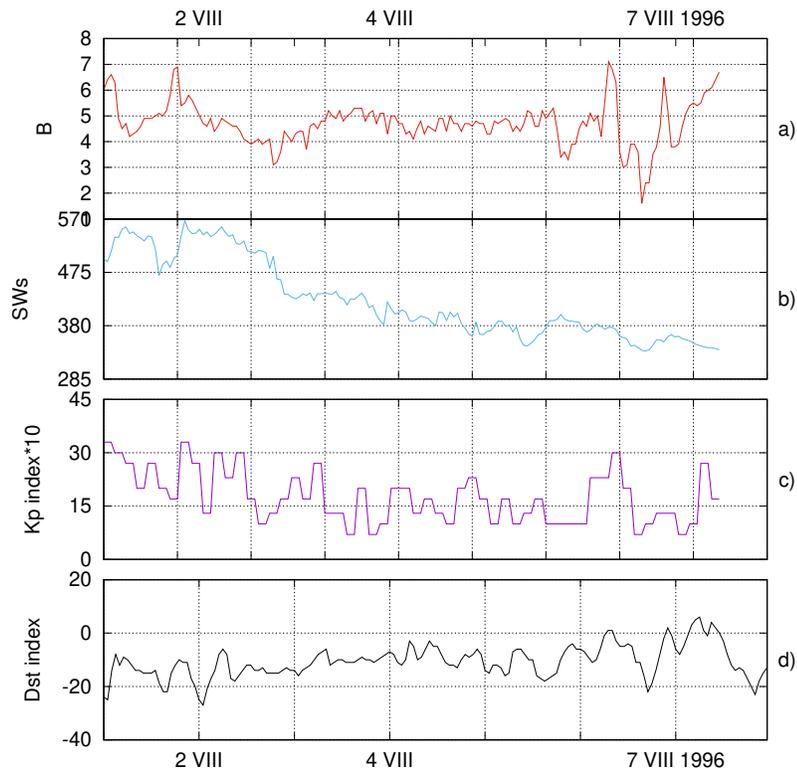


Fig. 4. Hourly data of a) strength of the heliospheric magnetic field (B [nT]), b) solar wind speed (SWs [km/s]), c) geomagnetic Kp index, d) geomagnetic Dst index [nT], during one week: 1–7 VIII 1996

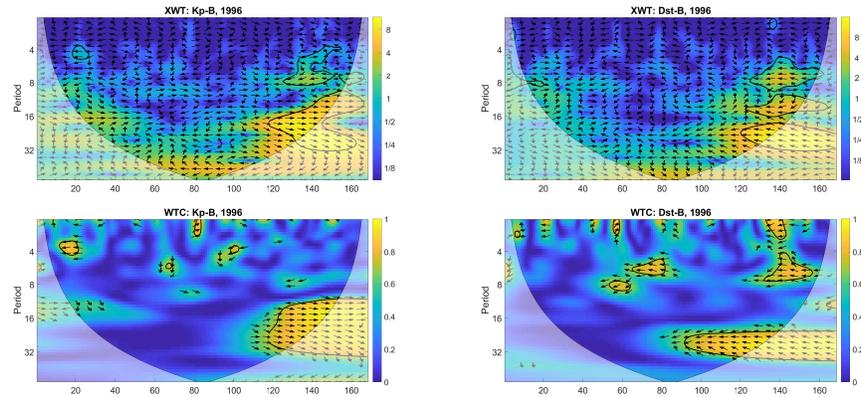


Fig. 5. Cross wavelet transform (a, b) and wavelet coherence (c, d) for geomagnetic indexes (Kp and Dst) and heliospheric magnetic field strength B. On the x-axis is time in hours during one week: 1–7 VIII 1996. Arrows directed to the right mean that time series agree in phase and arrows directed to the left mean that series are in anti-phase

7. Fuller, W. A.: Introduction to Statistical Time Series. John Wiley & Sons (2009)
8. Torrence, C., Compo, G. P.: A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79, 61–78 (1998)
9. Jevrejeva, S., Moore, J. C., Grinsted, A.: Influence of the Arctic Oscillation and El Niño-Southern Oscillation (ENSO) on ice conditions in the Baltic Sea: The wavelet approach. *Journal of Geophysical Research* 108, D21, 4677 (11 pp.) (2003)
10. Grinsted, A., Moore, J. C., Jevrejeva, S.: Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics* 11, 561–566 (2004)
11. Koopmans, L. H.: *The Spectral Analysis of Time Series: Probability and Mathematical Statistics*. Academic Press (2014)
12. Torrence, C., Webster, P. J.: Interdecadal changes in the ENSO-monsoon system. *Journal of Climate* 12, 2679–2690 (1999)

The Utility of Point of Interest Data for Crime Risk Prediction*

Paweł Cichosz¹, Zbigniew M. Wawrzyniak¹, Radosław Pytlak¹,
Grzegorz Borowik², Eliza Szczechła³, Paweł Michalak³, Dobiesław Ircha³,
Wojciech Olszewski³, and Emilian Perkowski³

¹ Warsaw University of Technology, Warsaw, Poland
p.cichosz@elka.pw.edu.pl, z.wawrzyniak@ise.pw.edu.pl,
r.pytlak@mini.pw.edu.pl

² Police Academy in Szczytno, Szczytno, Poland
g.borowik@wspol.edu.pl

³ Scott Tiger SA, Warsaw, Poland
eliza.szczechla@tiger.com.pl, pawel.michalak@tiger.com.pl,
dobieslaw.ircha@tiger.com.pl, wojciech.olszewski@tiger.com.pl,
emilian.perkowski@tiger.com.pl

Abstract. This paper examines the utility of Point of Interest (POI) data for learning crime prediction models. Crime event locations for a Polish city are aggregated into a grid and merged with selected data layers from *OpenStreetMap*. The resulting dataset contains crime count attributes and POI count attributes for particular areas of the city, obtained by aggregating over a rectangular grid. After identifying high-risk areas based on crime counts, POI count attributes are used for learning crime risk prediction models with the logistic regression, support vector machines, and random forest algorithms. The experimental results suggest that POI attributes have high predictive utility. Classification models using these attributes, without any form of location identification, exhibit very good predictive performance, which makes it possible to reuse them over different cities.

1 Introduction

Crime prediction using analytic algorithms from machine learning and statistics has recently become a popular area of research and practical applications [8, 2, 4, 27, 19, 17]. The potential contribution of achievements in this direction to the public security provides high motivation for applying a variety of methodologies to various kinds of potentially useful data. Besides police crime records, these may include, e.g., socio-demographic and economic statistics, or social media posts.

In this paper we examine the utility of Point of Interest (POI) locations for predicting crime hotspots – city areas with high crime risk. The capability

* Supported by the Polish National Center for Research and Development under grant DOB-BIO7/05/02/2015.

to predict which areas have particularly high risk of specific types of criminal events based on the number and categories of nearby points of interest may help to allocate limited law enforcement resources where they are most needed.

The assumed model creation and prediction scenario can be summarized as follows:

hotspot identification: label city areas as high/low risk areas with respect to the number of historical crime events,

hotspot modeling: create a model for predicting high/low risk labels based on attributes derived from POI locations,

hotspot prediction: apply the model to predict high/low risk labels for particular city areas based on attributes derived from POI locations.

This uses historical data to create models capable of predicting the future crime risk. It is noteworthy that city areas are not represented by identifiers or by coordinates but described exclusively by attributes derived from POI locations. This makes it possible to capture more general and reusable relationships, independent of a particular city topography.

An experimental verification of the above scenario is performed using crime event records for a Polish city and selected *OpenStreetMap* data layers, supposed to answer the following questions:

- what is the utility of attributes derived from POI locations for hotspot prediction?
- what level of prediction quality can be achieved?
- which algorithms deliver the best predictive performance?

2 Data Preparation

The process of data preparation combines data from two sources: police crime records and *OpenStreetMap* point of interest locations.

2.1 Crime Data

Crime data used for this work were extracted from anonymized internal police records for Białystok, a city of about 300,000 inhabitants located in northeastern Poland. They cover the period from January 2013 to April 2017 and contain rounded geographical coordinates of 12 categories of events, including in particular:

- hooliganism,
- fight,
- theft,
- robbery.

Table 1. Crime counts for the selected four categories.

category	fight	hooliganism	robbery	theft
count	880	1291	126	3669

This study is limited to the four categories listed above. The overall number of events of these categories within the area of Białystok is presented in Table 1. Their locations on the city map are presented in Figure 1.

For the purpose of simple hotspot identification, the city area was partitioned into a grid of rectangles with dimensions of 300×300 meters, as illustrated in Figure 2, which uses a grey shade scale to visualize the overall crime event count in particular rectangles, restricted to the selected most interesting categories. Average daily crime counts for particular categories were then converted to binary high/low risk labels using the 3rd quartile as the cutoff point (high risk if the event count is above the 3rd quartile, low risk otherwise). The resulting binary risk indicators serve as target attributes for prediction.

2.2 POI Data

Point of interest data was obtained as shapefiles with selected *OpenStreetMap* layer extracts, available from *Geofabrik.de* [26]. POI shapefile loading and pre-processing was performed using the `rgdal` [3] and `sp` [20] R packages. The following layers were used:

pois: POI objects represented as points,
pois_a: POI objects represented as polygons,
transport: transport objects represented as points,
transport_a: transport objects represented as polygons.

They contain a total of 4632 objects within the city area, divided into 115 categories. They were aggregated using the same 300×300 grid as crime events, and the corresponding object counts serve as input attributes for prediction.

2.3 Combined Data

The combined dataset was obtained by joining the crime data and POI data by grid cells. It contains rows corresponding to grid cells and columns corresponding to risk indicators and POI attributes.

3 Algorithms

An arbitrary classification algorithm can be used to predict high/low risk labels based on POI attributes. A selection of the most useful algorithms known from the literature is applied in this work: logistic regression, support vector machines, decision trees, and random forests [9].

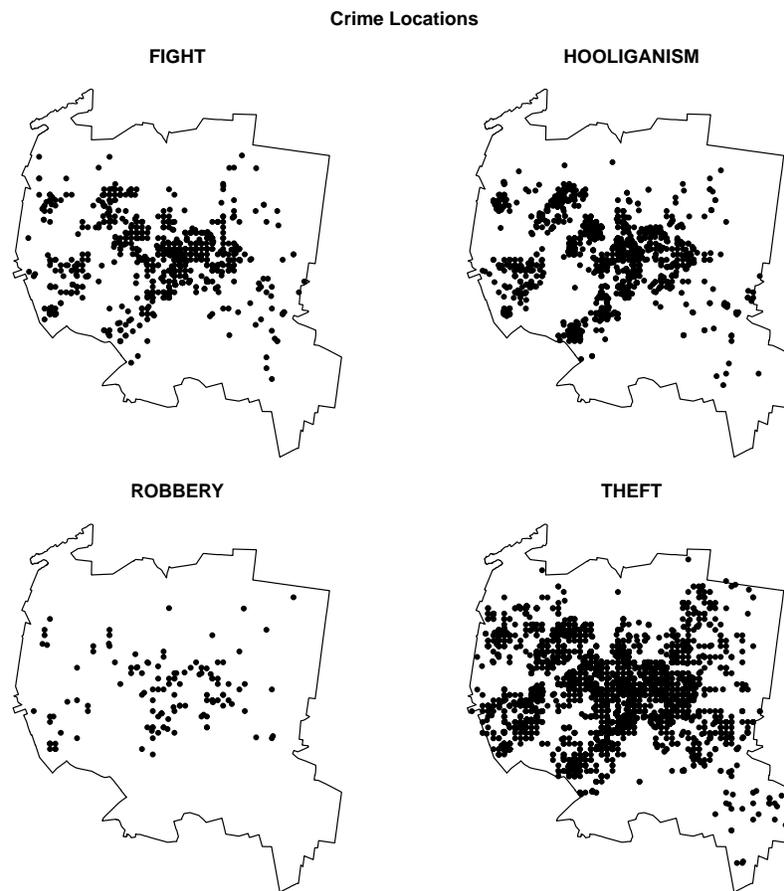


Fig. 1. Crime locations for the selected four categories.

3.1 Logistic Regression

Logistic regression is an instantiation of generalized linear models which adopts a composite model representation function, with an inner linear model and an outer logit transformation [15]. Training a logistic regression model consists in finding model parameters which maximize the log-likelihood of training set classes.

Due to the probabilistic objective function used for parameter estimation, logistic regression can generate well-calibrated probability predictions and is often the classification algorithm of choice where this is required. It is easy to apply and not overly prone to overfitting unless used for high-dimensional data. In our experiments, logistic regression serves as a natural comparison baseline for the more refined support vector machines algorithm which extends linear

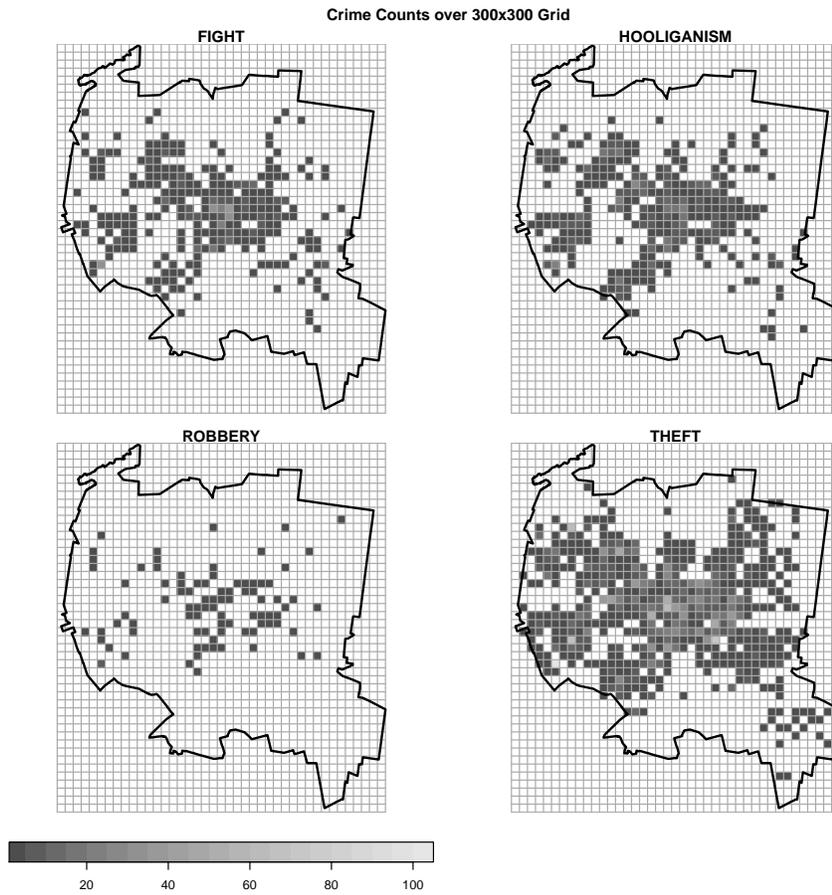


Fig. 2. Crime counts over 300×300 -meter grid for the selected four categories.

classification, achieving better overfitting resistance and permitting nonlinear relationships.

3.2 Support Vector Machines

Support Vector Machines (SVM), which often belong to the most effective general-purpose classification algorithms, can be viewed as a considerably strengthened version of a basic linear-threshold classifier with the following enhancements [10, 22, 14]:

- margin maximization:** the location of the decision boundary (separating hyperplane) is optimized with respect to the classification margin,
- soft margin:** incorrectly separated instances are permitted,

kernel trick: complex nonlinear relationships can be represented by representation transformation using kernel functions.

Instead of binary linear-threshold SVM predictions it may be often more convenient to use probabilistic predictions. This is possible by applying a logistic transformation to the signed distance of classified instances from the decision boundary, with parameters adjusted for maximum likelihood [23].

A noteworthy property of SVM is the insensitivity of model quality to data dimensionality, which – unlike for many other algorithms – does not increase the risk of overfitting because model complexity is related to the number of instances close to the decision boundary rather than to the number of attributes.

3.3 Decision Trees

A decision tree [7, 24] is a hierarchical structure that represents a classification model, i.e., a mapping of instances from a given domain to a finite set of classes. Internal tree nodes represent splits applied to decompose the domain into regions, and terminal nodes assign class labels or probabilities to regions believed to be sufficiently small or sufficiently uniform.

Decision trees are popular in many applications due to their capability of combining reasonably good prediction accuracy with the human readability of models. They may require appropriately tuned stop criteria or pruning to avoid overfitting. In our experiments, decision trees serve as a natural comparison baseline for the more refined random forest algorithm which combines multiple trees to achieve better prediction quality and overfitting resistance.

3.4 Random Forest

Random forests belong to the most popular ensemble modeling [11] algorithms, which achieve improved predictive performance by combining multiple diverse models for the same domain. A random forest [6] is an ensemble model represented by a set of unpruned decision trees, grown based on multiple bootstrap samples drawn with replacement from the training set, with randomized split selection. It can be considered an enhanced form of bagging [5], which additionally stimulates the diversity of individual models in the ensemble by randomizing the decision tree growing algorithm used to create them.

Random forest prediction is achieved by simple unweighted voting of individual trees from the model. Vote distribution can be also used to obtain class probability predictions. With sufficiently many diversified trees (typically hundreds) this simple voting mechanism usually makes random forests extremely accurate and resistant to overfitting. As a matter of fact, in many cases they belong to the most accurate classification models that can be achieved.

An additional capability of the random forest algorithm is providing measures of attribute predictive utility, referred to as variable importance. The most reliable of those is based on the decrease of prediction accuracy resulting from random attribute value permutation, estimated using out-of-bag training instances [28].

4 Experiments

The experimental evaluation of the utility of POI data for crime risk prediction is based on the assessment of the quality of classification models which predict risk indicators for the selected crime categories based on POI attributes.

4.1 Predictive Performance Evaluation

The most common classification quality measures such as the misclassification error or classification accuracy are not very useful whenever classes are unbalanced or likely to have different predictability. They also do not adequately capture the predictive power of probabilistic models which can be used in various operating points, corresponding to different probability cutoff values. This is why in the experiments reported in this section classification quality is visualized using ROC curves, presenting possible tradeoff points between the true positive rate and the false positive rate [12, 13], and summarized using the area under the ROC curve (AUC). To achieve reliable, low-bias and low-variance predictive performance estimates, the 10-fold cross-validation procedure repeated 10 times is applied [1].

4.2 Algorithm Implementations and Setup

The following algorithm implementations are used in the experiments:

logistic regression: the implementation provided by the standard `glm` R function [25],

SVM: the implementation provided by the `e1071` R package [18],

decision trees: the implementation provided by the `rpart` R package [29].

random forest: the implementation provided by the `randomForest` R package [16].

For the logistic regression and SVM algorithms parameters controlling the underlying optimization process were left at default values. The SVM parameters specifying the optimization problem were set as follows:

the cost of constraint violation (`cost`): 1,

the kernel type (`kernel`): `radial`,

the kernel parameter (`gamma`): the inverse of the number of attributes (input dimensionality),

class weights in constraint violation penalty (`class.weights`): 3 for the high risk class, 1 for the low risk class.

For the decision tree algorithm, the default stop criteria were modified to restrict the maximum tree depth (i.e., the number of split levels) to 3 using the `maxdepth` parameter. This is supposed to prevent overfitting and verify whether the relationships between crime risk and POI attributes are simple enough to enable successful prediction with small trees. Uniform prior probabilities for the two classes were set via the `prior` parameter.

For the random forest algorithm, the following setup was used:

- the number of trees (ntree):** 500,
- the number of attributes for split selection at each node (ntry):** the square root of the total number of available attributes,
- the stratified bootstrap sample size (sampsize):** the number of instances of the high risk class (the minority class).

It is worthwhile to notice that the parameter setups for the SVM, decision tree, and random forest algorithm include settings responsible for properly handling unbalanced classes (ensuring sufficient sensitivity to the minority class). This is achieved by specifying class weights for SVM (assigning a higher weight to the minority class when calculating the constraint violation penalty term in the optimization objective), setting uniform class priors for decision trees, and specifying stratified bootstrap sample size for the random forest algorithm (drawing the maximum possible number of minority class instances and the same number of the majority class instances). These settings were verified to indeed improve model quality. No form of class rebalancing is necessary for the logistic regression algorithm, since any class weights or priors would only shift the default class probability cutoff point used for predicted class label assignment. This would serve no useful purpose given the fact that the ROC analysis used for predictive performance evaluation is based on predicted class probabilities instead of class labels anyway.

4.3 Prediction Quality

The ROC curves visualizing the prediction quality are presented in Figure 3. The following observations can be made:

- high quality risk prediction for each of the four selected crime categories is possible using POI attributes, with AUC values above 0.85 (and true positive rate of about 0.8 possible with false positive rate of about 0.2 or less),
- the random forest algorithm achieves the best predictive performance regardless of the predicted crime category,
- the SVM algorithm is similarly successful except for the least frequent *robbery* category,
- logistic regression and decision tree models give clearly inferior prediction quality, which suggests that the relationship between crime risk and POI attributes is not simple enough to be adequately represented by a linear function or a small tree.

4.4 Attribute Predictive Utility

Figure 4 presents the variable plots, using the mean decrease of accuracy measure. For each crime category the 30 most useful POI attributes are shown. The rankings of predictive utility do not vary substantially across crime categories. The following are always among the most predictively useful POI objects: *playground*, *convenience*, *pitch*, *restaurant*, *kindergarten*, *school*, *camera_surveillance*, *supermarket*, *pharmacy*, and *atm*.

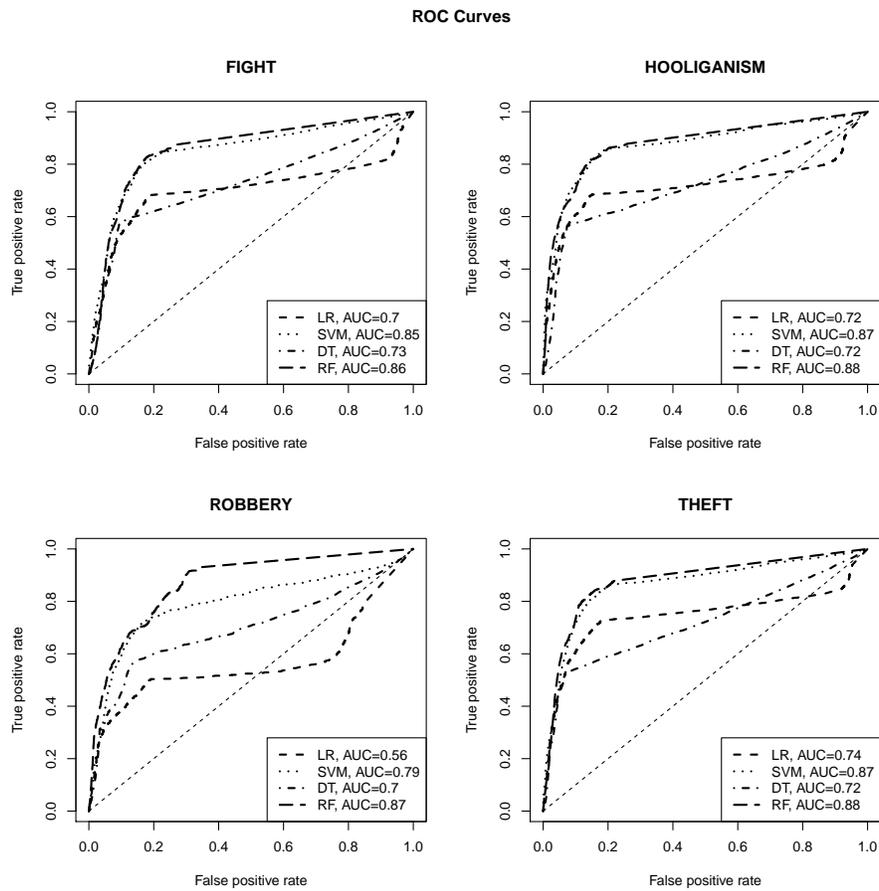


Fig. 3. The ROC curves for crime risk prediction.

5 Conclusions

This paper experimentally investigated the utility of point of interest data for crime risk prediction. The study was performed using police crime records for a mid-size Polish city and *OpenStreetMap* POI layers. A simple grid aggregation method for combining geotagged crime events with point of interest locations was applied. The results suggest that POI attributes are highly useful for crime prediction, making it possible to accurately discriminate between high-risk and low-risk areas. The patterns of relationship between crime risk and POI attributes are not trivial, though, since logistic regression and simple decision trees are outperformed by SVM and random forest models by a substantial margin. High-quality models for crime hotspot prediction based on POI attributes may become essential components of predictive policing systems, allowing law en-

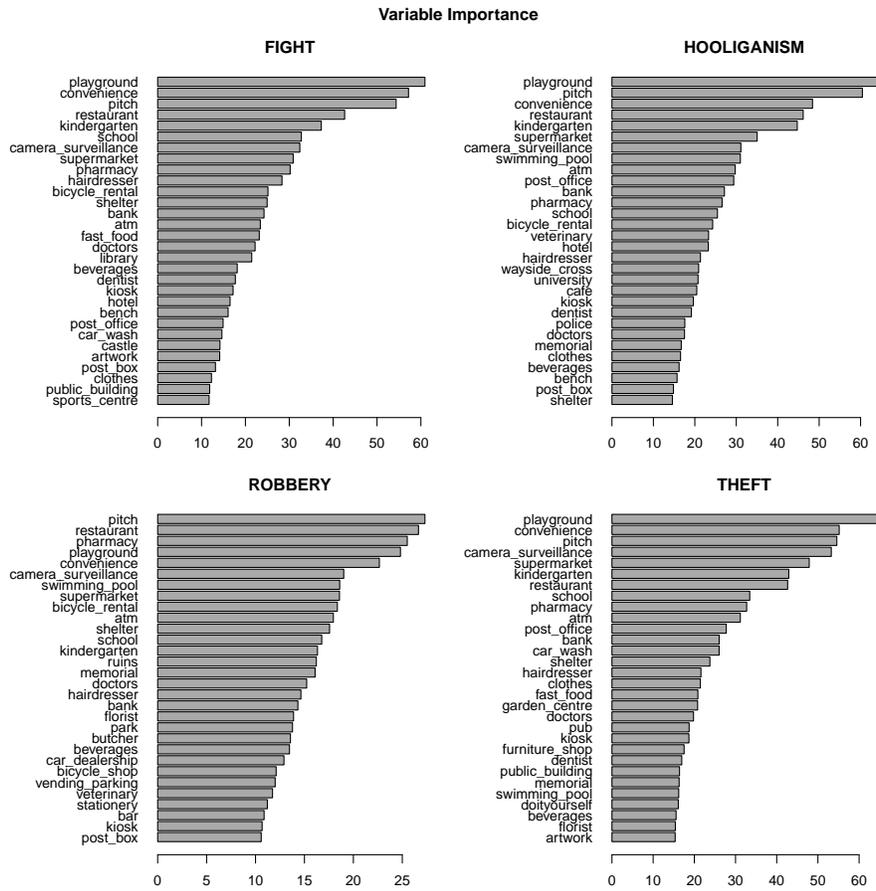


Fig. 4. The variable importance plots for crime risk prediction.

forcement agencies to more effectively prevent crime using constrained resources [21].

It is worthwhile to underline that crime risk predictions are made using per-category POI counts as the only type of area description, without any area identifiers or coordinates. The created models capture relationships between crime risk and POI density observed in historical data and use them to predict future crime risk, without being tied to a particular city topography.

Further work is needed to verify the usefulness of POI attributes for crime prediction for multiple diverse cities. It would be also interesting to examine the impact of grid resolution on prediction quality and the utility of more refined methods of spatial aggregation. It similarly remains to be verified whether and to what extent POI-based crime risk prediction models can be transferred between cities (i.e., whether and how well a model trained using crime data and POI

data for one city can serve for crime prediction in another city). Finally, by comparing crime records and predictions obtained using time-tagged POI data (e.g., from map extracts collected over a period of several months or years) it may be possible to observe how new city district development reflected by changing POI counts corresponds to emerging new risk areas.

The promising results obtained with classification models provide an encouragement to consider other types of modeling using crime records and POI attributes, such as regression for crime count prediction or clustering for identifying similarity patterns city with respect to POI and crime occurrence. It would be also worthwhile to verify the utility of other data sources from which attributes describing city areas can be derived, such as geotagged social media data and mobile network data.

While this research only considers the relationship of crime risk to area attributes, a natural extension would be to additionally consider attributes describing time. These could include, in particular, hour, day of week, month of year, as well as attributes describing actual or forecasted weather conditions. As our recent preliminary investigation suggests, such attributes may have high predictive utility, and combined with POI attributes might enable practically useful spatio-temporal crime risk prediction [30].

References

1. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**, 40–79 (2010)
2. Bernasco W., N.P.: How do residential burglars select target areas? a new approach to the analysis of criminal location choice. *The British Journal of Criminology* **45**, 296–315 (2005)
3. Bivand, R., Keitt, T., Rowlingson, B.: *rgdal: Bindings for the 'Geospatial' Data Abstraction Library* (2017), <https://CRAN.R-project.org/package=rgdal>
4. Bowers, K.J., Johnson, S.D., Pease, K.: Prospective hot-spotting: The future of crime mapping? *The British Journal of Criminology* **44**, 641–658 (2004)
5. Breiman, L.: Bagging predictors. *Machine Learning* **24**, 123–140 (1996)
6. Breiman, L.: Random forests. *Machine Learning* **45**, 5–32 (2001)
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Chapman and Hall (1984)
8. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: A general framework and some examples. *Computer* **37**, 50–56 (2004)
9. Cichosz, P.: *Data Mining Algorithms: Explained Using R*. Wiley (2015)
10. Cortes, C., Vapnik, V.N.: Support-vector networks. *Machine Learning* **20**, 273–297 (1995)
11. Dietterich, T.G.: Ensemble methods in machine learning. In: *Proceedings of the First International Workshop on Multiple Classifier Systems*. Springer (2000)
12. Egan, J.P.: *Signal Detection Theory and ROC Analysis*. Academic Press (1975)
13. Fawcett, T.: An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006)
14. Hamel, L.H.: *Knowledge Discovery with Support Vector Machines*. Wiley (2009)
15. Hilbe, J.M.: *Logistic Regression Models*. Chapman and Hall (2009)

16. Liaw, A., Wiener, M.: Classification and regression by randomForest. *R News* **2**(3), 18–22 (2002), <http://CRAN.R-project.org/doc/Rnews/>
17. Malleson, N., Heppenstall, A., See, L., Evans, A.: Using an agent-based crime simulation to predict the effects of urban regeneration on individual household burglary risk. *Environment and Planning B: Planning and Design* **40**, 405–426 (2013)
18. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2015), <https://CRAN.R-project.org/package=e1071>, R package version 1.6-7
19. Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. *Journal of the American Statistical Association* **106**, 100–108 (2011)
20. Pebesma, E.J., S., B.R.: Classes and methods for spatial data in R: The sp package. *R News* **5**, 9–13 (2005)
21. Perry, W.L., McInnis, B., Price, C.C., Smith, S., Hollywood, J.S.: The role of crime forecasting in law enforcement operations. Tech. Rep. RR-233-NJ, RAND Corporation (2013)
22. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) *Advances in Kernel Methods: Support Vector Learning*. MIT Press (1998)
23. Platt, J.C.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola, A.J., Barlett, P., Schölkopf, B., Schuurmans, D. (eds.) *Advances in Large Margin Classifiers*. MIT Press (2000)
24. Quinlan, J.R.: Induction of decision trees. *Machine Learning* **1**, 81–106 (1986)
25. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2016), <http://www.R-project.org>
26. Ramm, F.: OpenStreetMap data in layered GIS format (2015), access date: May 22, 2018
27. Short, M.B., DOrsogna, M.R., Brantingham, P.J., Tita, G.E.: Measuring and modeling repeat and near-repeat burglary effects. *Journal of Quantitative Criminology* **25**, 325–339 (2009)
28. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007)
29. Therneau, T., Atkinson, B., Ripley, B.: rpart: Recursive Partitioning and Regression Trees (2014), <http://CRAN.R-project.org/package=rpart>
30. Wawrzyniak, Z.M., Borowik, G., Szczechla, E., Michalak, P., Pytlak, R., Cichosz, P., Ircha, D., Olszewski, W., Perkowski, E.: Relationships between crime and everyday factors. In: *Proceedings of the Twenty-Second IEEE International Conference on Intelligent Engineering Systems (INES-2018)* (2018), to appear

Hawkes processes for credit indices time series analysis: How random are trades arrival times?

Achraf Bahamou, Maud Doumergue, and Philippe Donnat

Hellebore Capital Ltd
Michelin House, London

Abstract. Targeting a better understanding of credit market dynamics, the authors have studied a stochastic model named Hawkes process. Describing trades arrival times, this kind of model allows for the capture of self-excitement and mutual interactions phenomena. The authors propose here a simple yet conclusive method for fitting multidimensional Hawkes processes with exponential kernels, based on a maximum likelihood non-convex optimization. The method was successfully tested on simulated data, then used on new publicly available real trading data for three European credit indices, thus enabling quantification of self-excitement as well as volume impacts or cross indices influences.

Keywords: Point and counting processes, multidimensional Hawkes process, financial time series analysis, non-convex optimization, maximum likelihood optimization, credit indices.

1 Introduction

1.1 From credit derivative indices to Hawkes processes

Credit indices are financial instruments comprised of a set of credit securities, mainly used to hedge credit default risk. Each new index series is issued every 6 months and expires after a defined maturity. Though liquid, those indices are traded at a rather "mid-frequency" rate, with trades occurring at a minute scale. For European indices, the market is continuously open from 7:00 to 17:00, with a total of five to ten billion euros reported each day on average. Fig 1 gives an idea of the activity on a sampled day of trading. If recent regulations have led to a greater public reporting of trading activities, thus releasing more amount of traded data, this data is by essence quite sparse. Picturing such market behaviour over time represents a challenging opportunity: it originally motivated the work presented in this article.

This study focuses on three principal European credit indices, of the most traded 5-year maturity¹: the *Main Index*, noted here ITXEB, a combination of 125 equally weighted investment grade entities; the *Crossover Index*, noted ITXEX, with 75 sub-investment grade names; and the *Senior Financial*, noted ITXES, a subset of 30 financial entities from Main Index, referencing senior debt.

¹ Data provided by otcstreaming.com, and available on demand.

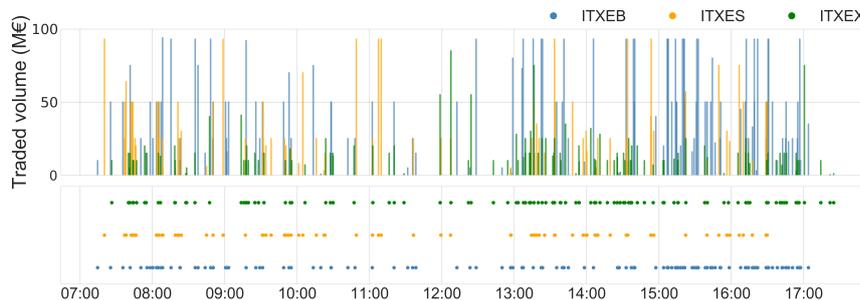


Fig. 1: Publicly reported trading activity on 22/05/2018 (*times, volumes*) for European credit indices ITXEB, ITXES & ITXEX.

When trying to capture the underlying dynamics of this market, one of the first questions that comes to mind is: how random is the timing of trades? A very naive approach is to model trades arrival times as purely random and uncorrelated processes, for example as a Poisson process by trying to fit an exponential law over the distribution of inter-arrival times. Yet, such a model fails to fit the data: from Fig 2, we can notice a high density of very short inter-arrival times, suggesting that some self-excitation phenomenon occurs, with trades triggering more trades; an intuition that every practitioner of the field would confirm.

	$\mathbb{E}(\Delta t)$	$\sigma(\Delta t)$	$Q_1(\Delta t)$	$Q_2(\Delta t)$	$Q_3(\Delta t)$	Number of trades
ITXEB	7 min	10 min	1 min	3 min	8 min	22 227
ITXEX	8 min	12 min	2 min	4 min	10 min	18 152
ITXES	17 min	24 min	3 min	9 min	21 min	7 194

Table 1: Statistics on inter-arrival times Δt (period: 03/01/2017 to 14/12/2017)

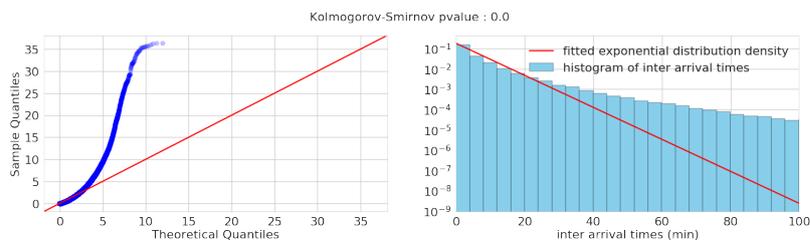


Fig. 2: ITXEB, from 15/01/2017 to 15/05/2018: *right*: log-scaled distribution of inter-arrival times; *left*: Q-Q plot. The *red lines* represent the best exponential law fitting. With a p-value of 0.0, this model is clearly inappropriate.

Literature review reveals that one model has recently driven considerable attention and a growing interest from both the scientific and quantitative communities: Hawkes processes. This model, introduced in 1971 by Hawkes ([7], [8])

to characterize earthquake tremors, can be used to describe timing of trades and their cross influences from a point process perspective.

After summarizing related work for Hawkes processes, we will recall the theoretical framework around this stochastic model. We will then focus on optimization methods for fitting such a process and describe Two Stage Hawkes Likelihood Optimization (2SHLO), a maximum likelihood based algorithm to fit multidimensional Hawkes processes with a parametric exponential kernel. Lastly, we will present our results, firstly on simulated data, then on credit derivative indices mentioned above. Hawkes processes in finance have mainly been applied to high frequency data. Is such a model appropriate for "mid-frequency" credit data? Can we specify the impact of volumes in trades or describe mutual influences between different indices?

1.2 Related work

Hawkes processes were originally designed for seismology analysis and deeply studied for that purpose by Ogota [13], [11], [12]. The concept has been utilized to model other effects where self and cross ignitions happen, such as social media tweets cascading [15], crime occurrences due to gang retaliations [10], or financial market events. As regards to financial applications, Bowsler [4] proposed in 2002 a generalized Hawkes process model taking into account night gaps, inter-day dependencies as well as intraday seasonality (with a piece-wise baseline intensity), which was used to model interactions between trades and price changes and also estimate of the price volatility based on mid-quote intensity for NYSE stock. Market price microstructure & market impact (influence of market orders on forthcoming prices) was modelled in [3], price impact was also studied in [1]. [14] focused on the impact of volume and order types on the limit order book. [2] (2015) exposes a full review of Hawkes process applications to finance, such as price & volatility modeling, market reflexivity measurement, order book modeling or risk contagion modeling.

2 Modeling Self and Cross Excitement with Hawkes Processes

This section re-frames the required formalism, as very clearly stated in [2], [9] or [16], starting from point and counting process definitions, through the key concept of intensity function. Hawkes model formulation is detailed for the reader, with a peculiar focus on exponential kernel structure.

2.1 Core concepts: from counting & point processes to intensity functions

Definition 1 (Point Process). *Let (Ω, \mathcal{F}, P) be a probability space. Let $(t_k)_{k \in \mathbb{N}^*}$ be a sequence of non-negative random variables such that $\forall k \in \mathbb{N}^*$, $t_k < t_{k+1}$. $(t_k)_{k \in \mathbb{N}^*}$ is called a (simple) point process on \mathbb{R}_+ .*

Definition 2 (Counting Process). Let $(t_k)_{k \in \mathbb{N}}$ be a point process. The right-continuous stochastic process defined for all $t \in \mathbb{R}_+$ as $N(t) = \sum_{k \in \mathbb{N}^*} 1_{t_k \leq t}$ is called the counting process associated with $(t_k)_{k \in \mathbb{N}^*}$.

The study of point processes goes through a single mathematical object, the *intensity function*, which is defined as the conditional probability density of occurrence of an event in the immediate future.

Definition 3 (Intensity Function). Let N be a counting process adapted to a filtration \mathcal{F}_t . The left-continuous intensity is heuristically defined as

$$\lambda(t|\mathcal{F}_t) = \lim_{h \rightarrow 0} \mathbb{E} \left[\frac{N(t+h) - N(t)}{h} \middle| \mathcal{F}_t \right] \quad (1)$$

The intensity function depends on the choice of filtration \mathcal{F}_t , which represents the amount of information available until time t (See [5] for a rigorous definition of the intensity function). In the context of Hawkes processes, we will simply use the natural filtration, with all previous information being available, and consider $\lambda(t)$.

Homogeneous Poisson process

One of the simplest point processes is the homogeneous Poisson process, for which the intensity function is constant over time: $\forall t \geq 0, \lambda(t) = \lambda$. In that case, durations (or inter-event waiting times) are independent and identically distributed (following an exponential distribution of hazard rate λ). As presented in the introduction, credit trades cannot be modeled by this memory-less model. Let's introduce Hawkes processes, a peculiar kind of non-homogeneous Poisson processes with linear dependencies over functions of past events.

2.2 Self-excitement: one-dimensional (or univariate) Hawkes processes

Definition 4 (One-dimensional Hawkes Processes).

Let $(t_k)_{k \in \mathbb{N}}$ be a point process and N the associated counting process, such that its intensity function λ is defined for each time $t \geq 0$ as

$$\lambda(t) = \mu(t) + \int_{-\infty}^t \phi(t - \tau) dN(\tau) = \mu(t) + \sum_{t_i < t} \phi(t - t_i) \quad (2)$$

where $\mu : \mathbb{R} \mapsto \mathbb{R}_+$ is an exogenous base intensity and $\phi : \mathbb{R}_+ \mapsto \mathbb{R}^+$ is a non-negative, measurable function such that $\|\phi\|_1 = \int_0^\infty \phi(s) ds < 1$.

N is called a Hawkes process with baseline μ and kernel ϕ .

The kernel ϕ expresses the positive influence of past events on the current value of the intensity. Each jump $dN(\tau) \neq 0$ increases the probability of future events through the kernel ϕ . Clustering effects and branching structure are well depicted by Hawkes processes, with the baseline activity generating immigrant events and descendant events enhancing the intensity.

For this study, we focus on *exponential kernels*, defined as $\phi(t) = \alpha\beta e^{-\beta t} \mathbf{1}_{t \geq 0}$, with parameters $\alpha, \beta \geq 0$, which allows easy interpretation: α or *adjacency* represents the weight of previous events while β is the *decay* / typical duration of influence for a past event.

The *branching ratio* $n = \|\phi\|_1 = \int_0^\infty \phi(s)ds$ represents the average number of descendants for any event. For exponential kernel, $n = \alpha$.

2.3 Including mutual-excitement with multidimensional Hawkes processes

Definition 5 (Multidimensional Hawkes Processes).

Let $M \in \mathbb{N}^*$ and $\{(t_k^i)_k\}_{i=1, \dots, M}$ be a M -dimensional point process.

We denote by $N_t = (N_t^1, \dots, N_t^M)$ the associated counting process such that its vector intensity function $\lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+^M$, is defined as, for all $t \geq 0, i \in 1, \dots, M$:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^M \int_0^t \phi_{i,j}(t - \tau) dN_j(\tau) = \mu_i + \sum_{j=1}^M \sum_{n=1}^{N_j(t)} \phi_{i,j}(t - t_{j,n}), \quad (3)$$

with $\mu = (\mu_i)_{i=1, \dots, M}$ an exogenous base intensity vector

and $\phi(t) = (\phi_{i,j})_{i,j=1, \dots, M}$ a matrix-valued kernel that is component-wise positive and causal (null values when $t < 0$) and with each component belonging to the space of L^1 -integrable functions.

N_t is called a multidimensional (or multivariate) Hawkes process with baseline μ and kernel ϕ .

The choice of an exponential kernel can be generalized for multivariate Hawkes processes, with kernel components $\phi_{i,j}(t) = \alpha_{i,j}\beta_{i,j}e^{-\beta_{i,j}t}\mathbf{1}_{t \geq 0}$.

This article focuses on Hawkes processes with exponential kernel and constant baselines, to which we propose a maximum likelihood based fitting method.

3 Hawkes Processes: Model Calibration

The most commonly used technique for parametric inference of Hawkes processes is a direct numerical optimization of the Maximum Likelihood, which was first introduced in the work of Ogata (1978) [11]. He proved the asymptotic consistency and efficiency of the Maximum Likelihood Estimator (MLE) under the assumption of stationary condition of the underlying point process. The negative log-likelihood function of a multidimensional Hawkes process over the time interval $[0, t]$ is given in Daley & Vere-Jones [5] Proposition 13.1.VI by :

$$\mathcal{L}_t(\lambda) := - \sum_{i=1}^M \left(\int_0^t \log \lambda_i(\tau) dN_i(\tau) - \int_0^t \lambda_i(\tau) d\tau \right). \quad (4)$$

Depending on the shape of kernels, \mathcal{L}_t is generally non-convex. Classic non-convex numerical optimization pitfalls are to be feared: a direct numerical optimization could converge to a merely local minimum. On the other hand, it may take too many iterations to converge as the negative log-likelihood function can be flat on some regions of the space of parameters. This problem is also faced when using an Expectation Maximization (EM) algorithm to find the MLE.

On the computational side, the repeated evaluation of the negative log-likelihood can be highly time consuming, essentially due to the nested sum in the first part of 4 where the conditional intensity is also expressed as a sum over the history.

3.1 Maximum Likelihood Estimation (MLE) for exponential multidimensional Hawkes processes

Let's consider a M-dimensional multivariate Hawkes model with constant baselines μ_i and kernel functions $\phi_{i,j}$ as defined in 2.3. The choice of an exponential kernel has proved to be very interesting as it allows intuitive and meaningful interpretations of its parameters and also reduces the computational cost of the evaluation of the negative log-likelihood function as noted by Ogata (1981) [12] who exhibited a recursive formula that eliminated the nested sum evaluation problem.

The existing methods to fit exponential parameters through MLE directly use non-linear optimization algorithms such as Nelder-Mead (also called downhill simplex method) or BFGS, with performance decreasing as the number of parameters increases. This represents quite an issue as $M(2M + 1)$ parameters describe a M-dimensional Hawkes process with exponential kernel.

In addition, other existing methods often make the strong assumption that the decays $\beta_{i,j}$ of the exponential kernel are given and fixed a priori. Consequently, the estimation of each kernel function is equivalent to the estimation of its adjacency coefficient $\alpha_{i,j}$ and since the conditional intensity $\lambda_i(t)$ is linear with respect to kernel functions $\phi_{i,1}(t), \dots, \phi_{i,p}(t)$, the negative log-likelihood function \mathcal{L}_t is convex with respect to all parameters. Therefore we can use the widely available convex optimization machinery to find the global minimum efficiently. The main drawback of this method is the possible inaccuracy of the decays parameters $\beta_{i,j}$ which can lead to model mismatch.

3.2 Introducing Two Stage Hawkes Likelihood Optimization (2SHLO):

In order to combine the benefits of both approaches described above, we propose the following estimation method to fit exponential Hawkes processes.

The negative log-likelihood function for exponential multivariate Hawkes processes over a period $[0, T]$ is given by ²:

² See Ogata[12] for the recursive formulation of the negative log-likelihood.

$$\begin{aligned} \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \sum_{m=1}^M \left(\mu_m T + \sum_{n=1}^M \alpha_{mn} \sum_{\{k:t_k^n < T\}} [1 - e^{-\beta_{mn}(T-t_k^n)}] \right. \\ & \left. - \sum_{\{k:t_k^n < T\}} \log[\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} \sum_{\{k:t_k^n < t_i^m\}} e^{-\beta_{mn}(t_i^m - t_k^n)}] \right) \end{aligned} \quad (5)$$

As we mentioned earlier, if β is fixed, then $\mathcal{L}_t(\cdot, \cdot, \beta)$ is convex with respect to parameters $(\boldsymbol{\mu}, \boldsymbol{\alpha})$. As a result it can be minimized using any convex optimization algorithm, for example an Accelerated Gradient Descent (AGD) or a Projected Newton Descent (see [?]) which converges to a global minimum $(\boldsymbol{\mu}^*(\boldsymbol{\beta}), \boldsymbol{\alpha}^*(\boldsymbol{\beta}))$ such that:

$$\mathcal{L}_t^*(\boldsymbol{\beta}) = \mathcal{L}_t(\boldsymbol{\mu}^*(\boldsymbol{\beta}), \boldsymbol{\alpha}^*(\boldsymbol{\beta}), \boldsymbol{\beta}) = \min_{(\boldsymbol{\mu}, \boldsymbol{\alpha})} \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (6)$$

$\mathcal{L}_t^*(\boldsymbol{\beta})$ is still a non-convex function but defined on a space with a lower dimension than the initial parameter space. Any non-linear heuristic algorithms can now be used to minimize $\mathcal{L}_t^*(\boldsymbol{\beta})$ leading to the MLE as³:

$$\min_{(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{(\boldsymbol{\beta})} \min_{(\boldsymbol{\mu}, \boldsymbol{\alpha})} \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (7)$$

We summarize the fitting method as follow :

Algorithm 1 Two Stage Hawkes Likelihood Optimization (2SHLO)

```

Start from mean of inter arrival times  $\boldsymbol{\beta}_0$ 
for each step of Nelder-Mead method until convergence do
  for each needed computation of  $\mathcal{L}_t^*(\boldsymbol{\beta}_i)$  do
    Start from a random  $(\boldsymbol{\mu}_0, \boldsymbol{\alpha}_0)$ 
    Use a convex optimization algorithm (AGD or Projected Newton Descent) to
    minimize  $\mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i)$ 
    Retrieve resulting  $\mathcal{L}_t^*(\boldsymbol{\beta}_i)$ 
  end for
end for
return last  $(\boldsymbol{\beta}, \boldsymbol{\mu}^*(\boldsymbol{\beta}), \boldsymbol{\alpha}^*(\boldsymbol{\beta}))$ 

```

3.3 Goodness of fit

The *compensator function* Λ of a point process with intensity λ is defined as $\forall t \geq 0, \Lambda(t) = \int_0^t \lambda(s|\mathcal{F}_t)ds$. To assess the goodness of fit of our estimation, we use the residual point process analysis theorem, as stated in [5]:

³ if we call $m = \min_{(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, the following inequality holds for all $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$: $m \leq \mathcal{L}_t(\boldsymbol{\mu}^*(\boldsymbol{\beta}), \boldsymbol{\alpha}^*(\boldsymbol{\beta}), \boldsymbol{\beta}) \leq \mathcal{L}_t(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 3.2 follows immediately by taking the minimum over $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ in each side of the inequality.

The transformed sequence $\{t_k^*\} = \{A(t_k)\}$ is a realization of a unit rate Poisson process if and only if the original sequence $\{t_k\}$ is a realization from the point process defined by λ .

The goodness of fit can be tested by comparing the transformed estimated times to a standard Poisson process using QQ plots, comparison of density functions and Kolmogorov-Smirnov test.

4 Results - 2SHLO in practice

We have been running our experiments in Python 3, relying on *tick*[17] library for Hawkes process simulations, AGD optimization and analytic tools. 2SHLO is also benefiting of well-known *scipy* scientific package.

To provide an idea of the computational performance of the fitting method, we mention that, on average, it takes 2 minutes to fit a one-dimensional exponential Hawkes process on a training data with 15000 observations.

4.1 Validating 2SHLO performances over simulated data

To generate Hawkes processes, a few methods are available (as summarized in [2]), either based on thinning, time-change or cluster algorithm. We have here used *tick* functions for simulations which are based on the thinning algorithm described in (Ogata, 1981, p.25, Algorithm 2) [12].

To validate the performance of the fitting algorithm, we ran a series of 100 simulations and fitting procedures on the simulated 2D Hawkes time-stamps with different training sizes and using the following simulation parameters :

$$\mu = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix} \quad \alpha = \begin{pmatrix} 0.5 & 0.00 \\ 0.4 & 0.3 \end{pmatrix} \quad \beta = \begin{pmatrix} 0.3 & 0.00 \\ 0.2 & 0.2 \end{pmatrix}$$

Fig. 3 confirms that the estimated parameters converge to their true optimal values for increasing N number of ticks in the training set.

4.2 Calibrating univariate Hawkes processes on "mid-frequency" credit derivative trades

We have fitted a modified one-dimensional Hawkes process over the period from 15/01/2017 to 15/12/2017 of reported trades data⁴ for each indice ITXEB, ITXES and ITXEX. The trading activity being discontinuous, the absence of trades overnight has to be accounted in the model: a null value is imposed to the intensity outside trading hours - in the "gaps" between days - , while a standard exponential kernel model is considered during trading hours. The model is thus slightly modified and described in the appendix section in a more detailed way. Before making this modeling choice, Bowsher's model [4] (see appendix for

⁴ The trading data have been cleaned by discarding transactions that do not reflect market signals such as roll and switch trades

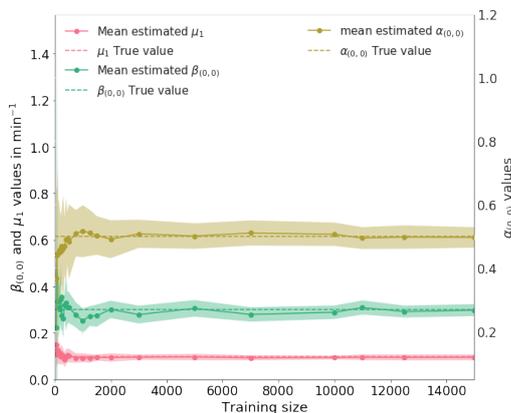


Fig. 3: Estimated parameters of a 2D exponential Hawkes process trained with 2SHLO on simulated datasets of increasing lengths.

more details) was experimented: this more elaborated model accounts for the "overnight spillover effect" where the intensity is recursively defined over days with a dependency on last trading day value, smoothed by a *spillover effect factor*. The fitting procedure resulted into a null parameter estimation of the spillover effect which motivated the modeling choice described above.

	Estimated μ^{-1}	Estimated α	Estimated β^{-1}
ITXEB	22 min	0.62	20 min
ITXEX	24 min	0.60	23 min
ITXES	58 min	0.65	36 min

Table 2: Estimated parameters for ITXEB, ITXES and ITXEX over the period (2017-01-15 to 2017-12-15)

To validate the universality of the estimates, the goodness of fit of each Hawkes model calibration was also tested on the out-of-sample period from 15/01/2018 to 15/07/2018. The model fitted well for the 3 index time series (as Q-Q plots show in Fig. 4) suggesting that each series has stable intrinsic parameters. However, highly demanding Kolmogorov-Smirnov (KS) tests failed on these large samples. Since with the KS test we try to find deviations from the exponential distribution, the larger the sample the better we are at detecting such deviations, even trivial small ones: outlier data points have a huge importance for passing KS-test. This issue has to be digged into in future work, to better quantify the heart of the distribution fit and to be able to detect outlier data points / days.

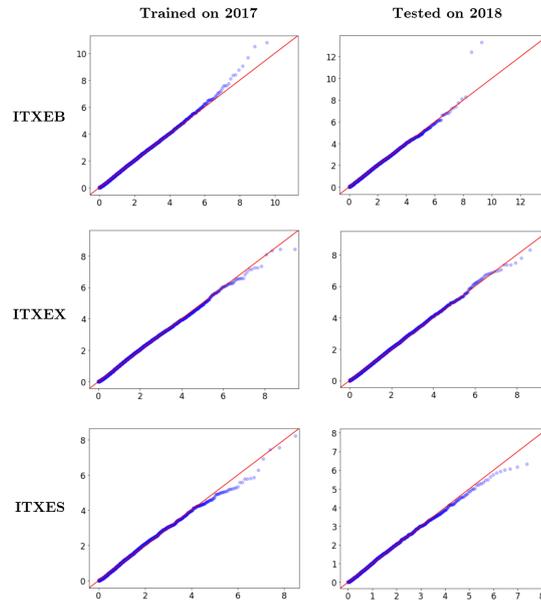


Fig. 4: Hawkes process fitted on 2017 data and tested on 2018 data for each index

4.3 Measuring the influence of traded volumes

When considering the distribution of trade volumes for credit indices, volume clusters are clearly outlined. For example, for index ITXEB, three bins can be distinguished: "small" trades, "medium" trades and "big" trades (See Fig. 5). Using multivariate Hawkes processes model, volume impacts have been modeled by considering 3 counting processes, splitting trades per volume size.

By using the same model as the last section to take overnight gaps into consideration, the fitting algorithm trained on 2018 data resulted in the following parameters estimates (μ and β expressed in minute^{-1}) :

$$\frac{1}{\mu} = \begin{pmatrix} 40 \\ 42 \\ 38 \end{pmatrix} \quad \alpha = \begin{pmatrix} 0.40 & 0.07 & 0.57 \\ 0.00 & 0.53 & 0.21 \\ 0.00 & 0.16 & 0.59 \end{pmatrix} \quad \frac{1}{\beta} = \begin{pmatrix} 12 & 21 & 53 \\ 71 & 18 & 12 \\ 53 & 14 & 34 \end{pmatrix}$$

Visualized in Fig. 5 the branching ratio, which corresponds the values of α , allows some intuitive interpretation of the fitted model as it shows that large trades are purely self-exciting process and they have a major influence in triggering small trades, we also notice that all trade categories have a non negligible self-excitation component.

4.4 Measuring cross interactions between traded indices

We have again adopted a multivariate Hawkes processes with null intensity overnight model to study the cross interaction between indices trades times.

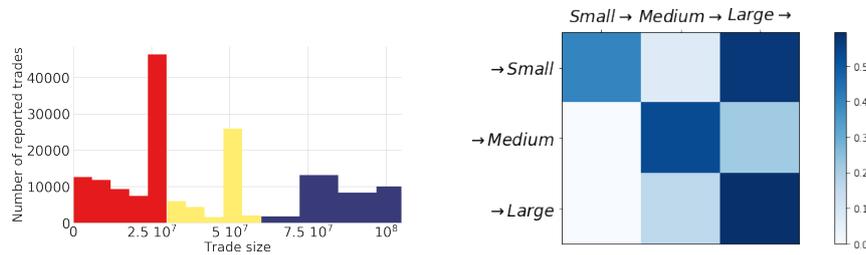


Fig. 5: ITXEB. *Left* : distribution of trade sizes; *right*: adjacency matrix $(\alpha_{i,j})$ of 3D exponential Hawkes fitted on 3 volume bin series.

The fitting algorithm trained on 2017 data resulted in the following parameters estimates (μ and β expressed in minute⁻¹) :

$$\frac{1}{\mu} = \begin{pmatrix} 29 \\ 30 \\ 125 \end{pmatrix} \quad \alpha = \begin{pmatrix} 0.44 & 0.25 & 0.20 \\ 0.23 & 0.37 & 0.18 \\ 0.07 & 0.08 & 0.45 \end{pmatrix} \quad \frac{1}{\beta} = \begin{pmatrix} 23 & 9 & 21 \\ 14 & 13 & 31 \\ 39 & 17 & 25 \end{pmatrix}$$

From Fig. 6 we can notice that ITXES index is the least influenced by other indices with very low cross excitation components and a high self excitation component which was expected due its low liquidity. We also notice that the estimated baselines of all indices are greater for this multivariate model compared to the univariate model. This can be explained by the fact that we decreased the 'Poissonian' behaviour of the arrival times by including more information from the influence of other indices.

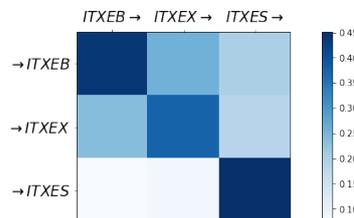


Fig. 6: Adjacency matrix $(\alpha_{i,j})$ of 3D exponential Hawkes fitted on 3 indices series (ITXEB, ITXES, ITXEX) for the period 2017-01-15 - 2017-12-15.

5 Conclusion

Through point process analysis, Hawkes processes allow a relatively simple and interpretable representation of time dependant events with self and mutual excitements. With 2SHLO; a maximum likelihood based algorithm mixing convex and non-convex optimizations, we have been able to fit fastly and properly exponential kernel multivariate Hawkes processes. The algorithm has been applied to credit derivative trading data, an unexplored "mid-frequency" market for such

processes. We have been able to emphasize self excitement for three index trades and to measure impacts of traded volumes. With the same framework, we were also able to quantify cross influences between indices. Yet, designing a proper test to assert the quality of fit to the heart of the data distribution and also to detect outliers are priority targets. Next prospects turns towards testing 2SHLO performances on high dimensional datasets, for instance studying both price & volume trade influences as well as market bid/ask dynamics on credit derivatives in a forecasting perspective.

References

1. Amaral, L., Papanicolaou, A.: Price Impact of Large Orders Using Hawkes Processes. In: NYU Tandon Research Paper No. 2874042 (2017)
2. Bacry, E., Mastromatteo, I., Muzy, J.F.: Hawkes Processes in Finance. In: Market Microstructure and Liquidity, vol. 01 (2015)
3. Bacry, E., Muzy, J.F.: Hawkes models for price and trades high-frequency dynamics. In: Quantitative Finance, vol. 14, pp. 1147–1166 (2013)
4. Bowsher, C. G.: Modelling Security Market Events in Continuous Time: Intensity Based, Multivariate Point Process Models. In: Nuffield Economics Working Paper (2003)
5. Daley, D.J., Vere-Jones, D.: An introduction to the theory of point processes, vol. 2. Springer (1988)
6. Bertsekas, D. P.: Projected Newton methods for optimization problems with simple constraints, SIAM Journal on Control and Optimization, vol. 20, pp. 221–226 (1982)
7. Hawkes, A.G.: Spectra of Some Self-Exciting and Mutually Exciting Point Processes. In: Biometrika, vol. 58, pp. 83–90 (1971)
8. Hawkes, A.G.: Point Spectra of Some Mutually Exciting Point Processes. In: Journal of the Royal Statistical Society. Series B (Methodological), vol. 33, pp. 438–443 (1971)
9. Laub, P., Taimre, T., Pollett, P.: Hawkes Processes. In: arXiv preprint arXiv:1507.02822 (2015)
10. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., Tita, G. E.: Self-Exciting Point Process Modeling of Crime. In: Journal of the American Statistical Association, vol. 106, pp. 100–108 (2011)
11. Ogata, Y., The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics, vol. 30, pp. 243–261 (1978)
12. Ogata, Y., On Lewis simulation method for point processes, IEEE Transactions on Information Theory, vol. 27, pp. 23–31 (1981)
13. Ogata, Y.: Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. In: Journal of the American Statistical Association, vol. 83, pp. 9–27 (1988)
14. Rambaldi, M., Bacry, E., Lillo, F.: The role of volume in order book dynamics: a multivariate Hawkes process analysis. In: Quantitative Finance (2016)
15. Rizoïu, M., Lee, Y., Mishra, S., Xie, L.: A Tutorial on Hawkes Processes for Events in Social Media. In: Frontiers of Multimedia Research, pp. 191–218 (2017)
16. Toke, I. M., An Introduction to Hawkes Processes with Applications to Finance (2011) http://lamp.ecp.fr/MAS/fiQuant/ioane_files/HawkesCourseSlides.pdf
17. Bacry E., Bompain M., Gaffas S., Poulsen S., Tick: a Python library for statistical learning. <https://x-datainitiative.github.io/tick/#>

Appendix: Hawkes Processes Extended Models

Univariate Bowsher Hawkes process model

Definition: The univariate Bowsher process, as described in [4], is defined recursively depending on the level of the stochastic parts in the intensity function at the end of the $(d - 1)$ th trading day, and the contributions of the events occurring on day d .

The observation time period $[0, T]$ is partitioned into different trading days; with $]\tau_{d-1}, \tau_d]$ being the interval defining day d .

The model is defined by the (scalar) stochastic intensity of the counting process $N(t)$:

$$\forall t \geq 0, \quad \lambda(t) = \mu + \tilde{\lambda}(t)$$

such that :

$$\forall t \in]\tau_{d-1}, \tau_d], \quad \tilde{\lambda}(t) = \pi \tilde{\lambda}(\tau_{d-1}) e^{-\rho(t-\tau_{d-1})} + \int_{[\tau_{d-1}, t)} \alpha e^{-\beta(t-u)} dN(u)$$

where :

- $\mu \in \mathbb{R}^+$ is the *baseline intensity*
- $\pi \in [0, 1]$ is the *spillover adjacency* coefficient
- $\rho \in \mathbb{R}^+$ is the *spillover decay*
- $\alpha \in [0, 1]$ is the self-excitement *adjacency* coefficient
- $\beta \in \mathbb{R}^+$ is the self-excitement *decay*

Log-likelihood: Computing the negative log-likelihood for Bowsher model results into the following closed form :

$$\begin{aligned} \mathcal{L}_T(\lambda) &= \mu T + \sum_{d \in \text{days}} \pi \tilde{\lambda}(\tau_{d-1}) (1 - e^{-\rho(\tau_d - \tau_{d-1})}) \\ &+ \alpha \sum_{t \in]\tau_{d-1}, \tau_d]} (1 - e^{-\rho(t - \tau_{d-1})}) - \sum_{t_i} \log \lambda(t_i) \end{aligned}$$

Fitting procedure: As the task of computing the gradient for each parameter of the log-likelihood is a fastidious one because of the recursive definition of the intensity, and also because we have only 5 parameters to fit $(\mu, \pi, \rho, \alpha, \beta)$, we choose to use the L-BFGS-B algorithm to minimize $\mathcal{L}_T(\lambda)$. It had delivered satisfying performance on fitting simulated data.

Day Gaps Multivariate Exponential Hawkes process model

Definition: To deal with gaps (nights, week-ends...) between trading days, we propose to tweak the standard M -multivariate Hawkes process with exponential kernel $\phi_{ij}(t) = \alpha_{ij}\beta_{ij} \exp(-\beta_{ij}t)1_{t>0}$ (as introduced in section 2.3): a null value is imposed for intensity functions outside trading hours as so as to take the day gaps into consideration.

For all t between two consecutive trading days, for all $i \in \llbracket 1, M \rrbracket$, $\lambda_i(t) = 0$.

Log-likelihood: Recalling from section 3, the negative log-likelihood of a Multivariate Point Process is defined as :

$$\mathcal{L}_t(\boldsymbol{\lambda}) = - \sum_{i=1}^M \left(\int_0^t \log \lambda_i(\tau) dN_i(\tau) - \int_0^t \lambda_i(\tau) d\tau \right).$$

Let's introduce parameters :

- D , the number of days in the data set
- δ , the fixed length of one trading day
- τ_d the last time of the day d

Computing the negative log-likelihood for Day Gaps Multivariate Exponential Hawkes process model results into the following closed form :

$$\begin{aligned} \mathcal{L}_t(\boldsymbol{\lambda}) := & D * \delta \sum_{m=1}^M \mu_m + \sum_{m=1}^M \sum_{n=1}^M \alpha_{mn} \sum_{d \in \text{days}} \sum_{t_i^n \in \text{day } d} (1 - e^{-\beta_{mn}(\tau_d - t_i^n)}) \\ & - \sum_{m=1}^M \sum_{t_i^m} \log(\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} R_{mn}(i)) \end{aligned}$$

where $R_{mn}(i)$ are defined using Ogata [12] recursive formula :

$$\begin{aligned} R_{mn}(i) = & \sum_{\{k: t_k^n < t_i^m\}} e^{-\beta_{mn}(t_i^m - t_k^n)} \\ = & e^{-\beta_{mn}(t_i^m - t_{i-1}^m)} R_{mn}(i-1) + \sum_{\{k: t_{i-1}^m \leq t_k^n < t_i^m\}} e^{-\beta_{mn}(t_i^m - t_k^n)} \end{aligned}$$

Fitting procedure: To fit this model, we use 2SHLO with the Projected Newton Descent [6] for the convex optimization part. The choice of Newton Descent is justified by the easy computation of the gradient and hessian matrix, which is sparse, and because this algorithm has a quadratic convergence rate compared to gradient descent methods.

The gradient of $\mathcal{L}_t(\boldsymbol{\lambda})$ is given by the following close formulas :
 Given $(\beta_{i,j})_{1 \leq i,j \leq M}$, for all $m, n \in \llbracket 1, M \rrbracket^2$:

$$\begin{aligned} \frac{\partial \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \mu_m} &= D * \delta - \sum_{t_i^m} \frac{1}{\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} R_{mn}(i)} \\ \frac{\partial \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \alpha_{mn}} &= \sum_{d \in \text{days}} \sum_{t_i^n \in \text{day } d} (1 - e^{-\beta_{mn}(\tau_d - t_i^n)}) - \sum_{t_i^m} \frac{\beta_{mn} R_{mn}(i)}{\mu_m + \sum_{l=1}^M \alpha_{ml} \beta_{ml} R_{ml}(i)} \end{aligned}$$

The hessian of $\mathcal{L}_t(\boldsymbol{\lambda})$ is given by the following close formulas :
 For all $m, n, l \in \llbracket 1, M \rrbracket^3$:

$$\begin{aligned} \frac{\partial^2 \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \mu_m \partial \mu_m} &= \sum_{t_i^m} \frac{1}{(\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} R_{mn}(i))^2} \\ \frac{\partial^2 \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \mu_m \partial \alpha_{ml}} &= \frac{\partial^2 \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \alpha_{ml} \partial \mu_m} = \sum_{t_i^m} \frac{\beta_{ml} R_{ml}(i)}{(\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} R_{mn}(i))^2} \\ \frac{\partial^2 \mathcal{L}_t(\boldsymbol{\lambda})}{\partial \alpha_{mk} \partial \alpha_{ml}} &= \sum_{t_i^m} \frac{\beta_{mk} R_{mk}(i) \beta_{ml} R_{ml}(i)}{(\mu_m + \sum_{n=1}^M \alpha_{mn} \beta_{mn} R_{mn}(i))^2} \end{aligned}$$

all other coefficients are null.

The Projected Newton Descent is described as follow :

Algorithm 2 Projected Newton Descent

Parameters: max_iterations, tolerance and shrink $\in [0, 1]$.

Given convex function f , convex set C and projection function *project* on C :

Starting from x_0 .

prev_x = x_0 , $x = x_0$

prev_v = $f(x_0)$, $v = f(x_0)$

repeat

 step = 1

 next_x = project($x - \text{step} * (\nabla^2 f(x))^{-1} \nabla f(x)$),

 next_v = $f(\text{next}_x)$

while next_v > v **do**

 step *= shrink

 next_x = project($x - \text{step} * (\nabla^2 f(x))^{-1} \nabla f(x)$)

 next_v = $f(\text{next}_x)$

end while

 Set prev_x = x, prev_v = v

 Set x = next_x, v = next_v

until Convergence is validated: $\text{abs}(v - \text{prev}_v) / \text{abs}(\text{prev}_v) < \text{tolerance}$; or
 max_iterations is reached.

return x

Tests for Segmented Cointegration: An Application to US Governments Budgets*

Luis F. Martins[†]

Paulo M. M. Rodrigues^{‡§}

June 2018

Abstract

There is a growing literature documenting that the persistence of economic time series may change over time (see, e.g. Martins and Rodrigues, 2012, ...). Hence, it is natural also to expect that changes occur in the long-run equilibrium of economic time series. For instance, Siklos and Granger (1997) suggest that failure to find a unique cointegration relationship between economic time series may be due to the testing procedures used, the span of the data set analysed, the choice of the lag length in generating the test statistics, the presence of structural breaks and the presence of cointegration only beyond some threshold. This lead Siklos and Granger (1997) to introduce the concept of regime-sensitive cointegration, according to which the underlying series need not be cointegrated at all times. The purpose of this paper is to propose tests that can be used in the context of segmented cointegration.

Keywords: Cointegration, Cointegration Breakdown, Nonstationarity, Breaks

JEL classification:C12, C22

[Preliminary version. Please do not quote without permission from authors.]

* **Acknowledgements:** ??????

[†]ISCTE-IUL, Business School and UNIDE. E-mail: luis.martins@iscte.pt.

[‡]Banco de Portugal and Nova School of Business and Economics, Universidade Nova de Lisboa. E-mail: pmrodrigues@bportugal.pt

[§]**Correspondence to:** Paulo M. M. Rodrigues, Banco de Portugal, Economics and Research Department, Av. Almirante Reis, 71-6th floor, 1150-012 Lisbon, Portugal (e-mail: pmrodrigues@bportugal.pt).

1 Introduction

Long-run equilibrium relationships are of considerable importance in economics and econometrics. Many economic theories are concerned with equilibrium relationships and as a result a large empirical literature on testing for cointegration has developed in economics and finance, particularly since the seminal works of Engel and Granger (1987) and Johansen (1988).

It is typically assumed that cointegrating relationships do not change over time, which however may be a restrictive assumption. Although, the impact of structural breaks in the deterministic kernels on cointegration have been widely analysed (see e.g. Hansen (1992); Quintos and Phillips (1993); Hao (1996); Andrews et al. (1996); Bai et al. (1998); Kuo (1998), Lütkepohl et al. (2003); Inoue (1999); and Johansen et al. (2000)), less attention has been given to the impact of changes in the actual long-run equilibrium.

Recently, a large literature documenting that economic time series may display persistence change over time has emerged (see, *inter alia*, Kim, 2000, Harvey, Leybourne and Taylor, 2006 and Martins and Rodrigues, 2012). Hence, it is natural to expect that changes in the persistence of economic time series may originate changes in their long-run equilibrium. According to Siklos and Granger (1997) failure to find a unique cointegration relationship between economic time series may be a consequence of the properties of the test procedures used, the span of the data set considered, the choice of the lag length in generating the test statistics, the presence of structural breaks and the presence of cointegration only beyond some threshold. This lead Siklos and Granger (1997) to introduce the concept of regime-sensitive cointegration, according to which the underlying series need not be cointegrated at all times. Siklos and Granger (1997) argue that "events or important changes in some of the institutional features of an economy can interrupt an underlying equilibrium-type relationship possibly for an extended period of time." Note that this view is different from the notion of time varying cointegration, since in the latter cointegration exists at all times. The regime-sensitive cointegration concept considers that economic relationships may occasionally fall in or out of equilibrium because of major events.

Two important challenges of modern empirical macroeconomics involve the incorporation of restrictions suggested by economic theory and the empirical need to allow for parameter changes in multivariate time series models (Jochmann and Koop, 2011). With regard to the former, cointegration has played an important role since economic theory often suggests particular cointegrating relationships which the researcher may impose or test; see e.g. Garratt, Lee, Pesaran and Shin (2003). With regards to the latter, Ang and Bekaert (2002) and Stock and Watson (1996), amongst

others, document widespread evidence of parameter changes in many macroeconomic time series. In the field of cointegration, there are a large number of theoretical and empirical papers that model breaks or other forms of nonlinearities in cointegrating relationships and present empirical results relating to cointegration work using subsamples of the data or attribute failures of cointegration tests to parameter changes (see, among others, Michael, Nobay and Peel, 1997, Quintos, 1997, Park and Hahn, 1999, Lettau and Ludvigson, 2004, Saikkonen and Choi, 2004, Andrade, Bruneau and Gregoir, 2005, Beyer, Haug and Dewald, 2009 and Bierens and Martins, 2010).

The remainder of the paper is organised as follows. Section 2 discusses the time series process under analysis, Section 3 introduces the tests for segmented cointegration, Section 4 presents the asymptotic properties of the tests, Section 5 provides an indepth Monte Carlo analysis, Section 6 presents an empirical analysis, Section 7 concludes and an Appendix includes detailed proofs of the main results provided in the text.

2 Segmented Cointegration

Consider the data generation process (DGP),

$$y_t = \mu_{yt} + \beta' \mathbf{x}_t + \varepsilon_t \quad (2.1)$$

$$\mathbf{x}_t = \mu_{xt} + \mathbf{v}_{xt} \quad (2.2)$$

$$\mathbf{v}_{xt} = \mathbf{v}_{x,t-1} + \mathbf{u}_t \quad (2.3)$$

where \mathbf{x}_t is a $K \times 1$ vector of regressors, y_t is a scalar and ε_t in (2.1) is such that,

$$\varepsilon_t = \alpha_j + \rho_j \varepsilon_{t-1} + v_t, \quad j = 1, \dots, m; \quad m \geq 1; \quad t = 2, \dots, T. \quad (2.4)$$

Furthermore, μ_{yt} in (2.1) and μ_{xt} in (2.2) are deterministic kernels. For the purpose of analysis we consider as, for instance, Perron and Rodriguez [PR] (2015), three empirically relevant cases, namely: $\mu_{yt} = 0$; $\mu_{yt} = \phi_{k0}$ and $\mu_{yt} = \phi_{k0} + \phi_{k1}t$; and similarly $\mu_{xt} = \mathbf{0}$, $\mu_{xt} = \phi_{\mathbf{x}0}$ and $\mu_{xt} = \phi_{\mathbf{x}0} + \phi_{\mathbf{x}1}\mathbf{t}$ with $\phi_{\mathbf{x}i}$, $i = 0, 1$, a $K \times 1$ vector of parameters.

Assumption A:

A.1: The errors $\{v_t\}$ in (2.4) are generated by a stationary linear process $v_t = \beta(L) a_t$, with $\beta(L) = \sum_{s=0}^{\infty} \beta_s L^s$ and $\sum_{s=0}^{\infty} s |\beta_s| < \infty$.

A.2: The process $\{a_t\}$ is a martingale difference sequence with $E(a_t^2 | a_{t-1}, \dots) = \sigma^2$, $E(|a_t|^r | a_{t-1}, \dots) =$

κ_r ($r = 3, 4$) and $\sup_t E(|a_t|^{4+\epsilon} |a_{t-1}, \dots) = \kappa < \infty$ for some $\epsilon > 0$.

A.3: All roots of $\beta(L)$ are outside the unit circle.

Assumption B:

B.1: The vector $\xi_t := (v_t, \mathbf{u}'_t)'$ a VMA(inf) representation, $\xi_t = \sum_{i=0}^{\infty} \Phi_i \eta_t$, with $\sum_{i=0}^{\infty} i \det |\Phi_i| < \infty$, $\Phi_0 = I_n$ and ξ_t is a martingale difference sequence with respect to the sigma-field \mathcal{F}_t , $E(\xi_t \xi'_t | \mathcal{F}_{t-1}) =: \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$, with Ω_{11} a $(m+1) \times (m+1)$ and Ω_{22} a $(n-1) \times (n-1)$ are positive semi-definite matrices.

We consider that under the null hypothesis there is no persistence change in $\{\varepsilon_t\}$, i.e., $H_0 : \rho_1 = \dots = \rho_{m+1} = 1$, but that under the alternative, m changes ($m+1$ regimes) of the cointegrating relationship are allowed. In other words, under the alternative, ρ_j in (2.4) is,

$$H_A : \rho_j = \begin{cases} 1 & \text{for } t \in \mathcal{N}_T \\ |\rho_j| < 1 & \text{for } t \in \mathcal{C}_T \end{cases}, \quad j = 1, \dots, m+1. \quad (2.5)$$

In a single segmented cointegration context (i.e., for $m = 1$), as considered for instance in Kim (2000) and Davidson and Monticini [DM] (2010), $\{\varepsilon_t\}$ follows a unit root process in subset $N_T := \lfloor \lambda T \rfloor$ of the sample, and a stationary process in the remainder of the sample, $C_T := \lfloor (1 - \lambda)T \rfloor$, with $\lambda \in (0, 1)$ and $T = N_T + C_T$. In a general context, as in e.g. Kejriwal, Perron and Zhou [KPZ] (2013), N_T and C_T may consist of multiple intervals so that $t = T_{j-1} + 1, \dots, T_j$, for $j = 1, \dots, m+1$ with $T_0 = 1$ and $T_{m+1} = T$. This framework allows for m structural breaks ($m+1$ regimes) which originate consecutive switches from nonstationarity (stationarity) to stationarity (nonstationarity). Thus, a sample of size T can be decomposed into s_N subsamples of size $N_{k_N T}$, $k_N = 1, \dots, s_N$ (the nonstationary parts) and s_C subsamples of size $C_{k_C T}$, $k_C = 1, \dots, s_C$, (the stationary parts), with $s_N + s_C = m+1$ and $|s_N - s_C| \leq 1$.

For the purpose of analysis, under the alternative hypothesis, we consider two specific cases for the process of $\{\varepsilon_t\}$ in (2.4):

1. $H_{1A} : \alpha_j = 0$ and $\rho_j = 1$ in odd regimes, and $|\rho_j| < 1$ in even regimes (the first regime is locally spurious);
2. $H_{1B} : \alpha_j = 0$ and $\rho_j = 1$ in even regimes, and $|\rho_j| < 1$ in odd regimes (the first regime is locally cointegrated).

To simplify our analysis we rule out changes in the dynamics (i.e., ρ_j are all the same in C_T)

and in the variance of the errors v_t in (2.4); see also KPZ (p. 293).

Remark 2.1: The inclusion of α_j in (2.4) (contrary to Kim, 2003, and PR) is to avoid spurious jumps at the breakpoints (see Kejriwal and Perron, 2012, and Leybourne, Kim and Taylor, 2007, for details), as well as to preserve the I(1) property of y_t in (2.1), as this process inherits the properties of $\{\varepsilon_t\}$ itself. \square

Remark 2.2: Contrary to KPZ, but similar to PR, we rule out concurrent changes in the level and slope of the trend function. Following Kim (2003) and DM, but contrary to Bai and Perron (1998) and Kejriwal and Perron, (2010) and others, we assume that μ_0, μ_1 , and β are not subject to shifts so that ε_t can be estimated using the entire sample by standard methods in a single-equation cointegration framework. \square

3 Testing for Segmented Cointegration

In what follows, we generalise existing persistence change tests developed for observed univariate time series to residual-based series, $\{e_t\}$, obtained from (2.1) to test for segmented cointegration. In particular, we generalise and evaluate the Wald test for persistence change proposed by KPZ and the local GLS modified unit root tests proposed by LKT and PR. The null hypothesis is, as previously indicated, that the whole sample is I(1), whereas the alternative hypothesis considers multiple shifts in the cointegrating relationship, as indicated in (2.4).

3.1 Wald type tests

The Wald-type procedures proposed by KPZ (p.293), to test for multiple persistence changes in univariate time series, compare the sum of squared residuals (SSR) computed under the null hypothesis of a spurious regression (i.e. $\{e_t\}$ is I(1) throughout the sample), against the SSR computed under the alternative hypothesis of m fixed changes.

In specific, considering Assumptions $\mathcal{A}.1 - \mathcal{A}.3$ and Assumptions $\mathcal{B}.1$, $\{e_t\}$ has an AR representation that can be approximated by a finite order $AR(p_T)$, with $p_T \rightarrow \infty$ as $T \rightarrow \infty$. Consequently, an augmented ADF type test regression of the OLS residuals of (2.1), which we define as e_t , can be considered, *viz.*,

$$\Delta e_t = c_j + \phi_j e_{t-1} + \sum_{i=1}^{p_T} \pi_i \Delta e_{t-i} + a_{p_T,t} \quad (3.1)$$

where $\phi_j := (\rho_j - 1)$ and $a_{p_T,t}$ is a white noise error term.

Following KPZ, if under the alternative hypothesis, the number of changes is fixed, $m = m^*$, the persistence change test statistics under H_{1A} and H_{1B} are respectively,

$$F_k(\tau, m^*) := \begin{cases} \frac{(T-m^*-2\delta_B-p_T)(SSR_0-SSR_{k,m^*})}{((m^*+2\delta_B)SSR_{k,m^*})} & \text{if } m^* \text{ is even} \\ \frac{(T-m^*-1-p_T)(SSR_0-SSR_{k,m^*})}{((m^*+1)SSR_{k,m^*})} & \text{if } m^* \text{ is odd} \end{cases}, \quad k = A, B. \quad (3.2)$$

Recall that under H_{1A} the first regime is I(1) whereas under H_{1B} the first regime is I(0). Moreover, $\tau := (\tau_1, \dots, \tau_{m^*})$, $\tau_j := T_j/T$, and $\delta_B = 1$ when H_{1B} is considered and $\delta_B = 0$, otherwise. SSR_0 in (3.2) denotes the restricted sum of squared OLS residuals computed from (3.1) imposing the null hypothesis that $c_j = 0$ and $\phi_j = 0$, for all j , and SSR_{k,m^*} , $k = A, B$, denote the unrestricted sum of squared OLS residuals from model (3.1) considering either H_{1A} or H_{1B} . In specific,

$$SSR_{A,m^*} := \begin{cases} SSR_{0,(1:T_1)} + SSR_{a,(T_1:T_2)}^2 + \dots + SSR_{0,(T_{m^*+1}:T_1)} & \text{if } m^* \text{ is even} \\ SSR_{0,(1:T_1)}^2 + SSR_{a,(T_1:T_2)}^2 + \dots + SSR_{a,(T_{m^*+1}:T_1)}^2 & \text{if } m^* \text{ is odd} \end{cases},$$

and

$$SSR_{B,m^*} := \begin{cases} SSR_{a,(1:T_1)}^2 + SSR_{0,(T_1:T_2)}^2 + \dots + SSR_{a,(T_{m^*+1}:T)}^2 & \text{if } m^* \text{ is even} \\ SSR_{a,(1:T_1)}^2 + SSR_{0,(T_1:T_2)}^2 + \dots + SSR_{0,(T_{m^*+1}:T)}^2 & \text{if } m^* \text{ is odd} \end{cases},$$

where

$$SSR_{0,(T_k:T_{k+1})} = \sum_{t=T_k+1}^{T_{k+1}} \left(\Delta e_t - \sum_{i=1}^{p_T} \hat{\pi}_{k+1,i} \Delta e_{t-i} \right)^2,$$

$$SSR_{a,(T_k:T_{k+1})}^2 := \sum_{t=T_k+1}^{T_{k+1}} \left(\Delta e_t - \hat{c}_{k+1} - \hat{\phi}_{k+1} e_{t-1} - \sum_{i=1}^{p_T} \hat{\pi}_{k+1,i} \Delta e_{t-i} \right)^2.$$

Recall that $T_0 = 1$ and $T_{m^*+1} = T$, $\hat{\pi}_{ki}, \hat{c}_k, \hat{\phi}_k$ are the OLS estimates computed from the $m^* + 1$ subsamples $t = T_{k-1}, \dots, T_k$, with $k = 1, \dots, m^* + 1$, and $T_0 = 1$ and $T_{m^*+1} = T$. Note that the autoregressive order p_T used is the same when estimating the models under the null and the alternative hypotheses (KPK p. 294).

Hence, considering $\Lambda_\epsilon^{m^*} := \{\tau : |\tau_{j+1} - \tau_j| \geq \epsilon, \tau_1 \geq \epsilon, \tau_{m^*} \leq 1 - \epsilon\}$ the sup-Wald test for segmented cointegration under H_{1A} and H_{1B} is,

$$\sup F_k(m^*) := \sup_{\tau \in \Lambda_\epsilon^{m^*}} F_k(\tau, m^*), \quad k = A, B. \quad (3.3)$$

However, since typically the order of integration (I(0) or I(1)) of the first regime under the alter-

native is unknown, the following test statistics for $m = m^*$ and fixed is considered,

$$\mathcal{W}(m^*) := \max[\sup F_A(m^*), \sup F_B(m^*)].$$

Moreover, to accommodate the case of an unknown number of m breaks, up to some maximal value \bar{m} , (e.g. $\bar{m} = 5$), we further consider,

$$\mathcal{W} \max := \max_{1 \leq m \leq \bar{m}} \mathcal{W}(m). \quad (3.4)$$

3.2 Sequential type tests

Leybourn, Kim and Taylor [LKT] (2007) also introduced a test for multiple persistence changes. LKT's test is based on sequences of doubly-recursive implementations of the regression-based unit root test statistic of Elliott, Rothenberg and Stock (1996). In this section, we generalise this approach to the present context and consider also the trinity of the so-called M unit root tests introduced by Stock (1999) and Perron and Ng (1996).

The test statistics are based on local GLS detrended residuals, which we denote by \tilde{e}_t . In specific, local GLS detrending follows from assuming a local to unity framework in (2.4). In specific, we consider $\bar{\rho}_1 = \dots = \bar{\rho}_m = \bar{\rho}$, where $\bar{\rho}$ is a local to unity parameter, such as,

$$\bar{\rho} = 1 + \bar{c}/T, \text{ for } \bar{c} < 0 \text{ and fixed.} \quad (3.5)$$

Note that \bar{c} is obtained by Monte Carlo simulations (see, for instance, ERS, LKT, PR and Rodrigues and Taylor, 2007, among others). In the unit root testing context with no break, ERS set $\bar{c} = -7$ if demeaning ($D_t = 1$) is considered and $\bar{c} = -13.5$ if detrending ($D_t = (1, t)'$) is applied. Ng and Perron (1996) consider the same values for \bar{c} since the asymptotic distribution of the M t-type test (MZ_t) is the same as that of the local GLS detrended DF t-statistic. In the context of cointegration ($n > 1$), but with no breaks, PR show that the "optimal" \bar{c} is larger in absolute value and that it depends on the number of regressors (n) and the number of deterministic components included in the test regression (the more of each, the larger \bar{c} becomes). LKT in their analysis consider $\bar{c} = -10$ when $D_t = 1$ or $D_t = (1, t)'$ is used.

Following LKT, the local GLS detrended residual, \tilde{e}_t , is obtained as,

$$\tilde{e}_t = e_t - \hat{\delta}_{\bar{\rho}}' D_t, \quad (3.6)$$

where $\widehat{\delta}_{\bar{\rho}}$ is the estimate obtained from regressing $(e_{\lfloor \lambda T \rfloor}, e_{\lfloor \lambda T \rfloor + 1} - \bar{\rho}e_{\lfloor \lambda T \rfloor}, \dots, e_{\lfloor \tau T \rfloor} - \bar{\rho}e_{\lfloor \tau T \rfloor - 1})'$ on $(D_{\lfloor \lambda T \rfloor}, D_{\lfloor \lambda T \rfloor + 1} - \bar{\rho}D_{\lfloor \lambda T \rfloor}, \dots, D_{\lfloor \tau T \rfloor} - \bar{\rho}D_{\lfloor \tau T \rfloor - 1})'$ with $\bar{\rho}$ as defined in (3.5).

Alternatively to (3.6), PR suggest detrending each variable separately. That is, each series is locally GLS detrended first using a $\bar{\rho}$ as in (3.5). For Model 1 (no deterministic), $\tilde{y}_t = y_t$ and $\tilde{\mathbf{x}}_t = \mathbf{x}_t$. For Model 2 (constant only), $\tilde{y}_t = y_t - \widehat{\delta}'_{1\bar{\rho}}$ and $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \widehat{\delta}'_{2\bar{\rho}}$ where $\widehat{\delta}_{1\bar{\rho}}$ is obtained from regressing $(y_1, y_2 - \bar{\rho}y_1, \dots, y_T - \bar{\rho}y_{T-1})'$ on $(1, 1 - \bar{\rho}, \dots, 1 - \bar{\rho})'$ and $\widehat{\delta}_{2\bar{\rho}}$ from regressing $(\mathbf{x}_1, \mathbf{x}_2 - \bar{\rho}\mathbf{x}_1, \dots, \mathbf{x}_T - \bar{\rho}\mathbf{x}_{T-1})'$ on $(1, 1 - \bar{\rho}, \dots, 1 - \bar{\rho})'$. For Model 3 (constant and time trend), $\tilde{y}_t = y_t - \widehat{\delta}'_{1\bar{\rho}}D_t$ and $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \widehat{\delta}'_{2\bar{\rho}}D_t$, with $D_t = (1, t)'$, where $\widehat{\delta}_{1\bar{\rho}}$ and $\widehat{\delta}_{2\bar{\rho}}$ are obtained as for Model 2 but the regressions are on $(D_1, D_2 - \bar{\rho}D_1, \dots, D_T - \bar{\rho}D_{T-1})'$ instead. Finally, $\tilde{e}_t = \tilde{y}_t - \widehat{\beta}'\tilde{\mathbf{x}}_t$ and in contrast to (3.6), $\widehat{\beta}$ is the OLS estimate of β obtained from $\tilde{y}_t = \beta'\tilde{\mathbf{x}}_t + \tilde{e}_t$, computed from the complete sample.

Remark 3.1: In our framework we compute \tilde{y}_t and $\tilde{\mathbf{x}}_t$ in exactly the same way as PR. The only difference is that our statistics use \tilde{e}_t , $t = \lfloor \lambda T \rfloor, \lfloor \lambda T \rfloor + 1, \dots, \lfloor \tau T \rfloor - 1, \lfloor \tau T \rfloor$. Alternatively, one could consider obtaining $\widehat{\delta}$ from regressing $(x_{\lfloor \lambda T \rfloor}, x_{\lfloor \lambda T \rfloor + 1} - \bar{\rho}x_{\lfloor \lambda T \rfloor}, \dots, x_{\lfloor \tau T \rfloor} - \bar{\rho}x_{\lfloor \tau T \rfloor - 1})'$ on $(D_{\lfloor \lambda T \rfloor}, D_{\lfloor \lambda T \rfloor + 1} - \bar{\rho}D_{\lfloor \lambda T \rfloor}, \dots, D_{\lfloor \tau T \rfloor} - \bar{\rho}D_{\lfloor \tau T \rfloor - 1})'$. \square

Similar to LKT (see also DM and Kim, 2000) the statistics in our paper are recursively computed using subsamples of the local GLS de-trended residuals, \tilde{e}_t . Let $\tau \in (\lambda, 1]$, for a given λ in $(0, 1)$ and define the trinity of M unit root test statistics computed from observations at time $t = \lfloor \lambda T \rfloor, \lfloor \lambda T \rfloor + 1, \dots, \lfloor \tau T \rfloor - 1, \lfloor \tau T \rfloor$ as,

$$MSB^{GLS}(\lambda, \tau) := \left(\frac{1}{s(\lambda, \tau)^2} \frac{1}{[(\tau - \lambda)T]^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\tilde{e}_{t-1})^2 \right)^{1/2} \quad (3.7)$$

$$MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) := \left[\frac{2}{[(\tau - \lambda)T]^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\tilde{e}_{t-1})^2 \right]^{-1} \left[\frac{1}{[(\tau - \lambda)T]} (\tilde{e}_{\tau T})^2 - s(\lambda, \tau)^2 \right] \quad (3.8)$$

and

$$MZ_{t\hat{\rho}}^{GLS}(\lambda, \tau) := MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \times MSB^{GLS}(\lambda, \tau) \quad (3.9)$$

where $s(\lambda, \tau)^2$ is a consistent subsample estimate of the long-run variance of the local GLS de-trended ε_t , *i.e.*,

$$s(\lambda, \tau)^2 := \frac{s_{\eta b_T}(\lambda, \tau)^2}{(1 - \sum_{i=1}^{p_T} \widehat{\pi}_i)^2}, \quad (3.10)$$

where $s_{\eta b_T}(\lambda, \tau)^2 := \frac{1}{[(\tau - \lambda)T] - p_T + 1} \sum_{t=\lfloor \lambda T \rfloor + p_T}^{\lfloor \tau T \rfloor} (\widehat{\eta}_t^*)^2$ and $\widehat{\pi}_i$ and $\widehat{\eta}_t^*$ are OLS estimates obtained

from the ADF type regression,

$$\Delta \tilde{e}_t = \phi_0 \tilde{e}_{t-1} + \sum_{i=1}^{p_T} \pi_i \Delta \tilde{e}_{t-i} + \eta_t^*, \text{ for } t = \lfloor \lambda T \rfloor, \lfloor \lambda T \rfloor + 1, \dots, \lfloor \tau T \rfloor - 1, \lfloor \tau T \rfloor.$$

Ng and Perron (2001) suggest the use of the modified information criteria (MIC) to select p_T , which is defined as,

$$p_{T,MIC} := \arg \min_{p_T \leq p_{\max}} MIC(p_T),$$

where, for a subperiod $t = \lfloor \lambda T \rfloor, \dots, \lfloor \tau T \rfloor$,

$$MIC(p_T) := \ln(s_{\eta b_T}^2(p_{\max})) + \frac{2[\varrho_T(p_{\max}) + p_T]}{[(\tau - \lambda)T] - p_{\max}}$$

with $\varrho_T(p_{\max}) := \frac{\hat{\rho}_0^2}{\hat{\sigma}_\eta^2(p_{\max})} \sum_{t=\lfloor \lambda T \rfloor + p_{\max}}^{\lfloor \tau T \rfloor} (\tilde{e}_{t-1})^2$ and $s_{\eta b_T}^2(p_{\max}) := \frac{1}{[(\tau - \lambda)T] - p_{\max} + 1} \sum_{t=\lfloor \lambda T \rfloor + p_{\max}}^{\lfloor \tau T \rfloor} (\hat{\eta}_t^*)^2$.

The maximum lag order, p_{\max} , considered is determined based on Schwert's rule.

Remark 3.3: Following Perron and Ng (1996) we can consider two alternative estimators for the long-run variance. Firstly, a non-parametric kernel estimator based on the sample autocovariances, $\hat{\lambda}^2 = s_{WA}^2$, with $s_{WA}^2 := \sum_{h=-T+1}^{T-1} \omega(h/b) \hat{\gamma}_h$, $\hat{\gamma}_h := n^{-1} \sum_{t=1}^{T-|h|} e_{0,t} e_{0,t+|h|}$, where $e_{0,t}$ are the OLS residuals from regressing e_t on e_{t-1} , with kernel function $\omega(\cdot)$ satisfying e.g. the general conditions reported in Jansson (2002, Assumption A3) and the bandwidth parameter $b \in (0, \infty)$ satisfying $1/b + b^2/T \rightarrow 0$ as $T \rightarrow \infty$ (which corresponds to Assumption A4 of Jansson, 2002). Secondly, a parametric autoregressive spectral density estimator, $\hat{\lambda}^2 := s_{AR}^2$, as suggested by Berk (1974), where $s_{AR}^2 := \hat{\sigma}_k^2 / \left(1 - \sum_{i=1}^k \hat{\alpha}_i\right)^2$, $\hat{\sigma}_k^2 := T^{-1} \sum e_{k,t}^2$, and k corresponds to the lag order used. It has been suggested by some authors (e.g. Haldrup and Jansson, 2006) that the M tests, when coupled with the modified AIC lag selection method of Ng and Perron (2001), are preferable to standard ADF tests due to their superior size properties, relative to the latter, in the presence of weak dependence in $\{\varepsilon_t\}$. Perron and Rodriguez [PR] (2015) have recently introduced these procedures for the no breaks case. Hence, in this section we extend PR by allowing for the possible presence of breaks. \square

Following LKT, to test for a shift from I(0) to I(1) the test statistics proposed are

$$MZ_{\hat{\rho}}^f(\lambda) := \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \text{ and } MZ_{\hat{\rho}}^f(\lambda) := \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau), \quad (3.11)$$

whereas to test for shifts from I(1) to I(0) time-reversed data, $e_t^r = e_{T-t+1}$, $t = 1, \dots, T$, is used to compute the statistics defined in (3.11) using e_t^r instead. We define the resulting statistics as

$MZ_{\hat{\rho}}^r(\lambda)$ and $MZ_{t_{\hat{\rho}}}^r(\lambda)$, respectively.

Moreover, when the direction of change is unknown, as is typically the case in empirical work, the tests

$$MZ_{\hat{\rho}}(\lambda) := \min \left\{ MZ_{\hat{\rho}}^f(\lambda), MZ_{\hat{\rho}}^r(\lambda) \right\} \quad \text{and} \quad MZ_{t_{\hat{\rho}}}(\lambda) := \min \left\{ MZ_{t_{\hat{\rho}}}^f(\lambda), MZ_{t_{\hat{\rho}}}^r(\lambda) \right\}$$

are used. LKT suggest setting $\lambda = \lambda_0 = 1/T$, and recommend that the sample period covers $t = 1, 2, \dots, \lfloor \tau T \rfloor$, thus starting at the first observation. Moreover, the breakdate is estimated consistently as, $\hat{\tau} := \arg \inf_{\tau \in (\lambda_0, 1]} MZ_{\hat{\rho}}(\lambda_0, \tau) = \arg \inf_{\tau \in (1/T, 1]} MZ_{\hat{\rho}}(1/T, \tau)$ or alternatively $\hat{\tau} := \arg \inf_{\tau \in (\lambda_0, 1]} MZ_{t_{\hat{\rho}}}(\lambda_0, \tau) = \arg \inf_{\tau \in (1/T, 1]} MZ_{t_{\hat{\rho}}}(1/T, \tau)$.

Assuming that there are multiple breaks, for instance, three regimes (I(1) - I(0) - I(1)), then the statistics defined previously are doubly-recursive, i.e., $MZ_{\hat{\rho}}^f(\lambda, \tau) := \inf_{\lambda \in (0, 1)} MZ_{\hat{\rho}}^f(\lambda) = \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}(\lambda, \tau)$ and $MZ_{t_{\hat{\rho}}}^f(\lambda, \tau) := \inf_{\lambda \in (0, 1)} MZ_{t_{\hat{\rho}}}^f(\lambda) = \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}(\lambda, \tau)$. These statistics deliver consistent estimates of τ_1 and τ_2 (the break fractions), and of the beginning and ending of the stationary regimes, by considering, either $(\hat{\lambda}, \hat{\tau}) := \arg \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}(\lambda, \tau)$ or $(\hat{\lambda}, \hat{\tau}) := \arg \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}(\lambda, \tau)$. This will include the case of a single shift. Note that $\hat{\lambda} = 0$ suggests I(0) - I(1); and $\hat{\tau} = 1$ implies I(1) - I(0).

After determining the most prominent I(0) regime, testing for further I(0) regimes can be done sequentially by analysing the subintervals $[0, \hat{\lambda}]$ and $[\hat{\tau}, 1]$ based on $MZ_{\hat{\rho}}^f$ and $MZ_{t_{\hat{\rho}}}^f$. In the subsample corresponding to $[0, \hat{\lambda}]$, the tests may not reject the null hypothesis (suggesting that the series is I(1) over the whole period) or reject it (because of a single change from I(0) to I(1) or from I(1) to I(0) or because of multiple changes I(1) - I(0) - I(1)), and similarly for period $[\hat{\tau}, 1]$. Repeating the procedure for the I(1) sub-periods until all tests do not reject the null, one concludes this testing scheme, identifying τ_1, \dots, τ_m , with m either odd or even, and whether the first regime is I(0) or I(1).

Note that for multiple regimes, the reversed statistics $MZ_{\hat{\rho}}^r(\lambda), MZ_{t_{\hat{\rho}}}^r(\lambda)$ are not required. For example, I(0) - I(1) - I(0) can be detected either (i) by analysing whether there is a change from I(1) to I(0) for $\hat{\tau} = 1$, or (ii) if one concludes for a change from I(0) to I(1) in the I(1) regime.

4 Asymptotic Properties

To characterize the asymptotic properties of the test statistics discussed in the previous section we consider that, under the null hypothesis, ε_t is generated as

$$\varepsilon_t = \varepsilon_{t-1} + v_t, \quad t = 1, \dots, T.$$

where v_t is as defined in Assumption A.... . .

4.1 Wald-type tests

The following theorem characterises the limit distributions of the segmented cointegration tests proposed in (3.2). First, consider (2.1) with no deterministic.

Theorem 4.1 *Under the null hypothesis and Assumptions A1 to A4 it follows, for m^* fixed and even, assuming $p_T = 0$, that as $T \rightarrow \infty$,*

$$F_A(\tau, m^*) \Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}'\Omega_0\mathbf{b}} \sum_{j=1}^{m^*/2} \left\{ \frac{\left[\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) d\mathbf{B}(r)' + (\tau_{2j} - \tau_{2j-1}) \Omega_1 \right) \mathbf{b} \right]^2}{\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) \mathbf{B}^{(2j)}(r)' \right) \mathbf{b}} + \mathbf{b}' \frac{[\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})] [\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})]'}{\tau_{2j} - \tau_{2j-1}} \mathbf{b} \right\}$$

where $\mathbf{B}^{(2j)}(r) := \mathbf{B}(r) - \frac{1}{\tau_{2j} - \tau_{2j-1}} \int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}(r) dr$ is the n -vector Brownian motion of regime $2j$.

Corollary 1 *Under the null hypothesis and Assumptions A.1 to A.4 it follows, for m^* fixed and even, and assuming $p_T = 0$, that as $T \rightarrow \infty$,*

$$F_A(\tau, m^*) \Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}'\Omega_0\mathbf{b}} \sum_{j=1}^{m^*/2} \left\{ \frac{\left[\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}'\Omega_1\mathbf{b} \right]^2}{\omega_{11.2} \left[\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r)^2 \right]} + \omega_{11.2} \frac{[Q(\tau_{2j}) - Q(\tau_{2j-1})]^2}{\tau_{2j} - \tau_{2j-1}} \right\}. \quad (4.1)$$

where $Q := W_1(r) - \left(\int_0^1 W_1(r) \mathbf{W}_2(r)' \right) \left(\int_0^1 \mathbf{W}_2(r) \mathbf{W}_2(r)' \right)^{-1} \mathbf{W}_2(r)$.

Remark 4.1: Under endogeneity, $\omega_{21} \neq 0$, but with ξ_t not autocorrelated, the distribution of $F_A(\tau, m^*)$ is free of nuisance parameters. Note that in this case, $\Omega_1 = 0, \Omega = \Omega_0$, and therefore,

for m^* fixed and even it follows that,

$$F_A(\tau, m^*) \Rightarrow \frac{1}{m^*} \frac{1}{\kappa' \kappa} \sum_{j=1}^{m^*/2} \left\{ \frac{\left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r)^2} + \frac{[Q(\tau_{2j}) - Q(\tau_{2j-1})]^2}{\tau_{2j} - \tau_{2j-1}} \right\}$$

and for m^* fixed and odd

$$F_A(\tau, m^*) \Rightarrow \frac{1}{m^* + 1} \frac{1}{\kappa' \kappa} \sum_{j=1}^{(m^*+1)/2} \left[\frac{\left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r)^2} + \frac{(Q(\tau_{2j}) - Q(\tau_{2j-1}))^2}{\tau_{2j} - \tau_{2j-1}} \right]$$

where Q is as defined in Corolary 4.1 and $\kappa := \left(1, - \left(\int_0^1 W_1(r) \mathbf{W}_2(r)' \right) \left(\int_0^1 \mathbf{W}_2(r) \mathbf{W}_2(r)' \right)^{-1} \right)'$.

Theorem 4.2 Under Assumptions A.1 to A.3, and B.1 it follows under the null hypothesis, as $T \rightarrow \infty$, for m^* fixed and even, that

$$F_B(\tau, m^*) \Rightarrow \frac{1}{m^* + 2} \frac{1}{\kappa' \kappa} \sum_{j=0}^{m^*/2} \left[\frac{\left(\int_{\tau_{2j}}^{\tau_{2j+1}} Q^{(2j+1)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j}}^{\tau_{2j+1}} Q^{(2j+1)}(r)^2} + \frac{(Q(\tau_{2j+1}) - Q(\tau_{2j}))^2}{\tau_{2j+1} - \tau_{2j}} \right]$$

and for m^* fixed and odd

$$F_B(\tau, m^*) \Rightarrow \frac{1}{m^* + 1} \frac{1}{\kappa' \kappa} \sum_{j=0}^{(m^*-1)/2} \left[\frac{\left(\int_{\tau_{2j}}^{\tau_{2j+1}} Q^{(2j+1)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j}}^{\tau_{2j+1}} Q^{(2j+1)}(r)^2} + \frac{(Q(\tau_{2j+1}) - Q(\tau_{2j}))^2}{\tau_{2j+1} - \tau_{2j}} \right].$$

Corollary 2 Under Assumptions A.1 to A.4 it follows under the null hypothesis, as $T \rightarrow \infty$, for m^* fixed and even, that

$$\sup F_k(m^*) := \sup_{\tau \in \Lambda_\varepsilon^{m^*}} \mathfrak{F}_k(\tau, m^*), \quad k = A, B \quad (4.2)$$

converge weakly, \Rightarrow , as $T \rightarrow \infty$, to the $\sup_{\tau \in \Lambda_\varepsilon^{m^*}}$ of the correspondent limiting laws defined above. The same applies to $W(m^*) := \max[\sup \mathfrak{F}_A(m^*), \sup \mathfrak{F}_B(m^*)]$ and

$$W \max := \max_{1 \leq m \leq \bar{m}} W(m). \quad (4.3)$$

Remark 4.2: Assume that the standard assumptions hold. Let $i = 1, 2$. In the case of model (2.1) with an intercept, replace $W_i(r)$ by $\bar{W}_i(r) = W_i(r) - \int_0^1 W_i(r) dr$ (demeaned Brownian

motions) and for model (2.1) with an intercept and a linear trend, replace $W_i(r)$ by $\widetilde{W}_i(r) = W_i(r) - \int_0^1 W_i(r) dr - 12 \left(r - \frac{1}{2}\right) \int_0^1 \left(r - \frac{1}{2}\right) W_i(r) dr$ (demeaned and detrended Brownian motions). This is a standard result (for example, Park and Phillips, 1988) and easy to show using least squares regression. The model (2.1) with an intercept is equivalent to model (2.1) without deterministic but with all variables demeaned. For the second case, variables need also to be detrended.

Table 1: Critical values for the Wald type tests

		no deterministic					demeaned					detrended				
		$W(1)$	$W(2)$	$W(3)$	$W(4)$	W_{\max}	$W(1)$	$W(2)$	$W(3)$	$W(4)$	W_{\max}	$W(1)$	$W(2)$	$W(3)$	$W(4)$	W_{\max}
n=2	90%	8.229	8.362	7.168	6.804	9.677	8.050	8.279	7.069	6.895	9.536	8.373	8.527	7.279	7.152	9.946
	95%	9.367	9.329	7.956	7.790	11.033	9.106	9.277	7.764	7.731	10.762	9.810	9.755	8.229	8.311	11.580
	97.5%	10.615	10.334	8.901	8.932	12.499	10.308	10.269	8.684	8.930	12.025	11.638	11.092	9.531	9.561	13.631
	99%	12.349	11.958	10.574	11.129	15.016	12.089	11.428	10.329	11.083	14.670	15.592	12.568	11.884	12.126	17.734
n=3	90%	7.812	8.216	6.872	6.657	9.403	7.711	8.036	6.830	6.664	9.214	7.666	8.085	6.863	6.713	9.298
	95%	8.915	9.165	7.660	7.612	10.539	8.761	9.090	7.645	7.619	10.552	8.756	9.070	7.597	7.792	10.430
	97.5%	9.829	9.942	8.583	8.575	12.085	9.661	10.050	8.405	8.602	11.797	9.843	10.110	8.467	9.017	11.671
	99%	11.476	11.298	9.972	10.942	13.887	11.103	11.470	9.855	10.713	13.973	11.365	11.222	9.612	11.072	13.690
n=4	90%	7.529	7.988	6.664	6.492	9.131	7.669	7.913	6.598	6.499	9.093	7.588	7.985	6.659	6.588	9.208
	95%	8.542	8.903	7.450	7.495	10.459	8.628	8.852	7.281	7.475	10.330	8.589	9.049	7.366	7.688	10.375
	97.5%	9.615	9.851	8.293	8.661	11.795	9.721	9.816	8.199	8.738	11.597	9.471	10.000	8.148	9.003	11.628
	99%	11.026	11.217	9.967	10.835	13.941	10.707	11.419	9.374	10.905	13.785	10.870	11.291	9.423	11.085	13.334
n=5	90%	7.546	7.942	6.516	6.393	8.952	7.994	7.928	6.658	6.418	9.194	7.947	7.903	6.646	6.435	9.279
	95%	8.448	8.921	7.171	7.530	9.989	8.936	8.936	7.364	7.449	10.407	9.000	8.969	7.289	7.528	10.461
	97.5%	9.320	9.734	7.915	8.750	11.180	9.876	9.801	8.104	8.683	11.867	10.037	9.785	8.024	8.784	12.084
	99%	10.718	10.741	9.175	10.640	13.109	11.179	11.204	9.720	11.017	13.893	11.751	11.230	9.500	11.696	14.375
n=6	90%	7.857	7.882	6.553	6.374	9.151	8.452	8.023	6.740	6.375	9.591	8.330	7.890	6.641	6.458	9.443
	95%	8.772	8.832	7.235	7.397	10.425	9.616	8.942	7.450	7.323	10.754	9.412	8.767	7.363	7.454	10.676
	97.5%	9.733	9.875	8.003	8.740	11.803	10.667	9.864	8.199	8.511	11.806	10.398	9.755	8.048	9.080	12.028
	99%	11.029	10.910	9.127	11.314	13.336	11.793	11.028	9.480	10.701	13.379	11.916	10.989	9.111	11.956	13.747

4.2 Sequential type test

Regarding the statistics in (3.8) and (3.9) it follows that:

Theorem 4.4 *Under the same assumptions of Theorem 1, for an $I(0) - I(1)$ shift it follows as $T \rightarrow \infty$, that*

$$MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \left(2 \int_{\lambda}^{\tau} Q(r)^2\right)^{-1} \left[\frac{\tau Q(\tau)^2}{(\tau - \lambda)} - \kappa' \kappa\right] =: \mathcal{MZ}_{\hat{\rho}}^{GLS}(\lambda, \tau), \quad (4.4)$$

$$MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \left(2 \left((\kappa' \kappa) \int_{\lambda}^{\tau} Q(r)^2\right)^{1/2}\right)^{-1} \left[\frac{\tau Q(\tau)^2}{(\tau - \lambda)} - \kappa' \kappa\right] =: \mathcal{MZ}_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \quad (4.5)$$

whereas for an $I(1)$ to $I(0)$ shift, considering time-reversed data, $\tilde{e}_t = e_{T-t+1}$, $t = 1, \dots, T$, so that

$$\begin{aligned} MZ_{\hat{\rho}}^{GLS,r}(\lambda, \tau) &\Rightarrow \left(\frac{2}{(\tau - \lambda)} \int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr\right)^{-1} \left[\frac{(1-\tau)}{(\tau - \lambda)} Q(1-\tau)^2 - \kappa' \kappa\right] \\ &=: \mathcal{MZ}_{\hat{\rho}}^{GLS,r}(\lambda, \tau) \end{aligned} \quad (4.6)$$

$$\begin{aligned} MZ_{t_{\hat{\rho}}}^{GLS,r}(\lambda, \tau) &\Rightarrow 2 \left(\frac{\kappa' \kappa}{(\tau - \lambda)} \int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr\right)^{-1/2} \left[\frac{(1-\tau)}{(\tau - \lambda)} Q(1-\tau)^2 - \kappa' \kappa\right] \\ &=: \mathcal{MZ}_{t_{\hat{\rho}}}^{GLS,r}(\lambda, \tau) \end{aligned} \quad (4.7)$$

where Q is as defined in Corolary 4.1 and $\kappa := \left(1, -\left(\int_0^1 W_1(r) \mathbf{W}_2(r)'\right) \left(\int_0^1 \mathbf{W}_2(r) \mathbf{W}_2(r)'\right)^{-1}\right)'$.

Remark 4.3: For $n = 1$, it follows from (4.4) and (4.5) that, $MZ_{\hat{\rho}}(\lambda, \tau) \Rightarrow \left(2 \int_{\lambda}^{\tau} W(r)^2 dr\right)^{-1} \left(\frac{\tau W(\tau)^2}{(\tau - \lambda)} - 1\right)$; and $MZ_{t_{\hat{\rho}}}(\lambda, \tau) \Rightarrow \left(4 \int_{\lambda}^{\tau} W(r)^2 dr\right)^{-1/2} \left(\frac{\tau W(\tau)^2}{(\tau - \lambda)} - 1\right)$, which is an alternative to LKT, and for $\tau = 1$ and $\lambda = 0$, $MZ_{\alpha}(\lambda, \tau)$ and $MZ_t(\lambda, \tau)$ are the Perron and Ng (1996)'s distributions, MZ_{α} and MZ_t . If $n > 1$, $\tau = 1$ and $\lambda = 0$ we have Pesavento and PR stats. \square

Theorem 4.5 *Under the assumptions of Theorem 1 the limit distributions of the statistics for segmented cointegration, when the shift is $I(0) - I(1)$, are,*

$$\begin{aligned} MZ_{\hat{\rho}}^f(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \mathcal{MZ}_{\hat{\rho}}^{GLS}(\lambda, \tau) \\ MZ_{t_{\hat{\rho}}}^f(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \mathcal{MZ}_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \end{aligned}$$

whereas for $I(1) - I(0)$,

$$\begin{aligned} MZ_{\hat{\rho}}^r(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{\hat{\rho}}^{GLS, r}(\lambda, \tau) \\ MZ_{t_{\hat{\rho}}}^r(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{t_{\hat{\rho}}}^{GLS, r}(\lambda, \tau) \end{aligned}$$

and

$$\begin{aligned} MZ_{\hat{\rho}}(\lambda) &= \min \left\{ MZ_{\hat{\rho}}^f(\lambda), MZ_{\hat{\rho}}^r(\lambda) \right\} \Rightarrow \min \left\{ \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{\hat{\rho}}^{GLS}(\lambda, \tau), \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{\hat{\rho}}^{GLS, r}(\lambda, \tau) \right\} \\ MZ_{t_{\hat{\rho}}}(\lambda) &= \min \left\{ MZ_{t_{\hat{\rho}}}^f(\lambda), MZ_{t_{\hat{\rho}}}^r(\lambda) \right\} \Rightarrow \min \left\{ \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau), \inf_{\tau \in (\lambda, 1]} \mathcal{M}Z_{t_{\hat{\rho}}}^{GLS, r}(\lambda, \tau) \right\} \end{aligned}$$

Finally, $MZ_{\alpha}^f \Rightarrow \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau)$ and $MZ_t^f \Rightarrow \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau)$.

When GLS residuals are considered, i.e.,

$$\tilde{e}_t = \hat{\mathbf{b}}' \mathbf{x}_t - \hat{\delta}_{\hat{\rho}} = \hat{\mathbf{b}}' \left(\mathbf{x}_t - \frac{\sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\mathbf{x}_t - \bar{\rho} \mathbf{x}_{t-1})}{(1 - \bar{\rho}) \lfloor (\tau - \lambda) T \rfloor} \right),$$

Since,

$$\begin{aligned} \frac{1}{\lfloor (\tau - \lambda) T \rfloor^{1/2} \hat{\delta}_{\hat{\rho}}} &= \frac{1}{(1 - \bar{\rho}) \lfloor (\tau - \lambda) T \rfloor^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\hat{\mathbf{b}}' \mathbf{x}_t - \bar{\rho} \hat{\mathbf{b}}' \mathbf{x}_{t-1}) \\ &= \hat{\mathbf{b}}' \frac{1}{(1 - \bar{\rho})} \left\{ \frac{1}{\lfloor (\tau - \lambda) T \rfloor^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \mathbf{x}_t - \frac{\bar{\rho}}{\lfloor (\tau - \lambda) T \rfloor^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \mathbf{x}_{t-1} \right\} \\ &\Rightarrow \mathbf{b}' \frac{1}{(1 - \bar{\rho})} \{ \mathbf{B}(\tau) - \mathbf{B}(\lambda) - \bar{\rho} [\mathbf{B}(\tau) - \mathbf{B}(\lambda)] \} = \mathbf{b}' [\mathbf{B}(\tau) - \mathbf{B}(\lambda)] \end{aligned}$$

and

$$\frac{1}{\lfloor (\tau - \lambda) T \rfloor^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\hat{\mathbf{b}}' \mathbf{x}_{t-1} - \hat{\delta}_{\hat{\rho}})^2 \Rightarrow \mathbf{b}' \left\{ \int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' - [\mathbf{B}(\tau) - \mathbf{B}(\lambda)] [\mathbf{B}(\tau) - \mathbf{B}(\lambda)]' \right\} \mathbf{b}.$$

the following results can be stated.

Theorem 4.6

$$MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \frac{\left\{ \left(\frac{\tau}{\tau - \lambda} \right)^{1/2} Q(\tau) - [Q(\tau) - Q(\lambda)] \right\}^2 - \kappa' \kappa}{2 \left\{ \int_{\lambda}^{\tau} Q^2 - [Q(\tau) - Q(\lambda)]^2 \right\}}$$

and

$$MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \frac{\left\{ \left(\frac{\tau}{\tau - \lambda} \right)^{1/2} Q(\tau) - [Q(\tau) - Q(\lambda)] \right\}^2 - \kappa' \kappa}{2(\kappa' \kappa)^{1/2} \left\{ \int_{\lambda}^{\tau} Q^2 - [Q(\tau) - Q(\lambda)]^2 \right\}^{1/2}}.$$

Note that $s(\lambda, \tau)^2$ has the same properties as before since $\Delta \tilde{e}_t = \Delta(e_t - \hat{\delta}_{\hat{\rho}}) = \Delta e_t$.

We take the same values for \bar{c} as in PR (p.93 Table 1) for the GLS detrended residuals. Further analysis is needed, but for now we assume that this choice of \bar{c} is also adequate in our context. Using \bar{c} we simulate the limit distributions of the tests under the null hypothesis that $c = 0$ to obtain the critical values. Under the null it is expected that the distributions do not depend on nuisance parameters.

Table 2: Critical Values for the LKT Tests

	LKT (0-1/1-0), Model 1					LKT (Multiple), Model 1				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.685	-3.971	-4.326	-4.589	-4.853	-3.857	-4.247	-4.568	-4.832	-5.105
2.5%	-3.392	-3.711	-4.058	-4.328	-4.589	-3.599	-3.962	-4.269	-4.558	-4.853
5.0%	-3.126	-3.472	-3.799	-4.096	-4.363	-3.360	-3.713	-4.049	-4.325	-4.612
7.5%	-2.962	-3.333	-3.634	-3.940	-4.213	-3.197	-3.540	-3.886	-4.178	-4.450
10.0%	-2.840	-3.217	-3.518	-3.818	-4.090	-3.068	-3.430	-3.773	-4.050	-4.341
15.0%	-2.663	-3.045	-3.351	-3.655	-3.918	-2.875	-3.264	-3.597	-3.869	-4.160
20.0%	-2.521	-2.907	-3.213	-3.510	-3.793	-2.735	-3.120	-3.451	-3.729	-4.027

	LKT (0-1/1-0), Model 2					LKT (Multiple), Model 2				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.596	-3.960	-4.313	-4.540	-4.871	-3.849	-4.194	-4.565	-4.787	-5.185
2.5%	-3.322	-3.703	-4.052	-4.307	-4.575	-3.544	-3.941	-4.265	-4.549	-4.879
5.0%	-3.084	-3.454	-3.831	-4.055	-4.348	-3.329	-3.736	-4.033	-4.340	-4.644
7.5%	-2.931	-3.302	-3.683	-3.915	-4.209	-3.187	-3.570	-3.881	-4.174	-4.483
10.0%	-2.811	-3.191	-3.562	-3.810	-4.092	-3.073	-3.437	-3.772	-4.061	-4.367
15.0%	-2.651	-3.042	-3.393	-3.643	-3.911	-2.872	-3.255	-3.611	-3.873	-4.173
20.0%	-2.506	-2.901	-3.254	-3.513	-3.779	-2.726	-3.118	-3.458	-3.748	-4.029

	LKT (0-1/1-0), Model 3					LKT (Multiple), Model 3				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.942	-4.196	-4.507	-4.815	-5.130	-4.201	-4.491	-4.834	-5.098	-5.409
2.5%	-3.690	-3.961	-4.258	-4.556	-4.830	-3.939	-4.219	-4.554	-4.841	-5.124
5.0%	-3.471	-3.779	-4.050	-4.332	-4.588	-3.670	-4.018	-4.324	-4.580	-4.867
7.5%	-3.326	-3.627	-3.909	-4.176	-4.439	-3.518	-3.866	-4.152	-4.438	-4.707
10.0%	-3.209	-3.517	-3.790	-4.054	-4.323	-3.395	-3.724	-4.063	-4.313	-4.597
15.0%	-3.029	-3.345	-3.612	-3.898	-4.153	-3.217	-3.551	-3.879	-4.134	-4.407
20.0%	-2.891	-3.212	-3.485	-3.749	-4.016	-3.078	-3.401	-3.744	-3.998	-4.277

Note: (PR: $p_y = 0$) and \mathbf{x}_{2t} with trend (PR: $p_x = 1$) : GLS: Same W_1 and replace W_2 by $(W'_{2,n-2}, r)'$

Note:(PR: $p_y = 1$) and \mathbf{x}_{2t} with trend (PR: $p_x = 1$) (according to PR same as $p_y = 1, p_x = 0$) : GLS: Replace W_1 by

$$W_1 - \left[\varphi W_1(1) + 3(1 - \varphi) \int_0^1 s W_1(s) ds \right] r, \text{ where } \varphi = \frac{1 - \bar{c}}{1 - \bar{c} + \bar{c}^2/3}$$

and W_2 by

$$W_2 - \left[\varphi W_2(1) + 3(1 - \varphi) \int_0^1 s W_2(s) ds \right] r, \text{ where } \varphi = \frac{1 - \bar{c}}{1 - \bar{c} + \bar{c}^2/3}$$

5 Monte Carlo

In this section we evaluate the finite sample performance of the tests introduced in the previous section.

Our DGP is essentially the same as in Perron and Rodriguez (2016). We consider a system given by

$$y_t = \mathbf{x}_t + u_t \quad (5.1)$$

$$u_t = \rho u_{t-1} + v_{2t} \quad (5.2)$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_{1t} \quad (5.3)$$

$$\mathbf{v}_t = A\mathbf{v}_{t-1} + \epsilon_t \quad (5.4)$$

where $\epsilon_t \sim iidN(0, \Sigma)$ with Σ such that the long-run variance of \mathbf{v}_t , $\Omega := (I - A)^{-1}\Sigma((I - A)^{-1})'$, is $\Omega = \begin{bmatrix} 1 & R \\ R & 1 \end{bmatrix}$, with $\mathbf{A} := \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

Same DGP as in PR, p.13, table 5 (Model (??), $p_y = 0$ and $p_x = 0$), $c = 0$. Also 5000 replications. $MZ_{\hat{\rho}}^{GLS}$ and $MZ_{t\hat{\rho}}^{GLS}$ tests with GLS detrended data taking \bar{c} as in PR (their Table 1) and using OUR critical values. Model (??), $p_y = 0$ and $p_x = 0$. We also consider $A_0 = 0$, $A_1 = \begin{bmatrix} 0.0 & 0.9 \\ 0.0 & 0.0 \end{bmatrix}$, $A_2 = \begin{bmatrix} 0.5 & 0.0 \\ 0.2 & 0.3 \end{bmatrix}$ and $A_3 = \begin{bmatrix} 0.5 & 0.0 \\ 0.2 & 0.7 \end{bmatrix}$.

5.1 Size

[TO BE COMPLETED]

The results are in the Appendix - Tables 10 and 11 in subsection 9.1. Size.

Main Conclusion: KPZ with good size for most of the DGPs; LKT/PR not so much.

5.2 Power

[TO BE COMPLETED]

The results are in the Appendix - Tables 12 up to 27 in subsection 9.2. Power. Bellow Table 27, there are several other Tables regarding the breakpoint estimates.

Main Conclusion: KPZ is the best procedure in terms of power; Largest power need I(0) dominant regime, as expected.

6 Empirical Application - US Government's Budgets

6.1 Motivation

The consistency of fiscal policy, the sustainability of the government deficit and how it relates to discounted or undiscounted debt has long been subject to intense scrutiny. Empirical applications using models with coefficients that are subject to structural breaks and variables such as the primary surplus to GDP and debt to income ratio include Martin (2000), Camarero, Carrion-i-Silvestre, and Tamarit (2015), and Nguyen, Suardi, and Chua (2017), among others. For example, Camarero, Carrion-i-Silvestre, and Tamarit (2015) build upon the concept of multicointegration to show that there exists weak fiscal sustainability and cointegration between deficit and debt for several OECD countries.

In this article we do not discuss sustainability but rather examine thoroughly the main components of the US Federal, and State and Local government's budget by means of the segmented cointegration methodology presented above. Some authors considered standard panel cointegration between revenues and expenditures studying the panel of 47/48 state-local government units (see, for example, Mahdavia and Westerlund, 2011, and Saunoris, 2015). Chen (2016) take a different approach by considering a quantile-dependent cointegrating relationship between government expenditures and revenues. Instead, we allow for the possibility of regime changes in cointegration between budget components and we do it for an extended amount of different budgetary variables for the US economy.

6.2 Data

We use data from two sources: The Office of Management and Budget (OMB), the largest office within the Executive Office of the President of the United States (EOP), and the U.S. Bureau of Economic Analysis (BEA), which produces the national income and product accounts (NIPAs).

The OMB provides the longest history of annual time series data. The "Historical Tables" provide data on budget receipts, outlays (expenditures), surpluses or deficits over an extended time period, generally from 1940 or earlier to 2017. Additionally, Table 1.1 - Federal Government provides data on total receipts, and total outlays, in millions of dollars, for 1901–1939 and for earlier multi-year periods (1789–1849 and 1850–1900).

The BEA provides data for several budget variables, Federal (F), State and Local (SL) Governments, and at a higher frequency basis, namely quarterly. We consider the "NIPA tables" from Section 3 - Government Current Receipts and Expenditures, quarterly data, and generally from

1947Q1 to 2016q3. For those variables which are only measured in nominal terms, we obtain real values by using the Implicit Price Deflator for GDP (2009=100, seasonally adjusted, BEA Account Code: A191RD3 from Section 1 - Domestic Product and Income). The list of variables include current receipts (F and SL, codes W005RC1 and W023RC1, respectively), current expenditures (F and SL, codes W013RC1 and W024RC1, respectively), total receipts (F and SL, codes W018RC1 and W077RC1, respectively), total expenditures (F and SL, codes W019RC1 and W079RC1, respectively), current tax receipts (F, code W006RC1), contributions for government social insurance (F, code W780RC1), personal current taxes (F and SL, codes A074RC1 and W071RC1, respectively), taxes on corporate income (F, code B075RC1), taxes on production and imports (SL, code W072RC1), current transfer receipts (SL, code W075RC1), consumption expenditures (F and SL, codes A957RC1 and A991RC1, respectively), current transfer payments (F and SL, codes W014RC1 and LA0000171, respectively), interest payments (F and SL, codes A091RC1 and B111RC1, respectively), gross investment (F and SL, codes A787RC1 and A799RC1, respectively), national defense (F, code A824RC1), and nondefense (F, code A825RC1), all in billions of dollars and seasonally adjusted at annual rates.

6.3 Results

We present results using annual data from 1901 to 2015 (115 observations) and quarterly data from 1969Q1 to 2016Q3 (191 data points), the moment former President Barack Obama leaves office. The period prior to 1969 is not considered in order to prevent the existence of a structural break that is solely due to a change of the data collection method.¹ Moreover, we show the results for the Wald-type tests (KPZ) which are not sequential and possess better statistical properties than the LKT/PR, according to the results in the Monte Carlo section. The trimming parameter ϵ is equal to 0.15, a standard value in this type of literature, and the maximum number of breaks allowed is 4. We present also the results using the variables in nominal terms because that is the usual way they are measured given the difficulty in finding a single deflator for all different budget definitions. Still, we computed the proposed tests for the variables in real terms using the implicit price deflator for GDP and the results do not change significantly (these are available upon request). Having all variables in nominal terms, we apply our tests in models (2.1) with an intercept (Model 2,

¹From the OMB: "The Federal Government has used the unified or consolidated budget concept as the foundation for its budgetary analysis and presentation since the 1969 Budget. The basic guidelines for the unified budget were presented in the Report of the President's Commission on Budget Concepts (October 1967). The Commission recommended the budget include all Federal fiscal activities unless there were exceptionally persuasive reasons for exclusion." Also, the BEA splits the data in two different excel files, prior and after 1969, so we took the one with most recent period. We tried with data starting in 1947Q1 but obtained some incongruent results.

demeaned) and (2.1) with an intercept and a linear trend (Model 3, demeaned and detrended). For the cases where the choice for the dependent variable is not obvious, we apply the tests considering all possible situations in this regard.

To test for nonstationarity of y_t and x_t , we employ the efficient unit root tests proposed by Elliot, Rothenberg and Stock (1996) and Ng and Perron (2001). These include the ERS and DF-GLS (GLS detrending) tests and the efficient versions of the modified PP tests of Perron and Ng (1996). We find evidence of nonstationarity for all variables with the exception of the F total receipts with annual data (two unit roots). For the following cases, evidence of a unit root process was not so obvious: F total expenditures with annual data and SL current transfer payments (evidence not found using modified PP tests); F and SL current and total expenditures with quarterly data, F current transfer payments, and F national defense (results of the DF-GLS tests); SL consumption expenditures (results of the ERS tests).

In this empirical exercise, we aim to do a complete analysis of the US government's budget historical situation. We start with the overall budget results and then focus on each main component: first, we study cointegration between the surplus/deficit at the Federal and at the State and Local levels; second, comparing Federal to State and Local, we look at cointegration between receipts and expenditures; third, comparing Federal to State and Local, we analyse cointegration between the main components of the receipts; finally, we do the same for the main components of the expenditures. The results are in Tables ?????? and ??????????????????

Table 3: Wald type tests - US Government's Budgets

<i>y</i>	<i>W</i> max	Model 2			<i>W</i> max	Model 3		
		R1	\hat{m}	$\hat{\tau}$		R1	\hat{m}	$\hat{\tau}$
			Surplus or Deficit					
Current SL	21.72***	I(1)	2	2001Q3,2002Q4	19.75***	I(1)	1	2002Q4
Total SL	12.80**	I(1)	2	2001Q1,2002Q4	16.90**	I(1)	1	2009Q2
			Receipts and Expenditures (Outlays)					
Total F Exp. (A)	29.62***	I(1)	1	1997		I(1)	1	1997
Current F Exp. (Q)	12.08**	I(1)	1	2000Q3	18.60***	I(1)	1	2007Q3
Total F Exp. (Q)	10.03*	I(1)	2	2001Q4, 2009Q1	20.63***	I(1)	1	2006Q3
Current SL Exp. (Q)			no breaks				no breaks	
Total SL Exp. (Q)	10.15*	I(1)	1	2009Q2			no breaks	
			Receipts					
Personal taxes (F)	17.74***	I(1)	1	2001Q1		I(1)	1	2001Q1
Corporate taxes (F)	25.32***	I(1)	1	2008Q4		I(1)	1	2008Q4
Social insurance (F)	9.74*	I(1)	1	2009Q2	13.94***	I(1)	1	2009Q1
Personal taxes (SL)			no breaks				no breaks	
Prod. and imp. taxes (SL)	23.97***	I(1)	2	2001Q1, 2008Q4		I(1)	2	2001Q1, 2008Q4
Transfer receipts (SL)	25.58***	I(1)	1	2008Q2		I(1)	1	2008Q2

Notes: *,**,*** denote 10%,5%,1% significant levels; R1 stands for the 1st regime; F for Federal; SL for State and Local; A for Annual; Q for Quarterly.

Table 4: wald type tests - US Government's Budgets (continued)

<i>y</i>	<i>W</i> max	Model 2			<i>W</i> max	Model 3		
		R1	\hat{m}	$\hat{\tau}$		R1	\hat{m}	$\hat{\tau}$
			Expenditures					
midrule Cons. expend. (F)	44.00***	I(1)	2	2002Q2,2008Q3		I(1)	2	2002Q2,2008Q3
Transfer pay. (F)	47.10***	I(1)	2	2002Q2,2008Q3		I(1)	2	2002Q2,2008Q3
Interest pay. (F)	10.00*	I(1)	1	2000Q3	30.45***	I(1)	1	2008Q4
Cons. expend. (SL)	13.89**	I(1)	1	2008Q2		same as Model 2		
Transfer pay. (SL)	15.52***	I(1)	2	1993Q2,2009Q2	13.52**	I(1)	1	2009Q2
Interest pay. (SL)	14.60***	I(1)	2	2000Q2,2009Q2		I(1)	2	2000Q2,2009Q2
Cons. expend. (F)			no breaks				no breaks	
Gross Invest. (F)			no breaks				no breaks	
Cons. expend. (SL)	13.49**	I(1)	2	1997Q4,2009Q1	15.30***	I(1)	2	1996Q1,2009Q1
Gross Invest. (SL)	13.70**	I(1)	2	1998Q1,2009Q1	13.89***	I(1)	2	1997Q4,2009Q1
Defense			no breaks				no breaks	
Nondefense	9.84640*	I(1)	3	1991Q1,2002Q2 2009Q2		no breaks		

Notes: *,**,*** denotes significance at the 10%,5%,1% significant levels; R1 stands for the 1st regime; F for Federal; SL for State and Local.

Surplus or Deficit First, we regress the surplus or deficit at the State and Local (SL) level on the Federal's (F) surplus or deficit. For each model, the results using current or total values are about the same. For the demeaned model we find two breaks and cointegration during years 2001 and 2002, whereas for the demeaned and detrended model we find a single break with cointegration since 2002Q4, for current, or 2009Q2, for total. That is, for model 2 State and Local and Federal's surplus or deficit were almost never tied together (only happened during the first two years of George W. Bush as president (Republican)), and for model 3 ever since George W. Bush's, for current, or Barack Obama's (Democratic), for total.

Receipts and Expenditures (Outlays) Second, we regress expenditures on receipts. We do it for Federal and for quarterly State and Local data. The Federal's annual total expenditures and total receipts are only cointegrated after 1997, the second term of Bill Clinton as president (Democratic), whereas for quarterly data it happens since 2001 (model 2) or 2006 (model 3), the George W. Bush's presidency. On the contrary, the State and Local's expenditures and receipts are never cointegrated.

Receipts Next, we analyse the main components of the receipts at the Federal level (personal current taxes, taxes on corporate income, and contributions for government social insurance²) and at the State and Local level (personal current taxes, taxes on production and imports, and current transfer receipts³). For the case of the Federal government we find a strong evidence of a single break with cointegration during the second regime. The date of the break is not unanimous being either 2001 (cointegration since George W. Bush's legacy) or 2009 (Barack Obama's). With respect to State and Local the results depend on the choice for the dependent variable. It can be two breaks, one break or even no breaks! When the dependent variable is the current transfer receipts the conclusion is similar to the Federal case: single break and cointegration since Barack Obama's presidency.

Expenditures Finally, we study the main components of the expenditures. For both Federal and State and Local, we first take the decomposition consumption expenditures, current transfer payments, and interest payments⁴ and, second, only consumption expenditures and gross invest-

²In 2016, 44%, 13%, and 35%, respectively, from the Federal government current receipts.

³In 2016, 17%, 47%, and 29%, respectively, from the State and Local government current receipts.

⁴In 2016, 24%, 64%, and 11%, respectively, from the Federal government current expenditures, and 65%, 27%, and 8% from the State and Local government current expenditures.

ment⁵. We also analyse cointegration between the discretionary (not mandatory) spending of federal consumption expenditures and gross investment in defense and nondefense⁶.

At the Federal level, we observe segmented cointegration between consumption expenditures, current transfer payments, and interest payments. With the exception of model 3 taking interest payments as the dependent variable, cointegration exists during the period 2002 - 2008 (George W. Bush). On the contrary, the null hypothesis of no cointegration throughout the whole period is never rejected for consumption expenditures and gross investment. At a 5% level, the same happens with national defense and nondefense federal consumption expenditures and gross investment.

For the State and Local governments, we also conclude that there is segmented cointegration between consumption expenditures, current transfer payments, and interest payments. The first regime is clearly of no cointegration but the existence of one or two breaks depends on the model and the choice for the dependent variable. The only solid conclusion that one can come to is that until 1993Q2 (Bill Clinton's presidency excluded) there was no cointegration. When focusing solely on consumption expenditures and gross investment we conclude that there are two breaks and cointegration during the period 1998 - 2009, that is, Bill Clinton's second term and George W. Bush's presidency.

Summary of Results A few strong conclusions can be drawn from this empirical application. First, until Bill Clinton's presidency the governments' budget components never moved together. Second, apparently, Federal and State and Local budget decisions seem not to be fully coordinated: their surplus or deficits might had been cointegrated only during George W. Bush and Barack Obama's presidencies and, at each government's level, the periods in which (no) cointegration between expenditures and receipts occurred barely coincide. Third, we observe an effort undertaken by the Federal government to adjust expenditures to receipts since George W. Bush's presidency whereas at the State and Local level the two never co-move. Fourth, and at the Federal's level, the main components of the receipts have been cointegrated since George W. Bush's presidency and those of the expenditures only during that legacy. But when we split expenditures in consumption and gross investment or in national defense and nondefense, cointegration for each pair is never found.

⁵In 2016, 79% is consumption expenditures and 21% is gross investment for Federal and 83% and 17%, respectively, for State and Local.

⁶In 2016, 59% is for national defense and 41% for nondefense.

7 Conclusion

In this paper, we have provided tests for segmented cointegration, i.e., tests for multiple changes in the long-run equilibrium of time series, the corresponding large sample distributions and an in depth Monte Carlo analysis of their finite sample behavior. In particular, the test proposed are Wald-type procedures in line with the persistence change tests of Kejriwal, Perron and Zhou (2012) and the M-class of tests proposed by Leybourne, Kim and Taylor (2007) and Perron and Rodriguez (2015).

In the Monte Carlo comparison of the finite sample behavior of the procedures we observe that the new Wald-type tests for changes in the cointegrating relationship introduced in this paper have superior finite sample size and power performance than the M-class of tests also proposed here.

From an empirical perspective the new tests put forward are important in that they help validate the stability of cointegration relationships frequently used in applied work.

References

- [1] Andrade, P., Bruneau, C. and Gregoir, S. (2005), Testing for the Cointegration Rank When Some Cointegrating Directions are Changing,” *Journal of Econometrics*, 124, 269-310.
- [2] Ang, A. and Bekaert, G. (2002), Regime Switches in Interest Rates,” *Journal of Business and Economic Statistics*, 20, 163-182.
- [3] Beyer, A., Haug, A. and Dewald, W. (2009), Structural Breaks, Cointegration and the Fisher Effect,” ECB working paper 1013.
- [4] Bierens, H.J. and Martins, L.F. (2010), Time Varying Cointegration,” *Econometric Theory*, 26, 1453-1490.
- [5] Camarero, M., Carrion-i-Silvestre, J.L., and Tamarit, C. (2015) The Relationship Between Debt Level and Fiscal Sustainability in Organization for Economic Cooperation and Development Countries, *Economic Inquiry*, 53, 129-149.
- [6] Chen, P.-F. (2016), US Fiscal Sustainability and the Causality Relationship between Government Expenditures and Revenues: A New Approach Based on Quantile Cointegration, *Fiscal Studies*, 37, 301-320.
- [7] Davidson, J. and A. Monticini (2010), Tests for cointegration with structural breaks based on subsamples, *Computational Statistics and Data Analysis* 54(11), 2498-2511.
- [8] Elliot, G., T.J. Rothenberg, and J.H. Stock (1996). “Efficient Tests for an Autoregressive Unit Root,” *Econometrica*, 64, 813-836.
- [9] Garratt, A., Lee, K., Pesaran, M.H. and Shin, Y. (2003), A Long-run Structural Macroeconomic Model of the UK Economy,” *Economic Journal*, 113, 412-455.
- [10] Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford: Oxford University Press.
- [11] Kejriwal, M., Perron, P. & Zhou, J.. Wald Tests for Detecting Multiple Structural Changes in Persistence. *Econometric Theory*.
- [12] Kejriwal, M. & Perron, P.. A Note on Estimating a Structural Change in Persistence. *Economics Letters*.

- [13] Lettau, M. and Ludvigson, S. (2004), Understanding Trend and Cycle in Asset Values: Reevaluating the Wealth Effect on Consumption,” *American Economic Review*, 94, 276-299.
- [14] Mahdavia, S., Westerlund, J. (2011) Fiscal stringency and fiscal sustainability: Panel evidence from the American state and local governments, *Journal of Policy Modeling*, 33, 953-969.
- [15] Martin, G. (2000), US Deficit Sustainability: A New Approach Based on Multiple Endogenous Breaks,” *Journal of Applied Econometrics*, 15, 83-105.
- [16] Michael, P., Nobay, A. and Peel, D. (1997), Transactions Costs and Nonlinear Adjustment in Real Exchange Rates: An Empirical Investigation,” *Journal of Political Economy*, 105, 862-879.
- [17] Ng, S., and P. Perron (2001). “Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power,” *Econometrica*, 69, 1519-1554.
- [18] Nguyen, T.D., Suardi, S., and Chua, C.L., (2017) The Behavior of U.S. Public Debt and Deficits During the Global Financial Crisis, *Contemporary Economic Policy*, 35, 201-215.
- [19] Park, J., and Hahn, H. (1999), Cointegrating Regressions with Time Varying Coefficients,” *Econometric Theory*, 15, 664-703.
- [20] Perron, P. and S. Ng. (1996). “Useful Modifications to Some Unit Root Tests with Dependent Errors and their Local Asymptotic Properties,” *Review of Economic Studies*, 63, 435-463.
- [21] Quintos, C.E. (1997), Stability Tests in Error Correction Models,” *Journal of Econometrics*, 82, 289-315.
- [22] Saikkonen, P. and Choi, I. (2004), Cointegrating Smooth Transition Regressions,” *Econometric Theory*, 20, 301-340.
- [23] Saunoris, J.W. (2015) The Dynamics of the Revenue– Expenditure Nexus: Evidence from US State Government Finances, *Public Finance Review*, 43(1), 108-134.
- [24] Sims, C. and Zha, T. (2006), Were There Regime Switches in Macroeconomic Policy?” *American Economic Review*, 96, 54-81.
- [25] Stock, J. and Watson, M. (1996), Evidence on Structural Instability in Macroeconomic Time Series Relations,” *Journal of Business and Economic Statistics*, 14, 11-30.

A Technical Appendix

A.1 Preliminary Results

Considering ξ_t as defined in Assumption $\mathcal{A}.4$ and $\varepsilon_0 = 0$, the following multivariate FCLT can be stated,

$$T^{-1/2} \sum_{t=1}^{[rT]} \xi_t \Rightarrow \boldsymbol{\Omega}^{1/2} \mathbf{W}(r) =: \mathbf{B}(r),$$

where $W(r) := (W_y(r), \mathbf{W}_x(r)')'$, $W_y(r)$ is a standard Wiener process, $W_x(r)$ is an $(n-1)$ vector of standard Wiener processes and $\boldsymbol{\Omega} := \begin{pmatrix} \omega_{yy} & \omega'_{yx} \\ \omega_{xy} & \boldsymbol{\Omega}_{xx} \end{pmatrix}$. Thus, $T^{-1/2} \varepsilon_{[rT]} \Rightarrow \omega_{11.2}^{1/2} W_{11.2}(r)$,

where $\omega_{11.2} := \omega_{yy} - \omega'_{xy} \boldsymbol{\Sigma}_{xx}^{-1} \omega_{xy}$ and $W_{11.2}(r) := W_y(r) + \left(\frac{R^2}{1-R^2}\right)^{1/2} W_{xx}^{\&}(r)$, is a scalar Wiener process, $W_{xx}^{\&}(r) := (n-1)^{-1/2} \sum_{i=1}^{n-1} W_i(r)$, n is the number of exogenous regressors considered in (2.1), and R^2 is as defined in Perron and Rodriguez (2015).

Under the null hypothesis $H_0 : c_j = 0$ and $\phi_j = 0$, for all j . Thus, considering $\widehat{\mathbf{b}}' := (1, -\widehat{\beta}')$, where $\widehat{\beta}$ is the OLS estimate of β in (2.1) and $\widehat{\mathbf{b}}' X_t = e_t$ so that under H_0 , $\widehat{\mathbf{b}} \rightarrow \mathbf{b}$, where $\mathbf{b} := (1, -\beta')$. Then, it follows that,

$$\begin{aligned} T^{-2} \sum_{t=1}^T e_t^2 &= T^{-2} \sum_{t=1}^T \widehat{\mathbf{b}}' \mathbf{X}_t \mathbf{X}_t' \widehat{\mathbf{b}} = \widehat{\mathbf{b}}' \left(\frac{1}{T^2} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \right) \widehat{\mathbf{b}} \\ &\Rightarrow \mathbf{b}' \left(\int_0^1 \mathbf{B}(r) \mathbf{B}(r)' dr \right) \mathbf{b}, \end{aligned}$$

where $\mathbf{X}_t := (y_t, \mathbf{x}_t)'$, $B(r) := \begin{bmatrix} B_1(r)_{1 \times 1} \\ \mathbf{B}_2(r)_{(n-1) \times 1} \end{bmatrix}$ is an n -vector Brownian motion with covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_1'$.

To remove the nuisance parameters present in the distributions consider as in Phillips and Ouliaris (1990) that,

$$\boldsymbol{\Omega} := \begin{pmatrix} \omega_{yy} & \omega'_{yx} \\ \omega_{xy} & \boldsymbol{\Omega}_{xx} \end{pmatrix} = \mathbf{L} \mathbf{L}';$$

where $\mathbf{L} := \begin{pmatrix} l_{11} & 0 \\ l_{21} & \mathbf{L}_{22} \end{pmatrix}$ and $l_{11} := (\omega_{yy} - \omega'_{xy} \boldsymbol{\Omega}_{xx}^{-1} \omega_{xy})^{1/2}$; $l_{21} := \boldsymbol{\Omega}_{xx}^{-1/2} \omega_{xy}$; $\mathbf{L}_{xx} := \boldsymbol{\Omega}_{xx}^{1/2}$.

Moreover,

$$\mathbf{B}(r) : = \begin{bmatrix} B_1(r)_{1 \times 1} \\ \mathbf{B}_2(r)_{(n-1) \times 1} \end{bmatrix};$$

$$\mathbf{A} : = \int_0^1 \mathbf{B}(r)\mathbf{B}(r)'dr = \begin{pmatrix} a_{yy} & a'_{xy} \\ a_{xy} & \mathbf{A}_{xx} \end{pmatrix};$$

$$\mathbf{F} : = \int_0^1 \mathbf{W}(r)\mathbf{W}(r)'dr = \begin{pmatrix} f_{yy} & f'_{xy} \\ f_{xy} & \mathbf{F}_{xx} \end{pmatrix};$$

$$\kappa : = (1, -f'_{xy}\mathbf{F}_{xx}^{-1}) = \left(1, -\left(\int_0^1 W_1(r)\mathbf{W}_2(r)'dr\right)\left(\int_0^1 \mathbf{W}_2(r)\mathbf{W}_2(r)dr\right)^{-1}\right);$$

$$Q(r) : = W_1(r) - \left(\int_0^1 W_1(r)\mathbf{W}_2(r)'dr\right)\left(\int_0^1 \mathbf{W}_2(r)\mathbf{W}_2(r)dr\right)^{-1} \mathbf{W}_2(r)$$

$$\omega_{11.2} : = \omega_{yy} - \omega'_{xy}\mathbf{\Omega}_{xx}^{-1}\omega_{xy} = l_{11}^2;$$

$$a_{11.2} : = a_{yy} - a'_{xy}\mathbf{A}_{xx}a_{xy}$$

where $\mathbf{W}(r)$ is a vector of standard Brownian motions. From Phillips and Ouliaris (1990) we further note that $\mathbf{B}(r) = \mathbf{L}'\mathbf{W}(r)$; $\mathbf{Lb} = l_{11}\kappa$; $\mathbf{b}'\mathbf{\Omega}\mathbf{b} = \omega_{11.2}\kappa'\kappa$; $\mathbf{b}'\mathbf{B}(r) = l_{11}Q(r)$; $\mathbf{b}'\left(\int_0^1 \mathbf{B}d\mathbf{B}'\right)\mathbf{b} = \omega_{11.2}\left(\int_0^1 Q(r)dQ(r)'\right)$; and $\mathbf{b}'\mathbf{Ab} = a_{11.2} = \omega_{11.2}\int_0^1 Q(r)^2$.

Moreover, note that,

$$\begin{aligned} \widehat{\mathbf{b}}' &= (1, -\widehat{\beta}') \\ &= \left(1, -\left(\sum_{t=1}^T y_t \mathbf{x}_t'\right)\left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t'\right)^{-1}\right) \\ &= \left(1, -\left(\frac{1}{T^2} \sum_{t=1}^T y_t \mathbf{x}_t'\right)\left(\frac{1}{T^2} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t'\right)^{-1}\right) \\ &\rightarrow (1, -a'_{xy}\mathbf{A}_{xx}^{-1}) = \mathbf{b}' \end{aligned} \tag{A.1}$$

Proof of Theorem 1

Considering Assumptions $\mathcal{A}.1 - \mathcal{A}.4$, under the null: $\varepsilon_t = \varepsilon_{t-1} + \nu_t$ with ν_t white noise (this assumption will be relaxed later), the test regression we consider is:

$$\Delta e_t = c_j + \phi_j e_{t-1} + \eta_t^*, \tag{A.2}$$

where e_t are the full sample LS residuals from (2.1).

Recalling that the KPZ test statistic for H_{1A} (where the first regime is $I(1)$) is,

$$F_A(\tau, m^*) := \begin{cases} (T - m^* - p_T)(SSR_0 - SSR_{A,m^*}) / (m^* SSR_{A,m^*}) & \text{if } m^* \text{ is even} \\ (T - m^* - 1 - p_T)(SSR_0 - SSR_{A,m^*}) / ((m^* + 1) SSR_{A,m^*}) & \text{if } m^* \text{ is odd} \end{cases},$$

where $\tau := (\tau_1, \dots, \tau_{m^*})$ for $\tau_j := T_j/T$ and the number of changes is fixed, $m = m^*$. Thus, under $H_0 : c_j = 0$ and $\phi_j = 0$, $j = 1, \dots, m + 1$, it follows from (A.2) that,

$$SSR_0 = \sum_{t=1}^T (\Delta e_t)^2 = \sum_{t=1}^T (\hat{\mathbf{b}}' \mathbf{X}_t - \hat{\mathbf{b}}' \mathbf{X}_{t-1})^2 = \sum_{t=1}^T (\hat{\mathbf{b}}' \xi_t)^2 = \hat{\mathbf{b}}' \left(\sum_{t=1}^T \xi_t \xi_t' \right) \hat{\mathbf{b}} \quad (\text{A.3})$$

where $\xi_t := (\varepsilon_t, u_t)'$. Moreover, under the alternative hypothesis and for m^* even we obtain,

$$SSR_{A,m^*} = \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} \left(\Delta e_t - \bar{\Delta e}_{2j} - \hat{\phi}_{2j} (e_{t-1} - \bar{e}_{2j,-1}) \right)^2 + \sum_{j=0}^{m^*/2} \sum_{t=T_{2j}+1}^{T_{2j+1}} (\Delta e_t)^2.$$

Noting that $\sum_{j=0}^{m^*/2} \sum_{t=T_{2j}+1}^{T_{2j+1}} (\Delta e_t)^2 = \sum_{t=1}^T (\Delta e_t)^2 - \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t)^2$, it follows that,

$$SSR_{A,m^*} = \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} \left(\Delta e_t - \bar{\Delta e}_{2j} - \hat{\phi}_{2j} (e_{t-1} - \bar{e}_{2j,-1}) \right)^2 + \sum_{t=1}^T (\Delta e_t)^2 - \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t)^2$$

Since,

$$\begin{aligned} \sum_{t=T_{2j-1}+1}^{T_{2j}} \left((\Delta e_t - \bar{\Delta e}_{2j}) - \hat{\phi}_{2j} (e_{t-1} - \bar{e}_{2j,-1}) \right)^2 &= \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t - \bar{\Delta e}_{2j})^2 \\ &\quad - 2\hat{\phi}_{2j} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t - \bar{\Delta e}_{2j}) (e_{t-1} - \bar{e}_{2j,-1}) \\ &\quad + \hat{\phi}_{2j}^2 \sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2 \end{aligned}$$

and given that $\hat{\phi}_{2j} = \frac{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})(\Delta e_t - \bar{\Delta e}_{2j})}{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2}$, and

$$\begin{aligned} \hat{\phi}_{2j}^2 \sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2 &= \hat{\phi}_{2j} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t - \bar{\Delta e}_{2j}) (e_{t-1} - \bar{e}_{2j,-1}) \\ &= \frac{\left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1}) (\Delta e_t - \bar{\Delta e}_{2j}) \right)^2}{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2} \\ &= : \Xi \end{aligned}$$

we establish that

$$\sum_{t=T_{2j-1}+1}^{T_{2j}} \left((\Delta e_t - \overline{\Delta e_{2j}}) - \hat{\phi}_{2j}(e_{t-1} - \bar{e}_{2j,-1}) \right)^2 = \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t - \overline{\Delta e_{2j}})^2 - \Xi.$$

Therefore,

$$\begin{aligned} SSR_{A,m^*} &= \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} \left(\Delta e_t - \overline{\Delta e_{2j}} - \hat{\phi}_{2j}(e_{t-1} - \bar{e}_{2j,-1}) \right)^2 + \sum_{t=1}^T (\Delta e_t)^2 - \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t)^2 \\ &= \sum_{j=1}^{m^*/2} \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t - \overline{\Delta e_{2j}})^2 - \Xi \right) + \sum_{t=1}^T (\Delta e_t)^2 - \sum_{j=1}^{m^*/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} (\Delta e_t)^2 \\ &= \sum_{j=1}^{m^*/2} \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} \left[(\Delta e_t - \overline{\Delta e_{2j}})^2 - (\Delta e_t)^2 \right] - \Xi \right) + \sum_{t=1}^T (\Delta e_t)^2 \\ &= \sum_{j=1}^{m^*/2} \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (\overline{\Delta e_{2j}})^2 - 2 \sum_{t=T_{2j-1}+1}^{T_{2j}} (\overline{\Delta e_{2j}}) (\Delta e_t) - \Xi \right) + \sum_{t=1}^T (\Delta e_t)^2 \\ &= \sum_{j=1}^{m^*/2} \left(-\frac{1}{T_{2j} - T_{2j-1}} \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} \Delta e_t \right)^2 - \Xi \right) + SSR_0 \end{aligned}$$

where

$$\begin{aligned} \bar{e}_{2j} &: = \frac{1}{T_{2j} - T_{2j-1}} \sum_{t=T_{2j-1}+1}^{T_{2j}} e_t = \hat{b}' \overline{\mathbf{X}}_{2j}, \\ \bar{e}_{2j,-1} &: = \frac{1}{T_{2j} - T_{2j-1}} \sum_{t=T_{2j-1}+1}^{T_{2j}} e_{t-1} = \hat{b}' \overline{\mathbf{X}}_{2j,-1}, \end{aligned}$$

for $j = 1, \dots, m^*/2$ and $\overline{\Delta \mathbf{X}}_{2j} := \frac{1}{T_{2j} - T_{2j-1}} \sum_{t=T_{2j-1}+1}^{T_{2j}} \Delta \overline{\mathbf{X}}_t = \bar{\xi}_{2j}$.

Hence,

$$\begin{aligned}
& SSR_0 - SSR_{A,m^*} \\
&= \sum_{j=1}^{m^*/2} \left[\Xi + \frac{1}{T_{2j} - T_{2j-1}} \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} \Delta e_t \right)^2 \right] \\
&= \sum_{j=1}^{m^*/2} \left[\frac{\left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1}) (\Delta e_t - \bar{\Delta} e_{2j}) \right)^2}{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2} + \frac{T}{T_{2j} - T_{2j-1}} \left(b' \frac{1}{\sqrt{T}} \sum_{t=T_{2j-1}+1}^{T_{2j}} \xi_t \right)^2 \right] \\
&\Rightarrow \sum_{j=1}^{m^*/2} \left[\frac{\left(\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) d\mathbf{B}(r)' + (\tau_{2j} - \tau_{2j-1}) \boldsymbol{\Omega}_1 \right) \mathbf{b} \right)^2}{\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) \mathbf{B}^{(2j)}(r)' dr \right) \mathbf{b}} \right. \\
&\quad \left. + \mathbf{b}' \frac{1}{\tau_{2j} - \tau_{2j-1}} (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})) (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1}))' \mathbf{b} \right].
\end{aligned}$$

Note from (A.2) that,

$$\frac{1}{T} SSR_{A,m^*} = \hat{\mathbf{b}}' \left(\frac{1}{T} \sum_{t=1}^T \xi_t \xi_t' \right) \hat{\mathbf{b}} + o_p(1) \Rightarrow \mathbf{b}' \boldsymbol{\Omega}_0 \mathbf{b}.$$

Hence, for m^* fixed and even, and assuming $p_T = 0$, it follows that,

$$\begin{aligned}
F_A(\tau, m^*) &= \frac{1}{m^*} \frac{(SSR_0 - SSR_{A,m^*})}{SSR_{A,m^*} / (T - m^*)} \\
&\Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}' \boldsymbol{\Omega}_0 \mathbf{b}} \sum_{j=1}^{m^*/2} \left[\frac{\left(\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) d\mathbf{B}(r)' + (\tau_{2j} - \tau_{2j-1}) \boldsymbol{\Omega}_1 \right) \mathbf{b} \right)^2}{\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) \mathbf{B}^{(2j)}(r)' dr \right) \mathbf{b}} \right. \\
&\quad \left. + \mathbf{b}' \frac{1}{\tau_{2j} - \tau_{2j-1}} (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})) (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1}))' \mathbf{b} \right]. \quad (\text{A.4})
\end{aligned}$$

We can establish that,

$$\begin{aligned}
\mathbf{b}' \mathbf{B}^{(2j)}(r) &= \mathbf{b}' \left(\mathbf{B}(r) - \frac{1}{\tau_{2j} - \tau_{2j-1}} \int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}(r) dr \right) \\
&= \mathbf{b}' \mathbf{B}(r) - \frac{1}{\tau_{2j} - \tau_{2j-1}} \int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{b}' \mathbf{B}(r) dr \\
&= l_{11} Q(r) - \frac{1}{\tau_{2j} - \tau_{2j-1}} \int_{\tau_{2j-1}}^{\tau_{2j}} l_{11} Q(r) dr \\
&= l_{11} Q^{(2j)}; \quad (\text{A.5})
\end{aligned}$$

$$\begin{aligned}
\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) \mathbf{B}^{(2j)}(r)' dr \right) \mathbf{b} &= l_{11}^2 \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r)' dr \right) \\
&= \omega_{11 \cdot 2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r)' dr \right) \quad (\text{A.6})
\end{aligned}$$

$$\begin{aligned}
& \mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) d\mathbf{B}(r)' + (\tau_{2j} - \tau_{2j-1}) \boldsymbol{\Omega}_1 \right) \mathbf{b} \\
&= \int_{\tau_{2j-1}}^{\tau_{2j}} l_{11} Q^{(2j)}(r) dQ(r)' l_{11} + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}' \boldsymbol{\Omega}_1 \mathbf{b} \\
&= \omega_{11.2} \left(\int_0^1 Q^{(2j)}(r) dQ(r)' \right) + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}' \boldsymbol{\Omega}_1 \mathbf{b} \tag{A.7}
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{b}' \frac{1}{\tau_{2j} - \tau_{2j-1}} (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})) (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1}))' \mathbf{b} \\
&= \frac{1}{\tau_{2j} - \tau_{2j-1}} (\mathbf{b}' \mathbf{B}(\tau_{2j}) - \mathbf{b}' \mathbf{B}(\tau_{2j-1})) (\mathbf{B}(\tau_{2j}) \mathbf{b} - \mathbf{B}(\tau_{2j-1}) \mathbf{b})' \\
&= \frac{1}{\tau_{2j} - \tau_{2j-1}} (l_{11} Q(\tau_{2j}) - l_{11} Q(\tau_{2j-1})) (Q(\tau_{2j}) l_{11} - Q(\tau_{2j-1}) l_{11})' \\
&= \frac{1}{\tau_{2j} - \tau_{2j-1}} \omega_{11.2} (Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))'. \tag{A.8}
\end{aligned}$$

Hence, from (A.6) to (A.8) we establish that under joint convergence,

$$\begin{aligned}
F_A(\tau, m^*) &\Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}' \boldsymbol{\Omega}_0 \mathbf{b}} \sum_{j=1}^{m^*/2} \left[\frac{\left(\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}' \boldsymbol{\Omega}_1 \mathbf{b} \right)^2}{\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r)' dr \right)} \right. \\
&\quad \left. + \frac{1}{\tau_{2j} - \tau_{2j-1}} \omega_{11.2} (Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))' \right] \\
&= \frac{1}{m^*} \frac{1}{\mathbf{b}' \boldsymbol{\Omega}_0 \mathbf{b}} \sum_{j=1}^{m^*/2} \left[\frac{\left(\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}' \boldsymbol{\Omega}_1 \mathbf{b} \right)^2}{\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r)' \right)} \right. \\
&\quad \left. + \omega_{11.2} \frac{1}{\tau_{2j} - \tau_{2j-1}} (Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))' \right] \\
&= \frac{1}{m^*} \frac{1}{\mathbf{b}' \boldsymbol{\Omega}_0 \mathbf{b}} \sum_{j=1}^{m^*/2} \left[\frac{\left(\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) + (\tau_{2j} - \tau_{2j-1}) \mathbf{b}' \boldsymbol{\Omega}_1 \mathbf{b} \right)^2}{\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r)^2 \right)} \right. \\
&\quad \left. + \omega_{11.2} \frac{1}{\tau_{2j} - \tau_{2j-1}} (Q(\tau_{2j}) - Q(\tau_{2j-1}))^2 \right]. \tag{A.9}
\end{aligned}$$

Remark A.1: Note that Q is a scalar and that the distribution of this distribution is non-standard and depends on the nuisance parameters Ω_0, Ω_1 . This is still the case even if Ω_0 and Ω_1 are block diagonal. This "problem" is typical in limit theory for (non)cointegrating regressions. In this context, we can propose transformations of the statistic which involve consistent estimates of the nuisance parameters (Phillips and Park, 1988) or we can try FM-OLS optimal estimation of Phillips and Hansen (1990) which introduces nonparametric corrections in the OLS estimator. \square

Note that if ξ_t is not autocorrelated, the distribution becomes free of nuisance parameters even if we have endogeneity, $\omega_{21} \neq 0$. Since in this case, $\Omega_1 = 0$, $\Omega = \Omega_0$, we can write (A.9) as,

$$\begin{aligned}
F_A(\tau, m^*) &\Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}'\Omega\mathbf{b}} \sum_{j=1}^{m^*/2} \left[\left[\omega_{11 \cdot 2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r) dr \right) \right]^{-1} \left(\omega_{11 \cdot 2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) \right)^2 \right. \\
&\quad \left. + \omega_{11 \cdot 2} \frac{(Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))'}{\tau_{2j} - \tau_{2j-1}} \right] \\
&= \frac{1}{m^*} \frac{1}{\omega_{11 \cdot 2} \kappa' \kappa} \sum_{j=1}^{m^*/2} \left[\left[\omega_{11 \cdot 2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r) dr \right) \right]^{-1} \left(\omega_{11 \cdot 2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) \right)^2 \right. \\
&\quad \left. + \omega_{11 \cdot 2} \frac{(Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))'}{\tau_{2j} - \tau_{2j-1}} \right] \\
&= \frac{1}{m^*} \frac{1}{\kappa' \kappa} \sum_{j=1}^{m^*/2} \left[\frac{\left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r) dr} + \frac{(Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))'}{\tau_{2j} - \tau_{2j-1}} \right] \\
&= \frac{1}{m^*} \frac{1}{\kappa' \kappa} \sum_{j=1}^{m^*/2} \left[\frac{\left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dr} + \frac{(Q(\tau_{2j}) - Q(\tau_{2j-1}))^2}{\tau_{2j} - \tau_{2j-1}} \right]. \tag{A.10}
\end{aligned}$$

Remark A.2: For $n = 1$, this becomes the KPZ distribution, as $Q = W_1 \equiv W$ and $\kappa = 1$. The case of $m^* = 0$ (standard cointegration) cannot be obtained from this distribution. \square

To correct for the nuisance parameters consider instead of SSR_{A,m^*} the following corrected sum of squared residuals,

$$SSR_{A,m^*}^\# = \sum_{j=1}^{m^*/2} \left[\sum_{t=T_{2j-1}+1}^{T_{2j}} \left(\Delta e_t - \bar{\Delta} e_{2j} - (\hat{\phi}_{2j} - \hat{\Upsilon}_1) (e_{t-1} - \bar{e}_{2j,-1}) \right)^2 \right] + \sum_{j=0}^{m^*/2} \sum_{t=T_{2j}+1}^{T_{2j+1}} (\Delta e_t)^2,$$

where the scalar correction term $\hat{\Upsilon}_1$ is defined as,

$$\begin{aligned}
\hat{\Upsilon}_1 &: = \frac{\left(\frac{T_{2j} - T_{2j-1}}{T} \right) \hat{\mathbf{b}}' \hat{\Omega}_1 \hat{\mathbf{b}}}{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2} \\
&= \frac{\left(\frac{T_{2j} - T_{2j-1}}{T} \right) \hat{\mathbf{b}}' \hat{\Omega}_1 \hat{\mathbf{b}}}{\hat{\mathbf{b}}' \left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (x_{t-1} - \bar{x}_{2j,-1}) (x_{t-1} - \bar{x}_{2j,-1})' \right) \hat{\mathbf{b}}}.
\end{aligned}$$

The statistic considering $p_T = 0$, and m^* even is,

$$F_A^\#(\tau, m^*) = (T - m^* - p_T) \frac{\left(SSR_0 - SSR_{A,m^*}^\# \right)}{m^* \left(SSR_{A,m^*}^\# + \hat{\Upsilon}_2 \right)}, \tag{A.11}$$

where the second (scalar) correction term $\widehat{\Upsilon}_2$ in the denominator of (A.11) is,

$$\widehat{\Upsilon}_2 := T\widehat{\mathbf{b}}' \left(\widehat{\boldsymbol{\Omega}}_1 + \widehat{\boldsymbol{\Omega}}_1' \right) \widehat{\mathbf{b}} = T\widehat{\mathbf{b}}' \left(\widehat{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}_0 \right) \widehat{\mathbf{b}}.$$

Hence, for the numerator of (A.11) we observe that,

$$\begin{aligned} SSR_0 - SSR_{A,m^*}^\# &= \sum_{j=1}^{m^*/2} \left[\frac{\left(\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1}) (\Delta e_t - \bar{\Delta} e_{2j}) - \left(\frac{T_{2j} - T_{2j-1}}{T} \right) \widehat{\mathbf{b}}' \widehat{\boldsymbol{\Omega}}_1 \widehat{\mathbf{b}} \right)^2}{\sum_{t=T_{2j-1}+1}^{T_{2j}} (e_{t-1} - \bar{e}_{2j,-1})^2} \right. \\ &\quad \left. + \frac{T}{T_{2j} - T_{2j-1}} \left(\mathbf{b}' T^{-1/2} \sum_{t=T_{2j-1}+1}^{T_{2j}} \xi_t \right)^2 \right] \\ &\Rightarrow \sum_{j=1}^{m^*/2} \left[\frac{\left(\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) d\mathbf{B}(r)' \right) \mathbf{b} \right)^2}{\mathbf{b}' \left(\int_{\tau_{2j-1}}^{\tau_{2j}} \mathbf{B}^{(2j)}(r) \mathbf{B}^{(2j)}(r)' \right) \mathbf{b}} \right. \\ &\quad \left. + \mathbf{b}' \frac{1}{\tau_{2j} - \tau_{2j-1}} (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1})) (\mathbf{B}(\tau_{2j}) - \mathbf{B}(\tau_{2j-1}))' \mathbf{b} \right] \end{aligned} \quad (\text{A.12})$$

and

$$\frac{SSR_{A,m^*}^\# + \widehat{\Upsilon}_2}{T} = \widehat{\mathbf{b}}' \left(\frac{1}{T} \sum_{t=1}^T \xi_t \xi_t' \right) \widehat{\mathbf{b}} + o_p(1) + \widehat{\mathbf{b}}' \left(\widehat{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}_0 \right) \widehat{\mathbf{b}} \Rightarrow \mathbf{b}' \boldsymbol{\Omega} \mathbf{b}. \quad (\text{A.13})$$

Thus, under joint convergence of (A.12) and (A.13) we establish that,

$$\begin{aligned} F_A^\#(\tau, m^*) &\Rightarrow \frac{1}{m^*} \frac{1}{\mathbf{b}' \boldsymbol{\Omega} \mathbf{b}} \sum_{j=1}^{m^*/2} \left[\frac{\left(\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right) \right)^2}{\omega_{11.2} \left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) Q^{(2j)}(r)' dr \right)} \right. \\ &\quad \left. + \omega_{11.2} \frac{1}{\tau_{2j} - \tau_{2j-1}} (Q(\tau_{2j}) - Q(\tau_{2j-1})) (Q(\tau_{2j}) - Q(\tau_{2j-1}))' \right] \\ &= \frac{1}{m^*} \frac{1}{\kappa' \kappa} \sum_{j=1}^{m^*/2} \left[\frac{\left(\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r) dQ(r)' \right)^2}{\int_{\tau_{2j-1}}^{\tau_{2j}} Q^{(2j)}(r)^2 dr} + \frac{1}{\tau_{2j} - \tau_{2j-1}} (Q(\tau_{2j}) - Q(\tau_{2j-1}))^2 \right], \end{aligned}$$

as in the case of $F_A(\tau, m^*)$ with ξ_t not autocorrelated. ■

$$\sup F_A(m^*) := \sup_{\tau \in \Lambda_\epsilon^{m^*}} F_A(\tau, m^*) \text{ and } \sup F_B(m^*) := \sup_{\tau \in \Lambda_\epsilon^{m^*}} F_B(\tau, m^*) \quad (\text{A.14})$$

converge weakly, \Rightarrow , as $T \rightarrow \infty$, to the $\sup_{\tau \in \Lambda_\epsilon^{m^*}}$ of the correspondent limiting laws defined above.

The same applies to

$$\mathcal{W}(m^*) := \max[\sup F_A(m^*), \sup F_B(m^*)]$$

and

$$\mathcal{W} \max := \max_{1 \leq m \leq \bar{m}} \mathcal{W}(m). \quad (\text{A.15})$$

Proof of Theorem 2

Consider first, the $MZ_{\hat{\rho}}^{GLS}(\lambda, \tau)$ and $MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau)$ statistics presented in (3.8) and (3.9). Recall that $\tilde{e}_t = e_t - \hat{\delta}_{\hat{\rho}}$ where $\hat{\delta}_{\hat{\rho}}$ is the LS estimator from regressing $(e_{\lfloor \lambda T \rfloor}, e_{\lfloor \lambda T \rfloor + 1} - \bar{\rho}e_{\lfloor \lambda T \rfloor}, \dots, e_{\lfloor \tau T \rfloor} - \bar{\rho}e_{\lfloor \tau T \rfloor - 1})'$ on $(1, 1 - \bar{\rho}, \dots, 1 - \bar{\rho})'$. Thus, ignoring the first observation on the estimation of $\delta_{\bar{\rho}}$, it follows that,

$$\begin{aligned} \tilde{e}_t &= e_t - \frac{1}{(1 - \bar{\rho})} \frac{1}{\lfloor (\tau - \lambda)T \rfloor} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (e_t - \bar{\rho}e_{t-1}) \\ &= \hat{\mathbf{b}}' \left(x_t - \frac{1}{(1 - \bar{\rho})} \frac{1}{\lfloor (\tau - \lambda)T \rfloor} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (x_t - \bar{\rho}x_{t-1}) \right). \end{aligned}$$

Then,

$$\begin{aligned} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) &= \left[2 \frac{1}{(\lfloor (\tau - \lambda)T \rfloor)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\hat{\mathbf{b}}' \mathbf{x}_{t-1})^2 \right]^{-1} \left[\frac{1}{\lfloor (\tau - \lambda)T \rfloor} (\hat{\mathbf{b}}' \mathbf{x}_{\tau T})^2 - s(\lambda, \tau)^2 \right] \\ &= \left[2 \hat{\mathbf{b}}' \left(\frac{1}{(\lfloor (\tau - \lambda)T \rfloor)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \mathbf{x}_{t-1} \mathbf{x}_{t-1}' \right) \hat{\mathbf{b}} \right]^{-1} \left[\hat{\mathbf{b}}' \left(\frac{\lfloor \tau T \rfloor}{\lfloor (\tau - \lambda)T \rfloor} \frac{1}{\lfloor \tau T \rfloor} \mathbf{x}_{\lfloor \tau T \rfloor} \mathbf{x}_{\lfloor \tau T \rfloor}' \right) \hat{\mathbf{b}} - s(\lambda, \tau)^2 \right] \\ &\Rightarrow \left[2 \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr \right) \mathbf{b} \right]^{-1} \left[\frac{\tau}{(\tau - \lambda)} \mathbf{b}' (\mathbf{B}(\tau) \mathbf{B}(\tau)') \mathbf{b} - \omega_{11.2} \kappa' \kappa \right] \\ &= \left[2 \omega_{11.2} \left(\int_{\lambda}^{\tau} Q(r) Q(r)' dr \right) \right]^{-1} \left[\omega_{11.2} Q(\tau) Q(\tau)' - \omega_{11.2} \kappa' \kappa \right] \\ &= \left[2 \int_{\lambda}^{\tau} Q(r) Q(r)' dr \right]^{-1} \left[\frac{\tau}{(\tau - \lambda)} Q(\tau) Q(\tau)' - \kappa' \kappa \right] \\ &= \left[2 \int_{\lambda}^{\tau} Q(r)^2 \right]^{-1} \left[\frac{\tau}{(\tau - \lambda)} Q(\tau)^2 - \kappa' \kappa \right], \end{aligned}$$

which does not depend on nuisance parameters. Q and $\kappa' \kappa$ are scalars. Recall that $Q(r) := W_1(r) - \int_0^1 (W_1(r) \mathbf{W}_2(r)') \left(\int_0^1 \mathbf{W}_2(r) \mathbf{W}_2(r)' \right)^{-1} \mathbf{W}_2(r)$, and $\kappa := \left(1, - \left(\int_0^1 (W_1(r) \mathbf{W}_2(r)') \right) \left(\int_0^1 \mathbf{W}_2(r) \mathbf{W}_2(r)' \right)^{-1} \right)'$.

By the same token,

$$MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \left(4(\kappa' \kappa) \int_{\lambda}^{\tau} Q(r)^2 dr \right)^{-1/2} \left(\frac{\tau Q(\tau)^2}{(\tau - \lambda)} - \kappa' \kappa \right).$$

When $n = 1$,

$$\begin{aligned} MZ_{\hat{\rho}}(\lambda, \tau) &\Rightarrow \left(2 \int_{\lambda}^{\tau} W(r)^2 dr \right)^{-1} \left(\frac{\tau W(\tau)^2}{(\tau - \lambda)} - 1 \right); \\ MZ_{t_{\hat{\rho}}}(\lambda, \tau) &\Rightarrow \left(4 \int_{\lambda}^{\tau} W(r)^2 dr \right)^{-1/2} \left(\frac{\tau W(\tau)^2}{(\tau - \lambda)} - 1 \right), \end{aligned}$$

which are alternatives to LKT. For $\tau = 1, \lambda = 0$, $MZ_{\alpha}(\lambda, \tau)$ and $MZ_t(\lambda, \tau)$ are the Perron and Ng (1996)'s distributions, MZ_{α} and MZ_t . If $n > 1$ and $\tau = 1, \lambda = 0$ we have Pesavento and PR stats.

Consequently,

$$\begin{aligned} MZ_{\hat{\rho}}^f(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \frac{\frac{\tau}{(\tau - \lambda)} Q(\tau)^2 - \kappa' \kappa}{2 \int_{\lambda}^{\tau} Q(r)^2 dr} \\ MZ_{t_{\hat{\rho}}}^f(\lambda) &= \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \frac{\frac{\tau}{(\tau - \lambda)} Q(\tau)^2 - \kappa' \kappa}{2(\kappa' \kappa)^{1/2} \left(\int_{\lambda}^{\tau} Q(r)^2 dr \right)^{1/2}}. \end{aligned}$$

For a shift from I(1) to I(0), we consider the time-reversed data, $\tilde{e}_t = e_{T-t+1}, t = 1, \dots, T$, so that

$$\begin{aligned} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) &= \frac{\frac{1}{[(\tau - \lambda)T]} \left(\hat{\mathbf{b}}' \mathbf{x}_{(1-\tau)T+1} \right)^2 - s(\lambda, \tau)^2}{\frac{1}{[(\tau - \lambda)T]^2} \sum_{t=[(1-\tau)T]}^{[(1-\lambda)T]} \left(\hat{\mathbf{b}}' \mathbf{x}_{t+2} \right)^2} \\ &= \frac{\hat{\mathbf{b}}' \left(\frac{[(1-\tau)T]+1}{[(\tau - \lambda)T]} \frac{1}{[(1-\tau)T]+1} \mathbf{x}_{(1-\tau)T+1} \mathbf{x}'_{(1-\tau)T+1} \right) \hat{\mathbf{b}} - s(\lambda, \tau)^2}{2 \hat{\mathbf{b}}' \left(\frac{1}{[(\tau - \lambda)T]^2} \sum_{t=[(1-\tau)T]}^{[(1-\lambda)T]} \mathbf{x}_{t+2} \mathbf{x}'_{t+2} \right) \hat{\mathbf{b}}} \\ &\Rightarrow \frac{\frac{(1-\tau)}{(\tau - \lambda)} \mathbf{b}' (\mathbf{B}(1-\tau) \mathbf{B}(1-\tau)') \mathbf{b} - \omega_{11.2} \kappa' \kappa}{2 \mathbf{b}' \left(\int_{(1-\tau)}^{(1-\lambda)} \mathbf{B}(r) \mathbf{B}(r)' dr \right) \mathbf{b}} \\ &= \frac{\frac{(1-\tau)}{(\tau - \lambda)} \omega_{11.2} Q(1-\tau) Q(1-\tau)' - \omega_{11.2} \kappa' \kappa}{2 \omega_{11.2} \left(\int_{(1-\tau)}^{(1-\lambda)} Q(r) Q(r)' dr \right)} \\ &= \frac{\frac{(1-\tau)}{(\tau - \lambda)} Q(1-\tau)^2 - \kappa' \kappa}{2 \int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr} \tag{A.16} \end{aligned}$$

$$MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \frac{\frac{(1-\tau)}{(\tau - \lambda)} Q(1-\tau)^2 - \kappa' \kappa}{2(\kappa' \kappa)^{1/2} \left(\int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr \right)^{1/2}}. \tag{A.17}$$

Consequently,

$$MZ_{\hat{\rho}}^r(\lambda) = \inf_{\tau \in (\lambda, 1]} MZ_{\hat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \frac{\frac{(1-\tau)}{(\tau-\lambda)} Q(1-\tau)^2 - \kappa' \kappa}{2 \int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr}$$

$$MZ_{t_{\hat{\rho}}}^r(\lambda) = \inf_{\tau \in (\lambda, 1]} MZ_{t_{\hat{\rho}}}^{GLS}(\lambda, \tau) \Rightarrow \inf_{\tau \in (\lambda, 1]} \frac{\frac{(1-\tau)}{(\tau-\lambda)} Q(1-\tau)^2 - \kappa' \kappa}{2 (\kappa' \kappa)^{1/2} \left(\int_{(1-\tau)}^{(1-\lambda)} Q(r)^2 dr \right)^{1/2}}$$

and

$$MZ_{\hat{\rho}}^f(\lambda) = \min \left\{ MZ_{\hat{\rho}}^f(\lambda), MZ_{\hat{\rho}}^r(\lambda) \right\} \Rightarrow \min \{ \text{see above} \}$$

$$MZ_{t_{\hat{\rho}}}^f(\lambda) = \min \left\{ MZ_{t_{\hat{\rho}}}^f(\lambda), MZ_{t_{\hat{\rho}}}^r(\lambda) \right\} \Rightarrow \min \{ \text{see above} \}$$

Finally,

$$MZ_{\alpha}^f \Rightarrow \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} \frac{\frac{\tau}{(\tau-\lambda)} Q(\tau)^2 - \kappa' \kappa}{2 \int_{\lambda}^{\tau} Q(r)^2 dr}$$

$$MZ_t^f \Rightarrow \inf_{\lambda \in (0, 1)} \inf_{\tau \in (\lambda, 1]} \frac{\frac{\tau}{(\tau-\lambda)} Q(\tau)^2 - \kappa' \kappa}{2 (\kappa' \kappa)^{1/2} \left(\int_{\lambda}^{\tau} Q(r)^2 dr \right)^{1/2}}.$$

Now, with our OLS-GLS residuals

$$\tilde{e}_t = \hat{\mathbf{b}}' \mathbf{x}_t - \hat{\delta}_{\bar{\rho}} = \hat{\mathbf{b}}' \left(\mathbf{x}_t - \frac{1}{(1-\bar{\rho})} \frac{1}{\lfloor (\tau-\lambda)T \rfloor} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\mathbf{x}_t - \bar{\rho} \mathbf{x}_{t-1}) \right).$$

Note that $s(\lambda, \tau)^2$ has the same properties as before since $\Delta \tilde{e}_t = \Delta (e_t - \hat{\delta}_{\bar{\rho}}) = \Delta e_t$. Here,

$$\begin{aligned} \frac{1}{(\lfloor (\tau-\lambda)T \rfloor)^{1/2}} \hat{\delta}_{\bar{\rho}} &= \frac{1}{(1-\bar{\rho})} \frac{1}{(\lfloor (\tau-\lambda)T \rfloor)^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} (\hat{\mathbf{b}}' \mathbf{x}_t - \bar{\rho} \hat{\mathbf{b}}' \mathbf{x}_{t-1}) \\ &= \frac{1}{(1-\bar{\rho})} \hat{\mathbf{b}}' \left(\frac{1}{(\lfloor (\tau-\lambda)T \rfloor)^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \mathbf{x}_t - \frac{\bar{\rho}}{(\lfloor (\tau-\lambda)T \rfloor)^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \mathbf{x}_{t-1} \right) \\ &\Rightarrow \frac{1}{(1-\bar{\rho})} \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda) - \bar{\rho} (\mathbf{B}(\tau) - \mathbf{B}(\lambda))) \\ &= \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) \end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \left(\widehat{\mathbf{b}}' \mathbf{x}_{t-1} - \widehat{\delta}_{\widehat{\rho}} \right)^2 \\
= & \frac{1}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \left(\widehat{\mathbf{b}}' \mathbf{x}_{t-1} \right)^2 + \frac{1}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \widehat{\delta}_{\widehat{\rho}}^2 - 2 \frac{1}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \widehat{\mathbf{b}}' \mathbf{x}_{t-1} \widehat{\delta}_{\widehat{\rho}} \\
= & \frac{1}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \left(\widehat{\mathbf{b}}' \mathbf{x}_{t-1} \right)^2 + \left(\frac{\widehat{\delta}_{\widehat{\rho}}}{([\tau - \lambda]T)^{1/2}} \right)^2 - 2 \frac{\widehat{\delta}_{\widehat{\rho}}}{([\tau - \lambda]T)^{1/2}} \frac{1}{([\tau - \lambda]T)^{3/2}} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \widehat{\mathbf{b}}' \mathbf{x}_{t-1} \\
\Rightarrow & \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr \right) \mathbf{b} + \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \mathbf{b} \\
& - 2 \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \mathbf{b} \\
= & \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr \right) \mathbf{b} - \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \mathbf{b} \\
= & \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr - (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \right) \mathbf{b}
\end{aligned}$$

Then,

$$\begin{aligned}
MZ_{\widehat{\rho}}^{GLS}(\lambda, \tau) &= \frac{\frac{1}{[\tau - \lambda]T} \left(\widehat{\mathbf{b}}' \mathbf{x}_{\tau T} - \widehat{\delta}_{\widehat{\rho}} \right)^2 - s(\lambda, \tau)^2}{\frac{2}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \left(\widehat{\mathbf{b}}' \mathbf{x}_{t-1} - \widehat{\delta}_{\widehat{\rho}} \right)^2} \\
&= \frac{\left(\widehat{\mathbf{b}}' \frac{([\tau T]^{1/2})}{([\tau - \lambda]T)^{1/2}} \frac{1}{([\tau T]^{1/2})} \mathbf{x}_{\tau T} - \frac{1}{([\tau - \lambda]T)^{1/2}} \widehat{\delta}_{\widehat{\rho}} \right)^2 - s(\lambda, \tau)^2}{\frac{2}{([\tau - \lambda]T)^2} \sum_{t=\lfloor \lambda T \rfloor}^{\lfloor \tau T \rfloor} \left(\widehat{\mathbf{b}}' \mathbf{x}_{t-1} - \widehat{\delta}_{\widehat{\rho}} \right)^2} \\
\Rightarrow & \frac{\left(\mathbf{b}' \left(\frac{\tau}{\tau - \lambda} \right)^{1/2} \mathbf{B}(\tau) - \mathbf{b}' (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) \right)^2 - \omega_{11.2} \kappa' \kappa}{2 \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr - (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \right) \mathbf{b}} \\
& \frac{\mathbf{b}' \left(\left(\frac{\tau}{\tau - \lambda} \right)^{1/2} \mathbf{B}(\tau) - (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) \right) \left(\left(\frac{\tau}{\tau - \lambda} \right)^{1/2} \mathbf{B}(\tau)' - (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \right) \mathbf{b} - \omega_{11.2} \kappa' \kappa}{2 \mathbf{b}' \left(\int_{\lambda}^{\tau} \mathbf{B}(r) \mathbf{B}(r)' dr - (\mathbf{B}(\tau) - \mathbf{B}(\lambda)) (\mathbf{B}(\tau) - \mathbf{B}(\lambda))' \right) \mathbf{b}} \\
& \frac{\omega_{11.2} \left(\left(\frac{\tau}{\tau - \lambda} \right)^{1/2} Q(\tau) - (Q(\tau) - Q(\lambda)) \right) \left(\left(\frac{\tau}{\tau - \lambda} \right)^{1/2} Q(\tau)' - (Q(\tau) - Q(\lambda))' \right) - \omega_{11.2} \kappa' \kappa}{2 \omega_{11.2} \left(\int_{\lambda}^{\tau} Q(r) Q(r)' dr - (Q(\tau) - Q(\lambda)) (Q(\tau) - Q(\lambda))' \right)} \\
& \frac{\left(\left(\frac{\tau}{\tau - \lambda} \right)^{1/2} Q(\tau) - (Q(\tau) - Q(\lambda)) \right)^2 - \kappa' \kappa}{2 \left(\int_{\lambda}^{\tau} Q(r)^2 dr - (Q(\tau) - Q(\lambda))^2 \right)}
\end{aligned}$$

and

$$MZ_{\widehat{\rho}}^{GLS}(\lambda, \tau) \Rightarrow \frac{\left(\sqrt{\frac{\tau}{\tau - \lambda}} Q(\tau) - (Q(\tau) - Q(\lambda)) \right)^2 - \kappa' \kappa}{2 \sqrt{\kappa' \kappa} \sqrt{\left(\int_{\lambda}^{\tau} Q(r)^2 dr - (Q(\tau) - Q(\lambda))^2 \right)}}$$

A Monte Carlo - TABLES

A.1 Size

Table 5: Wald type tests

	$n = 2$	$W(1)$	$W(2)$	$W(3)$	$W(4)$	$W \max$	$MZ_{\hat{\rho}}^{GLS}$		$MZ_{t_{\hat{\rho}}}^{GLS}$	
							0-1/1-0	Multiple	0-1/1-0	Multiple
$R^2 = 0, A_0$	T=200	0.028	0.096	0.047	0.047	0.047	0.049	0.065	0.048	0.066
	T=500	0.035	0.131	0.083	0.062	0.055	0.042	0.039	0.041	0.039
$R^2 = 0.8, A_1$	T=200	0.131	0.157	0.159	0.120	0.156	0.404	0.409	0.398	0.398
	T=500	0.084	0.142	0.121	0.072	0.090	0.260	0.334	0.255	0.331
$R^2 = 0.4, A_2$	T=200	0.029	0.105	0.052	0.059	0.045	0.098	0.118	0.092	0.116
	T=500	0.048	0.136	0.067	0.064	0.057	0.057	0.057	0.054	0.054
$R^2 = 0.4, A_3$	T=200	0.049	0.093	0.071	0.053	0.037	0.161	0.190	0.146	0.179
	T=500	0.061	0.185	0.102	0.087	0.077	0.101	0.098	0.091	0.093

A.2 Power

Single break at τ and I(0) regime AR(1) with coef $\phi.n = 2$. Only R^2, A with correct size (above).

Model (??), $p_y = 0$ and $p_x = 0$.

	T	$\mathcal{W}(1)$	$\mathcal{W}(2)$	$\mathcal{W}(3)$	$\mathcal{W}(4)$	\mathcal{W}_{\max}	$\mathcal{W}(1)$	$\mathcal{W}(2)$	$\mathcal{W}(3)$	$\mathcal{W}(4)$	\mathcal{W}_{\max}	$\mathcal{W}(1)$	$\mathcal{W}(2)$	$\mathcal{W}(3)$	$\mathcal{W}(4)$	\mathcal{W}_{\max}
		$\tau = 0.3; \phi = 0$					$\tau = 0.5; \phi = 0$					$\tau = 0.7; \phi = 0$				
I(0)-I(1)																
$R^2 = 0, A_0$	200	0.663	0.524	0.634	0.485	0.607	0.777	0.837	0.802	0.743	0.838	0.898	0.919	0.936	0.876	0.946
	500	0.715	0.699	0.722	0.654	0.720	0.789	0.881	0.872	0.848	0.870					
$R^2 = 0.4, A_1$	200	0.371	0.276	0.358	0.240	0.316	0.518	0.583	0.607	0.516	0.567	0.783	0.838	0.841	0.753	0.817
	500	0.460	0.452	0.475	0.420	0.452	0.605	0.741	0.722	0.704	0.703					
I(1)-I(0)																
$R^2 = 0, A_0$	200	0.992	0.995	0.995	0.980	0.997	0.961	0.983	0.981	0.966	0.989	0.945	0.947	0.957	0.939	0.958
	500	0.993	0.998	0.999	0.998	0.999	0.965	0.990	0.988	0.985	0.993					
$R^2 = 0.4, A_1$	200	0.964	0.991	0.989	0.983	0.989	0.886	0.956	0.943	0.913	0.948	0.846	0.884	0.875	0.861	0.895
	500	0.973	0.996	0.994	0.997	0.997	0.912	0.976	0.959	0.952	0.974					

$\tau = 0.3; \phi = 0.75$		$\mathcal{W}(1)$	$\mathcal{W}(2)$	$\mathcal{W}(3)$	$\mathcal{W}(4)$	\mathcal{W}_{\max}
I(0)-I(1)						
$R^2 = 0, A_0$	T=200	0.086	0.069	0.081	0.037	0.050
	T=500	0.334	0.196	0.297	0.157	0.231
$R^2 = 0.4, A_1$	T=200	0.068	0.093	0.081	0.038	0.052
	T=500	0.210	0.157	0.198	0.116	0.144
I(1)-I(0)						
$R^2 = 0, A_0$	T=200	0.894	0.766	0.621	0.413	0.833
	T=500	0.988	0.994	0.993	0.978	0.997
$R^2 = 0.4, A_1$	T=200	0.744	0.667	0.535	0.439	0.674
	T=500	0.954	0.980	0.975	0.929	0.983
I(0)-I(1)						
$R^2 = A = 0$	n=4	$\mathcal{W}(1)$	$\mathcal{W}(2)$	$\mathcal{W}(3)$	$\mathcal{W}(4)$	\mathcal{W}_{\max}
I(0)-I(1)						
$\tau = 0.5, \phi = 0$	T=200	0.786	0.769	0.791	0.681	0.814
	T=500	0.782	0.846	0.851	0.771	0.844
$\tau = 0.3, \phi = 0$	T=200	0.626	0.461	0.579	0.394	0.533
	T=500	0.703	0.609	0.675	0.566	0.656
$\tau = 0.3, \phi = 0.75$	T=200	0.152	0.122	0.145	0.053	0.087
	T=500	0.362	0.245	0.351	0.185	0.242
I(1)-I(0)						
$\tau = 0.5, \phi = 0$	T=200	0.875	0.948	0.928	0.883	0.957
	T=500	0.887	0.965	0.942	0.911	0.970
$\tau = 0.3, \phi = 0$	T=200	0.960	0.972	0.964	0.927	0.977
	T=500	0.974	0.988	0.987	0.978	0.988
$\tau = 0.3, \phi = 0.75$	T=200	0.813	0.705	0.599	0.355	0.730
	T=500	0.950	0.965	0.963	0.918	0.968

Wald type tests. Now with 3 regimes ϕ_1, ϕ_2, ϕ_3 ; 2 breaks τ_1, τ_2 . Results for "Multiple" test statistic. ϕ is the AR(1) coef for I(0) regime(s). (see Kim and LKT cases)

n=2		$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$
$\phi = 0$		$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$
		n=2, $\phi = 0$						n=2, $\phi = 0.75$					
I(1)-I(0)-I(1)		T=200						T=500					
$R^2 = 0, \mathbf{A}_0$	$\mathcal{W}(1)$	0.288	0.578	0.406	0.305	0.554	0.340	0.149	0.296	0.144	0.211	0.517	0.246
	$\mathcal{W}(2)$	0.964	0.986	0.972	0.979	0.997	0.980	0.431	0.712	0.498	0.778	0.975	0.869
	$\mathcal{W}(3)$	0.916	0.972	0.942	0.959	0.991	0.959	0.294	0.400	0.316	0.640	0.916	0.782
	$\mathcal{W}(4)$	0.832	0.981	0.926	0.928	0.992	0.947	0.179	0.363	0.281	0.435	0.904	0.719
	\mathcal{W}_{\max}	0.940	0.989	0.956	0.971	0.995	0.969	0.331	0.571	0.350	0.659	0.959	0.806
$R^2 = 0.4, \mathbf{A}_1$	$\mathcal{W}(1)$	0.294	0.410	0.268	0.293	0.446	0.266	0.199	0.259	0.164	0.224	0.409	0.205
	$\mathcal{W}(2)$	0.878	0.958	0.898	0.931	0.984	0.938	0.507	0.636	0.515	0.749	0.935	0.807
	$\mathcal{W}(3)$	0.831	0.932	0.840	0.894	0.961	0.911	0.407	0.421	0.379	0.630	0.836	0.725
	$\mathcal{W}(4)$	0.783	0.938	0.818	0.847	0.973	0.907	0.306	0.402	0.366	0.509	0.808	0.667
	\mathcal{W}_{\max}	0.852	0.942	0.869	0.908	0.975	0.917	0.428	0.509	0.398	0.660	0.899	0.729
I(0)-I(1)-I(0)													
$R^2 = 0, \mathbf{A}_0$	$\mathcal{W}(1)$	0.933	0.948	0.966	0.997	0.981	0.992	0.638	0.451	0.602	0.996	0.883	0.965
	$\mathcal{W}(2)$	0.999	0.959	0.985	1.000	0.967	0.997	0.443	0.245	0.381	0.997	0.806	0.959
	$\mathcal{W}(3)$	0.995	0.924	0.975	1.000	0.969	0.997	0.574	0.379	0.453	0.999	0.846	0.968
	$\mathcal{W}(4)$	0.978	0.949	0.964	1.000	0.962	0.994	0.392	0.218	0.322	0.996	0.787	0.912
	\mathcal{W}_{\max}	0.998	0.958	0.984	0.999	0.973	0.997	0.460	0.332	0.448	0.994	0.804	0.959
$R^2 = 0.4, \mathbf{A}_1$	$\mathcal{W}(1)$	0.986	0.911	0.941	0.996	0.946	0.973	0.503	0.512	0.570	0.972	0.827	0.891
	$\mathcal{W}(2)$	0.995	0.884	0.964	0.999	0.924	0.985	0.398	0.341	0.416	0.974	0.720	0.871
	$\mathcal{W}(3)$	0.999	0.885	0.963	1.000	0.928	0.985	0.500	0.472	0.464	0.984	0.785	0.900
	$\mathcal{W}(4)$	0.990	0.866	0.939	0.999	0.922	0.983	0.377	0.329	0.392	0.939	0.718	0.808
	\mathcal{W}_{\max}	0.996	0.890	0.960	0.999	0.935	0.984	0.407	0.418	0.470	0.953	0.746	0.880

Wald type tests

$R^2=0, A_0$		n=4	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$
I(1)-I(0)-I(1)			$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$
			T=200			T=500		
$\phi = 0$	$\mathcal{W}(1)$	0.359	0.693	0.541	0.389	0.641	0.477	
	$\mathcal{W}(2)$	0.901	0.975	0.941	0.945	0.990	0.965	
	$\mathcal{W}(3)$	0.813	0.959	0.898	0.888	0.967	0.944	
	$\mathcal{W}(4)$	0.706	0.955	0.870	0.819	0.973	0.926	
	$\mathcal{W} \max$	0.853	0.978	0.919	0.909	0.977	0.954	
$\phi = 0.75$	$\mathcal{W}(1)$	0.151	0.411	0.247				
	$\mathcal{W}(2)$	0.380	0.639	0.492				
	$\mathcal{W}(3)$	0.275	0.376	0.321				
	$\mathcal{W}(4)$	0.146	0.309	0.232				
	$\mathcal{W} \max$	0.276	0.513	0.361				
I(0)-I(1)-I(0)								
$\phi = 0$	$\mathcal{W}(1)$	0.914	0.892	0.907	0.990	0.921	0.938	
	$\mathcal{W}(2)$	0.977	0.866	0.966	0.994	0.873	0.954	
	$\mathcal{W}(3)$	0.981	0.847	0.944	0.996	0.888	0.970	
	$\mathcal{W}(4)$	0.936	0.838	0.905	0.994	0.848	0.939	
	$\mathcal{W} \max$	0.981	0.883	0.961	0.990	0.885	0.954	
$\phi = 0.75$	$\mathcal{W}(1)$	0.724	0.430	0.573				
	$\mathcal{W}(2)$	0.460	0.209	0.362				
	$\mathcal{W}(3)$	0.628	0.352	0.472				
	$\mathcal{W}(4)$	0.401	0.182	0.288				
	$\mathcal{W} \max$	0.504	0.274	0.392				

$\tau = 0.5; \phi = 0$	$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)	T=200			T=500		
$R^2 = 0, A_0$	0.462	1.396		0.435	1.500	
	(0.127)	(0.664)		(0.131)	(0.690)	
$R^2 = 0.4, A_1$	0.487	1.632		0.466	1.625	
	(0.171)	(0.770)		(0.175)	(0.702)	
I(1)-I(0)						
$R^2 = 0, A_0$	0.503	1.290		0.513	1.393	
	(0.061)	(0.517)		(0.063)	(0.539)	
$R^2 = 0.4, A_1$	0.502	1.615		0.505	1.669	
	(0.087)	(0.526)		(0.075)	(0.487)	

$\tau = 0.3; \phi = 0$	$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)	T=200			T=500		
$R^2 = 0, A_0$	0.333	1.312		0.312	1.321	
	(0.146)	(0.686)		(0.145)	(0.648)	
$R^2 = 0.4, A_1$	0.452	1.608		0.426	1.601	
	(0.222)	(0.831)		(0.222)	(0.855)	
I(1)-I(0)						
$R^2 = 0, A_0$	0.300	1.165		0.306	1.226	
	(0.054)	(0.444)		(0.057)	(0.465)	
$R^2 = 0.4, A_1$	0.295	1.387		0.308	1.481	
	(0.058)	(0.155)		(0.073)	(0.536)	

$\tau = 0.7; \phi = 0$	$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)	T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	0.665	1.435				
	(0.115)	(0.776)				
$R^2 = 0.4, \mathbf{A}_1$	0.649	1.472				
	(0.149)	(0.672)				
I(1)-I(0)						
$R^2 = 0, \mathbf{A}_0$	0.691	1.313				
	(0.053)	(0.493)				
$R^2 = 0.4, \mathbf{A}_1$	0.677	1.626				
	(0.094)	(0.510)				

$\tau = 0.3; \phi = 0.75$		$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
$R^2 = 0, A_0$		0.525	1.869		0.444	1.597	
		(0.220)	(0.878)		(0.209)	(0.803)	
$R^2 = 0.4, A_1$		0.605	1.968		0.581	1.764	
		(0.198)	(0.906)		(0.207)	(0.855)	
I(1)-I(0)		T=200			T=500		
$R^2 = 0, A_0$		0.323	1.242		0.299	1.156	
		(0.139)	(0.479)		(0.060)	(0.365)	
$R^2 = 0.4, A_1$		0.320	1.409		0.286	1.333	
		(0.148)	(0.551)		(0.071)	(0.502)	

n=4	$R^2 = 0, \mathbf{A}_0$	$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
	$\tau = 0.5$	0.456	1.464		0.431	1.522	
		(0.149)	(0.709)		(0.150)	(0.681)	
	$\tau = 0.3$	0.355	1.393		0.320	1.409	
		(0.189)	(0.717)		(0.164)	(0.719)	
	$\tau = 0.3$	0.533	1.788		0.445	1.710	
		(0.239)	(0.842)		(0.227)	(0.827)	
I(1)-I(0)		T=200			T=500		
	$\tau = 0.5$	0.520	1.593		0.535	1.691	
		(0.100)	(0.598)		(0.106)	(0.530)	
	$\tau = 0.3$	0.317	1.517		0.327	1.492	
		(0.094)	(0.748)		(0.099)	(0.572)	
	$\tau = 0.3$	0.355	1.379		0.327	1.355	
		(0.181)	(0.534)		(0.125)	(0.518)	

n=2		$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m
KPZ	I(1)-I(0)-I(1)	T=200, $\phi = 0$			T=500, $\phi = 0$			T=200, $\phi = 0.75$			T=500, $\phi = 0.75$		
$R^2 = 0, A_0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.376 (0.069)	0.540 (0.168)	2.009 (0.207)	0.379 (0.057)	0.544 (0.171)	2.000 (0.089)	0.362 (0.123)	0.548 (0.191)	1.940 (0.559)	0.368 (0.094)	0.554 (0.186)	2.030 (0.375)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.260 (0.066)	0.498 (0.264)	2.070 (0.381)	0.255 (0.052)	0.480 (0.265)	2.056 (0.323)	0.283 (0.116)	0.519 (0.263)	1.907 (0.500)	0.257 (0.076)	0.495 (0.273)	1.984 (0.264)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.261 (0.061)	0.507 (0.223)	2.009 (0.225)	0.257 (0.050)	0.505 (0.227)	2.013 (0.157)	0.312 (0.140)	0.542 (0.233)	2.033 (0.629)	0.271 (0.096)	0.532 (0.240)	2.014 (0.331)
$R^2 = 0.4, A_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.377 (0.086)	0.520 (0.177)	2.026 (0.281)	0.371 (0.077)	0.516 (0.173)	2.007 (0.225)	0.368 (0.120)	0.527 (0.192)	2.010 (0.586)	0.366 (0.108)	0.522 (0.191)	1.994 (0.400)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.256 (0.076)	0.463 (0.261)	2.045 (0.317)	0.250 (0.061)	0.463 (0.262)	2.024 (0.231)	0.289 (0.136)	0.490 (0.268)	1.984 (0.515)	0.251 (0.089)	0.472 (0.273)	2.018 (0.322)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.262 (0.085)	0.485 (0.238)	2.027 (0.310)	0.254 (0.066)	0.473 (0.237)	2.022 (0.267)	0.315 (0.152)	0.514 (0.245)	2.069 (0.626)	0.281 (0.129)	0.500 (0.251)	2.036 (0.420)
I(0)-I(1)-I(0)													
$R^2 = 0, A_0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.398 (0.044)	0.547 (0.096)	1.879 (0.427)	0.399 (0.044)	0.553 (0.086)	1.950 (0.351)	0.390 (0.109)	0.568 (0.147)	1.408 (0.730)	0.401 (0.058)	0.564 (0.104)	1.262 (0.521)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.269 (0.075)	0.656 (0.190)	1.386 (0.554)	0.251 (0.061)	0.674 (0.180)	1.411 (0.519)	0.365 (0.129)	0.666 (0.156)	1.368 (0.706)	0.319 (0.112)	0.677 (0.168)	1.150 (0.423)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.295 (0.112)	0.575 (0.093)	1.578 (0.590)	0.306 (0.133)	0.586 (0.077)	1.759 (0.511)	0.341 (0.116)	0.569 (0.133)	1.304 (0.616)	0.322 (0.115)	0.583 (0.095)	1.198 (0.439)
$R^2 = 0.4, A_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.417 (0.078)	0.572 (0.081)	1.417 (0.565)	0.424 (0.082)	0.574 (0.068)	1.501 (0.567)	0.409 (0.109)	0.588 (0.114)	1.472 (0.753)	0.419 (0.078)	0.580 (0.081)	1.291 (0.551)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.316 (0.118)	0.703 (0.132)	1.169 (0.450)	0.286 (0.111)	0.711 (0.129)	1.118 (0.377)	0.406 (0.131)	0.687 (0.123)	1.407 (0.761)	0.406 (0.133)	0.712 (0.100)	1.152 (0.465)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.386 (0.164)	0.582 (0.079)	1.486 (0.540)	0.405 (0.176)	0.594 (0.063)	1.537 (0.539)	0.371 (0.125)	0.575 (0.103)	1.456 (0.683)	0.398 (0.138)	0.585 (0.077)	1.357 (0.560)

n=4	$R^2 = 0, \mathbf{A}_0$ I(1)-I(0)-I(1)	$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m
		T=200	T=500				
$\phi = 0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.363 (0.089)	0.530 (0.182)	1.991 (0.273)	0.369 (0.079)	0.531 (0.179)	1.990 (0.209)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.267 (0.084)	0.452 (0.251)	2.090 (0.531)	0.267 (0.080)	0.468 (0.255)	2.102 (0.455)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.263 (0.074)	0.467 (0.226)	1.998 (0.331)	0.262 (0.065)	0.487 (0.229)	2.010 (0.214)
$\phi = 0.75$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.352 (0.140)	0.550 (0.214)	1.948 (0.689)			
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.293 (0.131)	0.488 (0.259)	1.852 (0.598)			
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.309 (0.144)	0.519 (0.248)	1.936 (0.678)			
I(0)-I(1)-I(0)							
$\phi = 0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.392 (0.076)	0.536 (0.114)	1.971 (0.538)	0.402 (0.090)	0.538 (0.109)	1.986 (0.422)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.296 (0.119)	0.646 (0.199)	1.376 (0.564)	0.271 (0.104)	0.637 (0.212)	1.345 (0.557)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.339 (0.154)	0.556 (0.132)	1.741 (0.621)	0.372 (0.175)	0.577 (0.127)	1.846 (0.539)
$\phi = 0.75$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.379 (0.129)	0.555 (0.165)	1.417 (0.752)			
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.391 (0.148)	0.638 (0.185)	1.434 (0.725)			
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.362 (0.138)	0.555 (0.164)	1.468 (0.733)			

On-line Appendix

Sequential Tests

$\tau = 0.5; \phi = 0$		0-1/1-0	Multiple	0-1/1-0	Multiple
I(0)-I(1)		T=200		T=500	
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.497	0.486	0.415	0.446
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.489	0.477	0.410	0.442
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.309	0.291	0.361	0.353
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.302	0.285	0.356	0.348
I(1)-I(0)					
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.535	0.483	0.526	0.460
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.529	0.478	0.519	0.460
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.407	0.284	0.442	0.335
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.396	0.280	0.432	0.332

Sequential Tests

$\tau = 0.3; \phi = 0$		0-1/1-0	Multiple	0-1/1-0	Multiple
I(0)-I(1)		T=200		T=500	
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.328	0.282	0.292	0.288
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.319	0.276	0.286	0.284
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.128	0.140	0.200	0.186
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.122	0.136	0.193	0.182
I(1)-I(0)					
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.861	0.845	0.877	0.829
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.855	0.842	0.872	0.830
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.791	0.703	0.835	0.761
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.784	0.700	0.828	0.760

Sequential Tests

$\tau = 0.7; \phi = 0$		0-1/1-0	Multiple	0-1/1-0	Multiple
I(0)-I(1)		T=200		T=500	
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.716	0.735	0.613	0.686
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.709	0.728	0.608	0.683
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.609	0.577	0.612	0.644
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.603	0.569	0.605	0.639
I(1)-I(0)					
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.264	0.186	0.231	0.204
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.261	0.176	0.226	0.202
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.223	0.092	0.208	0.122
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.211	0.089	0.199	0.119

Sequential Tests

$\tau = 0.3; \phi = 0.75$		0-1/1-0	Multiple	0-1/1-0	Multiple
I(0)-I(1)		T=200		T=500	
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.154	0.127	0.408	0.338
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.151	0.123	0.404	0.333
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.160	0.167	0.239	0.200
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.147	0.158	0.229	0.194
I(1)-I(0)					
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.810	0.723	0.946	0.936
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.795	0.718	0.942	0.935
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.622	0.571	0.825	0.803
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.609	0.566	0.821	0.802

Sequential Tests

$R^2 = A = 0$		n=4	0-1/1-0	Multiple	0-1/1-0	Multiple
I(0)-I(1)			T=200		T=500	
$\tau = 0.5, \phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.471	0.434	0.357	0.400	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.460	0.427	0.351	0.395	
$\tau = 0.3, \phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.220	0.190	0.196	0.208	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.211	0.183	0.192	0.203	
$\tau = 0.3, \phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.094	0.119	0.211	0.165	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.090	0.115	0.207	0.161	
I(1)-I(0)						
$\tau = 0.5, \phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.404	0.354	0.318	0.351	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.397	0.350	0.311	0.348	
$\tau = 0.3, \phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.754	0.731	0.671	0.701	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.747	0.726	0.660	0.699	
$\tau = 0.3, \phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.503	0.421	0.825	0.782	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.483	0.414	0.815	0.780	

Sequential Tests

n=2		$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$
$\phi = 0$		$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$
I(1)-I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.163	0.389	0.414	0.245	0.667	0.442
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.159	0.379	0.406	0.239	0.666	0.439
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.111	0.414	0.179	0.154	0.521	0.270
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.107	0.406	0.174	0.149	0.518	0.266
I(0)-I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.899	0.606	0.660	0.880	0.469	0.668
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.895	0.598	0.647	0.878	0.466	0.664
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.799	0.142	0.395	0.891	0.365	0.612
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.795	0.135	0.389	0.888	0.358	0.606

Sequential Tests

n=2		$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$
$\phi = 0.75$		$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$
I(1)-I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.091	0.384	0.190	0.159	0.753	0.466
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.087	0.375	0.182	0.155	0.753	0.463
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.140	0.342	0.206	0.107	0.547	0.294
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.133	0.332	0.198	0.104	0.542	0.287
I(0)-I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	$MZ_{\hat{\rho}}^{GLS}$	0.654	0.174	0.332	0.955	0.468	0.721
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.649	0.172	0.326	0.953	0.462	0.717
$R^2 = 0.4, \mathbf{A}_1$	$MZ_{\hat{\rho}}^{GLS}$	0.621	0.210	0.345	0.901	0.337	0.588
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.612	0.201	0.337	0.897	0.329	0.582
$R^2 = 0, \mathbf{A}_0$		n=4					
I(1)-I(0)-I(1)		$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$	$\tau_1 = 0.4$	$\tau_1 = 0.25$	$\tau_1 = 0.25$
		$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$	$\tau_2 = 0.6$	$\tau_2 = 0.75$	$\tau_2 = 0.6$
		T=200			T=500		
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.108	0.470	0.267	0.137	0.515	0.290
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.107	0.461	0.256	0.134	0.511	0.285
$\phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.077	0.211	0.127	0.061	0.528	0.233
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.075	0.207	0.123	0.061	0.522	0.228
I(0)-I(1)-I(0)							
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.750	0.233	0.444	0.797	0.284	0.506
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.743	0.229	0.439	0.794	0.281	0.501
$\phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.488	0.141	0.245	0.774	0.197	0.427
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.481	0.136	0.242	0.771	0.194	0.424

Break Dates

Mean and standard deviation in brackets

m : # breakpoints according to w_{\max} (m s.t. $\max w_{\max}(m)$)

Model (??), $p_y = 0$ and $p_x = 0$

$\tau = 0.5; \phi = 0$		$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	KPZ		0.462 (0.127)	1.396 (0.664)		0.435 (0.131)	1.500 (0.690)
	$MZ_{\hat{\rho}}^{GLS}$	0.094 (0.145)	0.601 (0.233)		0.099 (0.148)	0.591 (0.248)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.092 (0.142)	0.599 (0.231)		0.098 (0.147)	0.590 (0.247)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.487 (0.171)	1.632 (0.770)		0.466 (0.175)	1.625 (0.702)
	$MZ_{\hat{\rho}}^{GLS}$	0.130 (0.211)	0.626 (0.228)		0.101 (0.182)	0.585 (0.245)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.125 (0.207)	0.626 (0.227)		0.097 (0.177)	0.588 (0.244)	
I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	KPZ		0.503 (0.061)	1.290 (0.517)		0.513 (0.063)	1.393 (0.539)
	$MZ_{\hat{\rho}}^{GLS}$	0.411 (0.227)	0.922 (0.140)		0.441 (0.196)	0.893 (0.135)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.408 (0.227)	0.923 (0.137)		0.440 (0.196)	0.894 (0.131)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.502 (0.087)	1.615 (0.526)		0.505 (0.075)	1.669 (0.487)
	$MZ_{\hat{\rho}}^{GLS}$	0.385 (0.215)	0.895 (0.198)		0.424 (0.189)	0.897 (0.159)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.382 (0.216)	0.896 (0.192)		0.422 (0.189)	0.897 (0.155)	
<hr/>							
$\tau = 0.3; \phi = 0$		$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	KPZ		0.333 (0.146)	1.312 (0.686)		0.312 (0.145)	1.321 (0.648)
	$MZ_{\hat{\rho}}^{GLS}$	0.084 (0.155)	0.616 (0.270)		0.099 (0.159)	0.594 (0.280)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.082 (0.152)	0.617 (0.269)		0.097 (0.156)	0.594 (0.279)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.452 (0.222)	1.608 (0.831)		0.426 (0.222)	1.601 (0.855)
	$MZ_{\hat{\rho}}^{GLS}$	0.188 (0.229)	0.658 (0.269)		0.145 (0.214)	0.609 (0.283)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.181 (0.225)	0.660 (0.268)		0.140 (0.211)	0.612 (0.282)	
<hr/>							
I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	KPZ		0.300 (0.054)	1.165 (0.444)		0.306 (0.057)	1.226 (0.465)
	$MZ_{\hat{\rho}}^{GLS}$	0.336 (0.151)	0.925 (0.124)		0.374 (0.152)	0.890 (0.135)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.334 (0.151)	0.925 (0.124)		0.373 (0.151)	0.889 (0.135)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.295 (0.058)	1.387 (0.155)		0.308 (0.073)	1.481 (0.536)
	$MZ_{\hat{\rho}}^{GLS}$	0.315 (0.133)	0.941 (0.128)		0.342 (0.124)	0.908 (0.124)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.313 (0.132)	0.939 (0.128)		0.341 (0.124)	0.908 (0.123)	

$\tau = 0.7; \phi = 0$		$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	KPZ		0.665 (0.115)	1.435 (0.776)			
	$MZ_{\hat{\rho}}^{GLS}$	0.096 (0.155)	0.648 (0.206)		0.121 (0.163)	0.602 (0.230)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.096 (0.154)	0.646 (0.205)		0.121 (0.163)	0.600 (0.228)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.649 (0.149)	1.472 (0.672)			
	$MZ_{\hat{\rho}}^{GLS}$	0.077 (0.165)	0.672 (0.196)		0.085 (0.140)	0.601 (0.226)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.076 (0.162)	0.668 (0.196)		0.085 (0.139)	0.600 (0.225)	
I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	KPZ		0.691 (0.053)	1.313 (0.493)			
	$MZ_{\hat{\rho}}^{GLS}$	0.279 (0.307)	0.878 (0.215)		0.356 (0.321)	0.861 (0.212)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.273 (0.306)	0.878 (0.211)		0.350 (0.321)	0.861 (0.211)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.677 (0.094)	1.626 (0.510)			
	$MZ_{\hat{\rho}}^{GLS}$	0.209 (0.277)	0.761 (0.274)		0.292 (0.313)	0.777 (0.262)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.202 (0.276)	0.766 (0.271)		0.285 (0.312)	0.782 (0.259)	
<hr/>							
$\tau = 0.3; \phi = 0.75$		$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m
I(0)-I(1)		T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	KPZ		0.525 (0.220)	1.869 (0.878)		0.444 (0.209)	1.597 (0.803)
	$MZ_{\hat{\rho}}^{GLS}$	0.071 (0.165)	0.666 (0.265)		0.045 (0.136)	0.556 (0.270)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.067 (0.162)	0.665 (0.263)		0.042 (0.131)	0.555 (0.269)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.605 (0.198)	1.968 (0.906)		0.581 (0.207)	1.764 (0.855)
	$MZ_{\hat{\rho}}^{GLS}$	0.142 (0.220)	0.691 (0.265)		0.097 (0.193)	0.620 (0.281)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.135 (0.217)	0.693 (0.262)		0.092 (0.188)	0.622 (0.280)	
I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	KPZ		0.323 (0.139)	1.242 (0.479)		0.299 (0.060)	1.156 (0.365)
	$MZ_{\hat{\rho}}^{GLS}$	0.286 (0.168)	0.953 (0.113)		0.313 (0.094)	0.960 (0.063)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.285 (0.167)	0.951 (0.114)		0.312 (0.093)	0.959 (0.064)	
$R^2 = 0.4, \mathbf{A}_1$	KPZ		0.320 (0.148)	1.409 (0.551)		0.286 (0.071)	1.333 (0.502)
	$MZ_{\hat{\rho}}^{GLS}$	0.298 (0.197)	0.909 (0.155)		0.322 (0.114)	0.937 (0.092)	
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.294 (0.197)	0.910 (0.153)		0.322 (0.114)	0.936 (0.092)	

$R^2 = 0, \mathbf{A}_0$	n=4	$\hat{\lambda}$	$\hat{\tau}$	m	$\hat{\lambda}$	$\hat{\tau}$	m	
I(0)-I(1)		T=200			T=500			
$\tau = 0.5$	KPZ		0.456 (0.149)	1.464 (0.709)		0.431 (0.150)	1.522 (0.681)	
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.076 (0.137)	0.677 (0.246)		0.090 (0.133)	0.634 (0.268)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.075 (0.137)	0.677 (0.245)		0.089 (0.132)	0.637 (0.268)		
$\tau = 0.3$	KPZ		0.355 (0.189)	1.393 (0.717)		0.320 (0.164)	1.409 (0.719)	
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.068 (0.150)	0.748 (0.268)		0.084 (0.138)	0.725 (0.289)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.067 (0.149)	0.751 (0.264)		0.083 (0.137)	0.727 (0.286)		
$\tau = 0.3$	KPZ		0.533 (0.239)	1.788 (0.842)		0.445 (0.227)	1.710 (0.827)	
$\phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.065 (0.163)	0.812 (0.238)		0.036 (0.124)	0.732 (0.281)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.064 (0.162)	0.813 (0.235)		0.034 (0.120)	0.731 (0.279)		
I(1)-I(0)								
$\tau = 0.5$	KPZ		0.520 (0.100)	1.593 (0.598)		0.535 (0.106)	1.691 (0.530)	
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.318 (0.260)	0.935 (0.126)		0.373 (0.267)	0.910 (0.126)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.314 (0.260)	0.935 (0.124)		0.370 (0.267)	0.910 (0.122)		
$\tau = 0.3$	KPZ		0.317 (0.094)	1.517 (0.748)		0.327 (0.099)	1.492 (0.572)	
$\phi = 0$	$MZ_{\hat{\rho}}^{GLS}$	0.307 (0.163)	0.934 (0.117)		0.391 (0.175)	0.884 (0.128)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.306 (0.163)	0.934 (0.116)		0.389 (0.176)	0.885 (0.127)		
$\tau = 0.3$	KPZ		0.355 (0.181)	1.379 (0.534)		0.327 (0.125)	1.355 (0.518)	
$\phi = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.243 (0.196)	0.948 (0.123)		0.312 (0.126)	0.961 (0.063)		
	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.241 (0.195)	0.946 (0.123)		0.311 (0.125)	0.959 (0.064)		
n=2								
KPZ	$\phi = 0$		$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m
$R^2 = 0, \mathbf{A}_0$	I(1)-I(0)-I(1)		T=200			T=500		
	$\tau_1 = 0.4, \tau_2 = 0.6$	0.376 (0.069)	0.540 (0.168)	2.009 (0.207)	0.379 (0.057)	0.544 (0.171)	2.000 (0.089)	
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.260 (0.066)	0.498 (0.264)	2.070 (0.381)	0.255 (0.052)	0.480 (0.265)	2.056 (0.323)	
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.261 (0.061)	0.507 (0.223)	2.009 (0.225)	0.257 (0.050)	0.505 (0.227)	2.013 (0.157)	
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.377 (0.086)	0.520 (0.177)	2.026 (0.281)	0.371 (0.077)	0.516 (0.173)	2.007 (0.225)	
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.256 (0.076)	0.463 (0.261)	2.045 (0.317)	0.250 (0.061)	0.463 (0.262)	2.024 (0.231)	
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.262 (0.085)	0.485 (0.238)	2.027 (0.310)	0.254 (0.066)	0.473 (0.237)	2.022 (0.267)	
I(0)-I(1)-I(0)								
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.398 (0.044)	0.547 (0.096)	1.879 (0.427)	0.399 (0.044)	0.553 (0.086)	1.950 (0.351)	
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.269 (0.075)	0.656 (0.190)	1.386 (0.554)	0.251 (0.061)	0.674 (0.180)	1.411 (0.519)	
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.295 (0.112)	0.575 (0.093)	1.578 (0.590)	0.306 (0.133)	0.586 (0.077)	1.759 (0.511)	
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.417 (0.078)	0.572 (0.081)	1.417 (0.565)	0.424 (0.082)	0.574 (0.068)	1.501 (0.567)	
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.316 (0.118)	0.703 (0.132)	1.169 (0.450)	0.286 (0.111)	0.711 (0.129)	1.118 (0.377)	
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.386 (0.164)	0.582 (0.079)	1.486 (0.540)	0.405 (0.176)	0.594 (0.063)	1.537 (0.539)	

n=2	$\phi = 0$		$\hat{\lambda}$	$\hat{\tau}$	$\hat{\lambda}$	$\hat{\tau}$	
	I(1)-I(0)-I(1)		T=200		T=500		
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.236 (0.205)	0.809 (0.191)	0.265 (0.193)	0.775 (0.179)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.231 (0.204)	0.806 (0.188)	0.261 (0.193)	0.775 (0.177)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.237 (0.161)	0.777 (0.159)	0.244 (0.137)	0.738 (0.159)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.234 (0.160)	0.773 (0.159)	0.242 (0.137)	0.736 (0.159)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.185 (0.163)	0.739 (0.192)	0.199 (0.143)	0.699 (0.186)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.182 (0.163)	0.735 (0.191)	0.196 (0.142)	0.699 (0.184)	
	$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.265 (0.231)	0.773 (0.227)	0.274 (0.207)	0.757 (0.203)
		$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.258 (0.229)	0.775 (0.222)	0.270 (0.207)	0.758 (0.199)
		$\tau_1 = 0.25$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.231 (0.159)	0.800 (0.167)	0.232 (0.124)	0.768 (0.152)
		$\tau_2 = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.228 (0.159)	0.799 (0.164)	0.230 (0.124)	0.766 (0.151)
		$\tau_1 = 0.25$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.220 (0.192)	0.745 (0.206)	0.205 (0.158)	0.709 (0.195)
		$\tau_2 = 0.6$	$MZ_{\hat{\rho}}^{GLS}$	0.214 (0.190)	0.743 (0.204)	0.201 (0.156)	0.707 (0.193)
I(0)-I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.177 (0.255)	0.711 (0.275)	0.230 (0.284)	0.629 (0.284)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.181 (0.257)	0.718 (0.277)	0.231 (0.285)	0.631 (0.286)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.127 (0.243)	0.630 (0.312)	0.187 (0.295)	0.591 (0.321)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.128 (0.243)	0.637 (0.313)	0.188 (0.295)	0.595 (0.323)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.295 (0.287)	0.762 (0.301)	0.312 (0.299)	0.684 (0.326)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.295 (0.287)	0.768 (0.299)	0.314 (0.299)	0.688 (0.325)	
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.138 (0.244)	0.740 (0.286)	0.181 (0.272)	0.606 (0.289)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.137 (0.244)	0.744 (0.287)	0.182 (0.272)	0.608 (0.291)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.126 (0.238)	0.626 (0.316)	0.171 (0.294)	0.570 (0.327)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.122 (0.235)	0.634 (0.317)	0.170 (0.294)	0.573 (0.329)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.245 (0.285)	0.757 (0.319)	0.270 (0.300)	0.662 (0.344)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.242 (0.284)	0.764 (0.316)	0.270 (0.299)	0.664 (0.343)	

n=2	$\phi = 0.75$	$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m
KPZ	I(1)-I(0)-I(1)	T=200			T=500		
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.362 (0.123)	0.548 (0.191)	1.940 (0.559)	0.368 (0.094)	0.554 (0.186)	2.030 (0.375)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.283 (0.116)	0.519 (0.263)	1.907 (0.500)	0.257 (0.076)	0.495 (0.273)	1.984 (0.264)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.312 (0.140)	0.542 (0.233)	2.033 (0.629)	0.271 (0.096)	0.532 (0.240)	2.014 (0.331)
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.368 (0.120)	0.527 (0.192)	2.010 (0.586)	0.366 (0.108)	0.522 (0.191)	1.994 (0.400)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.289 (0.136)	0.490 (0.268)	1.984 (0.515)	0.251 (0.089)	0.472 (0.273)	2.018 (0.322)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.315 (0.152)	0.514 (0.245)	2.069 (0.626)	0.281 (0.129)	0.500 (0.251)	2.036 (0.420)
I(0)-I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.390 (0.109)	0.568 (0.147)	1.408 (0.730)	0.401 (0.058)	0.564 (0.104)	1.262 (0.521)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.365 (0.129)	0.666 (0.156)	1.368 (0.706)	0.319 (0.112)	0.677 (0.168)	1.150 (0.423)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.341 (0.116)	0.569 (0.133)	1.304 (0.616)	0.322 (0.115)	0.583 (0.095)	1.198 (0.439)
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.409 (0.109)	0.588 (0.114)	1.472 (0.753)	0.419 (0.078)	0.580 (0.081)	1.291 (0.551)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.406 (0.131)	0.687 (0.123)	1.407 (0.761)	0.406 (0.133)	0.712 (0.100)	1.152 (0.465)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.371 (0.125)	0.575 (0.103)	1.456 (0.683)	0.398 (0.138)	0.585 (0.077)	1.357 (0.560)

n=2	$\phi = 0.75$		$\hat{\lambda}$	$\hat{\tau}$	$\hat{\lambda}$	$\hat{\tau}$	
	I(1)-I(0)-I(1)		T=200		T=500		
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.216 (0.224)	0.805 (0.207)	0.266 (0.202)	0.771 (0.183)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.210 (0.222)	0.803 (0.202)	0.261 (0.202)	0.772 (0.180)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.198 (0.165)	0.836 (0.148)	0.223 (0.110)	0.803 (0.113)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.196 (0.164)	0.831 (0.147)	0.222 (0.110)	0.801 (0.112)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.166 (0.171)	0.781 (0.191)	0.188 (0.132)	0.733 (0.164)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.162 (0.169)	0.777 (0.188)	0.187 (0.132)	0.731 (0.162)	
	$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.213 (0.235)	0.786 (0.229)	0.243 (0.218)	0.780 (0.207)
		$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.203 (0.232)	0.788 (0.223)	0.235 (0.216)	0.783 (0.202)
		$\tau_1 = 0.25$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.211 (0.197)	0.815 (0.183)	0.221 (0.138)	0.796 (0.147)
		$\tau_2 = 0.75$	$MZ_{\hat{\rho}}^{GLS}$	0.205 (0.195)	0.812 (0.181)	0.219 (0.138)	0.793 (0.147)
		$\tau_1 = 0.25$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.183 (0.206)	0.777 (0.210)	0.180 (0.162)	0.744 (0.196)
		$\tau_2 = 0.6$	$MZ_{\hat{\rho}}^{GLS}$	0.175 (0.203)	0.777 (0.206)	0.175 (0.159)	0.743 (0.193)
I(0)-I(1)-I(0)							
$R^2 = 0, \mathbf{A}_0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.083 (0.201)	0.824 (0.261)	0.090 (0.214)	0.739 (0.287)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.081 (0.200)	0.826 (0.259)	0.089 (0.213)	0.738 (0.288)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.088 (0.215)	0.679 (0.309)	0.106 (0.252)	0.545 (0.317)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.085 (0.212)	0.683 (0.309)	0.104 (0.250)	0.545 (0.317)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.171 (0.271)	0.798 (0.295)	0.244 (0.300)	0.721 (0.330)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.169 (0.269)	0.804 (0.291)	0.242 (0.299)	0.723 (0.329)	
$R^2 = 0.4, \mathbf{A}_1$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.128 (0.234)	0.758 (0.285)	0.134 (0.249)	0.687 (0.295)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.124 (0.232)	0.763 (0.285)	0.133 (0.249)	0.690 (0.297)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.113 (0.223)	0.674 (0.299)	0.126 (0.262)	0.571 (0.314)	
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.108 (0.219)	0.680 (0.299)	0.125 (0.260)	0.575 (0.316)	
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.176 (0.265)	0.732 (0.310)	0.223 (0.293)	0.667 (0.336)	
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.172 (0.264)	0.742 (0.307)	0.223 (0.293)	0.671 (0.336)	

n=4	$R^2 = 0, \mathbf{A}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$	m	$\hat{\tau}_1$	$\hat{\tau}_2$	m
KPZ	I(1)-I(0)-I(1)	T=200			T=500		
$\phi = 0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.363 (0.089)	0.530 (0.182)	1.991 (0.273)	0.369 (0.079)	0.531 (0.179)	1.990 (0.209)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.267 (0.084)	0.452 (0.251)	2.090 (0.531)	0.267 (0.080)	0.468 (0.255)	2.102 (0.455)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.263 (0.074)	0.467 (0.226)	1.998 (0.331)	0.262 (0.065)	0.487 (0.229)	2.010 (0.214)
$\phi = 0.75$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.352 (0.140)	0.550 (0.214)	1.948 (0.689)			
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.293 (0.131)	0.488 (0.259)	1.852 (0.598)			
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.309 (0.144)	0.519 (0.248)	1.936 (0.678)			
I(0)-I(1)-I(0)							
$\phi = 0$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.392 (0.076)	0.536 (0.114)	1.971 (0.538)	0.402 (0.090)	0.538 (0.109)	1.986 (0.422)
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.296 (0.119)	0.646 (0.199)	1.376 (0.564)	0.271 (0.104)	0.637 (0.212)	1.345 (0.557)
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.339 (0.154)	0.556 (0.132)	1.741 (0.621)	0.372 (0.175)	0.577 (0.127)	1.846 (0.539)
$\phi = 0.75$	$\tau_1 = 0.4, \tau_2 = 0.6$	0.379 (0.129)	0.555 (0.165)	1.417 (0.752)			
	$\tau_1 = 0.25, \tau_2 = 0.75$	0.391 (0.148)	0.638 (0.185)	1.434 (0.725)			
	$\tau_1 = 0.25, \tau_2 = 0.6$	0.362 (0.138)	0.555 (0.164)	1.468 (0.733)			

n=4	$R^2 = A = 0$		$\hat{\lambda}$	$\hat{\tau}$	$\hat{\lambda}$	$\hat{\tau}$
	I(1)-I(0)-I(1)		T=200		T=500	
$\phi = 0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.164 (0.203)	0.884 (0.168)	0.186 (0.198)	0.861 (0.172)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.159 (0.201)	0.882 (0.165)	0.182 (0.197)	0.860 (0.169)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.207 (0.171)	0.816 (0.162)	0.248 (0.154)	0.765 (0.170)
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.204 (0.170)	0.814 (0.161)	0.246 (0.154)	0.764 (0.170)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.154 (0.165)	0.821 (0.188)	0.182 (0.150)	0.768 (0.202)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.151 (0.165)	0.819 (0.187)	0.181 (0.150)	0.767 (0.201)
$\phi = 0.75$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.146 (0.212)	0.872 (0.189)	0.174 (0.209)	0.868 (0.177)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.141 (0.209)	0.871 (0.184)	0.171 (0.208)	0.868 (0.172)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.161 (0.178)	0.875 (0.154)	0.212 (0.133)	0.837 (0.130)
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.159 (0.177)	0.871 (0.152)	0.210 (0.133)	0.835 (0.129)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.133 (0.178)	0.861 (0.177)	0.161 (0.143)	0.814 (0.176)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.131 (0.176)	0.857 (0.177)	0.159 (0.143)	0.813 (0.175)
I(0)-I(1)-I(0)						
$\phi = 0$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.143 (0.235)	0.768 (0.267)	0.265 (0.303)	0.680 (0.300)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.144 (0.235)	0.775 (0.267)	0.265 (0.302)	0.681 (0.300)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.087 (0.187)	0.782 (0.271)	0.143 (0.251)	0.721 (0.301)
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.085 (0.184)	0.785 (0.269)	0.141 (0.249)	0.725 (0.301)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.225 (0.271)	0.844 (0.245)	0.315 (0.312)	0.784 (0.292)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.222 (0.269)	0.847 (0.243)	0.312 (0.311)	0.786 (0.290)
$\phi = 0.75$	$\tau_1 = 0.4$	$MZ_{\hat{\rho}}^{GLS}$	0.060 (0.173)	0.886 (0.223)	0.074 (0.201)	0.827 (0.266)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.058 (0.171)	0.889 (0.220)	0.073 (0.201)	0.830 (0.265)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.062 (0.165)	0.812 (0.259)	0.064 (0.191)	0.732 (0.304)
	$\tau_2 = 0.75$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.060 (0.163)	0.815 (0.257)	0.062 (0.188)	0.733 (0.303)
	$\tau_1 = 0.25$	$MZ_{\hat{\rho}}^{GLS}$	0.104 (0.221)	0.886 (0.228)	0.186 (0.288)	0.852 (0.264)
	$\tau_2 = 0.6$	$MZ_{t_{\hat{\rho}}}^{GLS}$	0.101 (0.219)	0.888 (0.223)	0.182 (0.287)	0.852 (0.262)

LKT CVs

Table 6: Critical Values for the LKT Tests

	LKT (0-1/1-0), Model 1					LKT (Multiple), Model 1				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
	$MZ_{\hat{\rho}}^{GLS}$					$MZ_{\hat{\rho}}^{GLS}$				
1.0%	-27.237	-31.615	-37.459	-42.219	-47.221	-29.953	-36.345	-41.903	-46.796	-52.448
2.5%	-23.115	-27.688	-33.016	-37.584	-42.169	-26.120	-31.568	-36.582	-41.701	-47.248
5.0%	-19.627	-24.196	-28.919	-33.636	-38.134	-22.812	-27.764	-32.959	-37.549	-42.697
7.5%	-17.639	-22.288	-26.503	-31.127	-35.571	-20.640	-25.258	-30.379	-35.097	-39.810
10.0%	-16.202	-20.780	-24.834	-29.253	-33.517	-19.087	-23.687	-28.617	-33.012	-37.869
15.0%	-14.278	-18.605	-22.552	-26.784	-30.761	-16.753	-21.415	-26.040	-30.088	-34.725
20.0%	-12.792	-16.977	-20.733	-24.742	-28.857	-15.116	-19.636	-23.977	-27.969	-32.598
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.685	-3.971	-4.326	-4.589	-4.853	-3.857	-4.247	-4.568	-4.832	-5.105
2.5%	-3.392	-3.711	-4.058	-4.328	-4.589	-3.599	-3.962	-4.269	-4.558	-4.853
5.0%	-3.126	-3.472	-3.799	-4.096	-4.363	-3.360	-3.713	-4.049	-4.325	-4.612
7.5%	-2.962	-3.333	-3.634	-3.940	-4.213	-3.197	-3.540	-3.886	-4.178	-4.450
10.0%	-2.840	-3.217	-3.518	-3.818	-4.090	-3.068	-3.430	-3.773	-4.050	-4.341
15.0%	-2.663	-3.045	-3.351	-3.655	-3.918	-2.875	-3.264	-3.597	-3.869	-4.160
20.0%	-2.521	-2.907	-3.213	-3.510	-3.793	-2.735	-3.120	-3.451	-3.729	-4.027
	LKT (0-1/1-0), Model 2					LKT (Multiple), Model 2				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	n=2	n=3	n=4	n=5	n=6
	$MZ_{\hat{\rho}}^{GLS}$					$MZ_{\hat{\rho}}^{GLS}$				
1.0%	-25.908	-31.458	-37.334	-41.317	-47.494	-29.872	-35.510	-41.824	-46.145	-53.840
2.5%	-22.154	-27.514	-32.888	-37.204	-41.981	-25.265	-31.212	-36.520	-41.589	-47.766
5.0%	-19.109	-23.917	-29.407	-32.993	-37.908	-22.337	-28.046	-32.750	-37.851	-43.231
7.5%	-17.307	-21.901	-27.165	-30.749	-35.501	-20.478	-25.618	-30.324	-35.039	-40.287
10.0%	-15.871	-20.455	-25.467	-29.108	-33.556	-19.023	-23.771	-28.610	-33.163	-38.373
15.0%	-14.127	-18.582	-23.092	-26.611	-30.672	-16.699	-21.344	-26.281	-30.150	-35.058
20.0%	-12.662	-16.897	-21.220	-24.779	-28.636	-15.008	-19.585	-24.046	-28.258	-32.619
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.596	-3.960	-4.313	-4.540	-4.871	-3.849	-4.194	-4.565	-4.787	-5.185
2.5%	-3.322	-3.703	-4.052	-4.307	-4.575	-3.544	-3.941	-4.265	-4.549	-4.879
5.0%	-3.084	-3.454	-3.831	-4.055	-4.348	-3.329	-3.736	-4.033	-4.340	-4.644
7.5%	-2.931	-3.302	-3.683	-3.915	-4.209	-3.187	-3.570	-3.881	-4.174	-4.483
10.0%	-2.811	-3.191	-3.562	-3.810	-4.092	-3.073	-3.437	-3.772	-4.061	-4.367
15.0%	-2.651	-3.042	-3.393	-3.643	-3.911	-2.872	-3.255	-3.611	-3.873	-4.173
20.0%	-2.506	-2.901	-3.254	-3.513	-3.779	-2.726	-3.118	-3.458	-3.748	-4.029
	LKT (0-1/1-0), Model 3					LKT (Multiple), Model 3				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
	$MZ_{\hat{\rho}}^{GLS}$					$MZ_{\hat{\rho}}^{GLS}$				
1.0%	-31.139	-35.274	-40.825	-46.468	-52.837	-35.591	-40.358	-46.812	-52.033	-58.732
2.5%	-27.334	-31.466	-36.356	-41.545	-46.725	-31.192	-35.813	-41.621	-46.930	-52.722
5.0%	-24.165	-28.654	-32.932	-37.641	-42.187	-27.070	-32.500	-37.494	-42.111	-47.531
7.5%	-22.239	-26.417	-30.642	-34.963	-39.466	-24.957	-30.150	-34.728	-39.534	-44.478
10.0%	-20.657	-24.798	-28.860	-32.949	-37.459	-23.279	-27.927	-33.167	-37.464	-42.443
15.0%	-18.436	-22.451	-26.173	-30.455	-34.598	-20.866	-25.386	-30.239	-34.351	-39.022
20.0%	-16.788	-20.691	-24.363	-28.190	-32.332	-19.088	-23.352	-28.222	-32.124	-36.738
	$MZ_{t_{\hat{\rho}}}^{GLS}$					$MZ_{t_{\hat{\rho}}}^{GLS}$				
1.0%	-3.942	-4.196	-4.507	-4.815	-5.130	-4.201	-4.491	-4.834	-5.098	-5.409

One-pass incremental-Learning of temporal patterns with a bounded memory constraint

Koki Ando*, Koichiro Yamauchi

Chubu University
1200, Matsumoto-cho,
Kasugai-shi, Aichi 487-8501, JAPAN
tp17003@sti.chubu.ac.jp, k_yamauchi@isc.chubu.ac.jp
<http://sakura.cs.chubu.ac.jp>

Abstract. This paper proposes a scheme for predicting time series that utilizes one-pass learning on a bounded memory constraint. The term "one-pass" refers to a method of machine learning in which a machine can learn to predict target outputs without a large number of rehearsals. The proposed network consists of the echo state network (ESN) followed by a kernel perceptron. We show that this scheme is suitable not only for realizing one-pass learning of the time series data, but also for reducing the size of ESN without degrading the performance. Simulation results suggest that the proposed method reduces interference caused by the additional learning for the new time-series inputs.

Keywords: Time Series, One-pass Learning, Learning with memory constraint, Reservoir Computing, Learning on a memory budget

1 Introduction

Incremental machine learning has been well explored in various practical applications. For learning static patterns, there are a large number of learning methods, (e.g. [1], [2] and [3]). Although there are numerous methods for incremental learning of temporal patterns, including those presented in [4] and [5], their number is much less than those of static patterns. The major difficulty of incremental learning is avoiding forgetting due to the interference caused by the additional modifications of parameters. In the case of temporal pattern processing, the adverse effects caused by the interference of current time also affect follow-on processing.

Temporal pattern processing applies various methods depending on the task. In the case of sequence recognition, the recognizers only need state transition information, and time duration information about each state can be eliminated. For speech recognition tasks, the recognizer should be robust to deformation not only in time but also in formant frequency pattern. To tackle such deformations, hidden Markov models (HMMs) are often used as the recognizer. In time series prediction, the learner is expected to approximate the relationship between the output signals from the echo state network (ESN) and the desired outputs of a target network.

In this study, we assume that the learner can oversee the prediction of the next output value for an unknown target system. This task is frequently performed by the forward

controller of motor systems. The forward controller predicts optimal control signal with which to control the target object. The control signal is not only a product the current situation, it is also influenced by the past status of the target object. Such controllers are usually embedded into small devices. To enable installation in such small devices, the learning algorithm must be executable within a fixed memory constraint.

In this paper, we propose a one-pass learning method for a fixed memory constraint for time series prediction based on an ESN. Normally, the outputs of the ESN are sent to a single layered perceptron. For successful one-pass learning, the catastrophic forgetting due to the new learning of new time series must be avoided. In Section 2.1, we will discuss the introduction of kernel machines to overcome the aforementioned issue. Hence, a weighted projection method, which was proposed in our previous paper [6], is applied to enable the learning of temporal patterns. The important management issue for an ESN-based system in embedded systems reducing the size of reservoir and realizing the hyper-parameter optimization of the learner.

The remainder of this paper is organized as follows: In Section 2, related works are described. In Section 3.1 the proposed system and methods are outlined and discussed. In Section 4, the experimental results are discussed in detail. Finally in Section 5, we present our conclusions.

2 Related works

Recently, transfer learning methods have been used to adjust pretrained deep neural networks to current new instances [7]. Some researchers might believe that incremental learning is equivalent to the transfer learning. However, the transfer learning methods, which are intended to progress learning with a small number of samples, are different from the incremental learning. Therefore, they do not need to consider that the fact that the current environment is going to change. On the other hand, in the case of incremental learning, the system must consider the possibility that the current environment might change. If we assume that the environment is going back a previous state, then the system must not forget previously learned skills. Several incremental learning algorithms for time series patterns have been developed. Some of them use the phrase "incremental" to refer to the on-line updating of internal parameters (e.g. [8]). In [8], the reservoir networks are incrementally adjusted to maximize entropy for better learning in the subsequent backpropagation-decorrelation learning process. In [9], a fuzzy learning method for incremental learning of time series is presented. In this method, they presented a light weighted gradient based learning method for the fuzzy learning module. Although their model does not use the reservoir, they demonstrated that their proposed method learns the target time series well. Their model does not aim to reduce interference due the additional learning. Incremental learning based on the extreme learning method of an ESN was proposed in [10]. In their method, the size of elements of an ESN is increased by the learning process, where the added unit's parameters are randomly determined. The weights connected to the output units are adjusted by the least square method. Their method, however, relies on creating matrices to optimize the parameters. This is the equivalent to storing all given samples.

Conversely, our proposed method stores only the parameters for the kernels and the pre-determined ESN. Moreover, our method realizes one-pass learning of the time series data.

An ESN with Gaussian process was proposed in [11]. In their method, the adjustable parameters are optimized using the Gaussian process. They also introduced a recursive kernel. The hyperparameters for the recursive kernels are also optimized through the learning process. The learning process continues until the hyperparameters converge at the optimal values. Although the resultant network is well optimized, this method does not realize one-pass learning.

2.1 Learning machines suitable for incremental learning with the echo state network

Generally, the outputs from the reservoir are sent to a single layered neural network. The network learns new input patterns by using the Hebbian learning method. Thus, the output of the network is described as follows. For simplicity, let us assume that the output function is a linear function and $y[\mathbf{x}_t] = \mathbf{W}_{out}^T \mathbf{x}_t$, where \mathbf{W}_{out} is the output connections between reservoir and the output cell and is updated incrementally without referencing past samples. For example, let us assume that the single layered perceptron learns the samples by applying Hebbian learning and it has learned the first learning sample as $\mathbf{W}_{out} = \eta y_1 \mathbf{x}_1$, where $0 < \eta < 1$. Here, $\mathbf{W}_{out}^T \mathbf{x}_1 = \eta y_1 \|\mathbf{x}_1\|^2$. Hence, it functions as an associative memory that recalls the value which proportional to y_1 . Next it learns the new (\mathbf{x}_2, y_2) incrementally, the weight vector is $\mathbf{W}_{out} = \eta(y_1 \mathbf{x}_1) + y_2 \mathbf{x}_2$. In this case, the output for the previous input is $\mathbf{W}_{out}^T \mathbf{x}_1 = \eta(y_1 \|\mathbf{x}_1\|^2 + y_2 \mathbf{x}_2^t \mathbf{x}_1)$. If \mathbf{x}_1 and \mathbf{x}_2 are independent of each other, we obtain $\mathbf{W}_{out}^T \mathbf{x}_1 = \eta y_1 \|\mathbf{x}_1\|^2$. This means that the single layered perceptron successfully recalls y_1 even after the incremental learning of (\mathbf{x}_2, y_2) .

However, if \mathbf{x}_1 is linearly dependent on \mathbf{x}_2 , the network cannot recall the previously learned sample correctly because $\mathbf{x}_2^T \mathbf{x}_1 \neq 0$. Thus, we have to employ decorrelation backpropagation[8] to avoid this situation. However, the past samples must be repeatedly provided for rehearsal. To overcome this difficulty, we can use the kernel method. Using the kernel method, we can convert \mathbf{x}_t to a high dimensional $\mathbf{k}(\mathbf{x}_t, \cdot)$, whose number of dimension is infinite. To recall a specified pattern, we can use a mathematical technique, namely ‘kernel trick.’ The kernel trick means that the dot product of two infinite dimensional vectors $\mathbf{k}(\mathbf{x}, \cdot)$ and $\mathbf{k}(\mathbf{y}, \cdot)$ is derived by $\langle \mathbf{k}(\mathbf{x}_t, \cdot), \mathbf{k}(\mathbf{y}, \cdot) \rangle = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_t\|^2}{\sigma^2}\right)$. Assume that $\mathbf{W}_{out} = a_1 \mathbf{k}(\mathbf{x}_1, \cdot) + a_2 \mathbf{k}(\mathbf{x}_2, \cdot) + \dots$. If the converted input $\mathbf{k}(\mathbf{x}_t, \cdot)$ of the new input \mathbf{x}_t is similar to $\mathbf{k}(\mathbf{x}_1, \cdot)$, then $\langle \mathbf{k}(\mathbf{x}_t, \cdot), \mathbf{W} \rangle \simeq a_1$. Note that $\langle \mathbf{k}(\mathbf{x}_t, \cdot), \mathbf{k}(\mathbf{W}, \cdot) \rangle = \sum_i a_i \exp(-\gamma \|\mathbf{x}_t - \mathbf{x}_i\|^2)$. Additionally, the weighted sum of the Gaussian kernels is one of the forms of the kernel perceptron, whose kernel function is one of the reproductive kernels. This means that the kernel perceptron is suitable for learning the relationships between the input time series and target output. Many online learning algorithms for kernel perceptrons have been proposed [12] [13] [14] [15]. Almost every methods is aimed at solving clustering problems. We have already proposed an incremental learning algorithm using the kernel perceptron for solving regression problems [6]. The method aims to record a novel instance within a

fixed number of kernels without forgetting the previously learned memory. To realize this, we have introduced an adaptive projection method, called the weighted projection method (W_{prj}), to minimize the forgotten memory. In this paper, we introduce a modified version of our previous method for the one-pass learning of time series data.

3 One-pass learning of temporal patterns

3.1 Outline

The proposed method consists of two parts: the ESN and the kernel perceptron. As stated in the introduction, the ESN generates various time series features.

The output from each unit is described by the vector at time t as $\mathbf{S}(t)$, which is also regarded as the current status. The next output status is derived from the current status by the following equation:

$$\mathbf{S}(t) = f[\mathbf{W}_{in}\mathbf{U}(t) + \mathbf{W}\mathbf{S}(t-1)], \quad (1)$$

where $\mathbf{U}(t)$ denotes the input vector to the ESN at time t , \mathbf{W}_{in} and \mathbf{W} are the connection weights between the inputs and the units, and the connections between units, respectively. \mathbf{W}_{in} and \mathbf{W} are determined randomly and fixed beforehand. In fact, the final performance of the system is sensitive to the weights.

In this study, we proposed a new method that adjusts a weight parameter to determine an appropriate sensitivity for each output feature. Given that the output vector from the ESN includes various redundant features, the learner must reduce the importance weight of these redundant dimensions. In this paper, an additional method to solve this problem is proposed. The method is described in detail in Section 3.3.

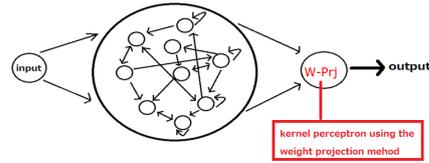


Fig. 1: Proposed method.

3.2 Weighted projection method

The outputs from echo state network ($\mathbf{S}(t)$) are sent to a learning machine that performs incremental learning (one-pass learning) on a constraint.

In this study, we chose a weighted projection method for kernel perceptron, as proposed in our previous study [16].

The output function of the kernel perceptron is represented by

$$f_{t-1}(\mathbf{S}(t)) = \sum_{i=1}^B w_i k(\mathbf{S}(t), \mathbf{x}_i), \quad (2)$$

where $k(\mathbf{x}, \mathbf{x}_i)$ denotes a reproductive kernel function and

$$k(\mathbf{S}(t), \mathbf{x}_i) \equiv \exp\left(-\frac{\|\mathbf{S}(t) - \mathbf{x}_i\|^2}{2\sigma^2}\right), \quad (3)$$

The number of kernels is limited to B . This reflects the fact that the weighted projection method aims to learn the new data within memory constraint B .

The Gaussian kernel is a type of reproductive kernel, hence, its learning algorithm is represented by algebraic manipulation by using the vector on Hilbert space. Let us denote Equation (2). Equation (4) define f_{t-1} as follows:

$$f_t \equiv \sum_{i=1}^B w_i k(\cdot, \mathbf{x}_i) \quad (4)$$

The learning algorithm of W_{prj} will now be described using the vector f_t .

In the early steps of learning, the W_{prj} algorithm is the same as the original kernel perceptron. When a new instance (\mathbf{x}_t, y_t) is presented, W_{prj} adds a new kernel function, whose centroid is \mathbf{x}_t . Now, f_t is given as follows:

$$f_t = f_{t-1} + e_t k(\cdot, \mathbf{x}_t) \quad (5)$$

where $k(\cdot, \mathbf{x}_t)$ denotes a converted vector of \mathbf{x}_t and $e_t = y_t - \langle f_{t-1}, k(\cdot, \mathbf{x}_t) \rangle$. where $\langle k(\cdot, \mathbf{x}_t), k(\cdot, \mathbf{x}) \rangle = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_t\|^2}{2\sigma^2}\right)$ Note that e_t denotes the current error of the kernel perceptron, which is different from the label of the sample. This is suitable for one-pass learning because it reduces the resultant error even if the distribution of kernels is nested. Let S_t be the support set: then by applying Equation (5), S_{t-1} is updated as $S_t = S_{t-1} \cup \{t\}$.

However, $|S_t|$ is bounded to B . If $|S_t| = B$, then W_{prj} cannot append a new kernel. In this case, the W_{prj} prunes one of the existing kernels and replaces it with the new kernel. To determine which kernel to prune, W_{prj} calculates an approximated linear dependency for each kernel. If the vector $k(\cdot, \mathbf{u}_i)$ is linearly dependent on other kernels, and $k(\cdot, \mathbf{u}_i)$ can be approximated by the linear combination of the other kernels, then the output of W_{prj} can be represented without the i -th kernel and it is pruned. The approximated linear dependency of the i -th kernel is

$$\delta_i = \arg \min_{\mathbf{a}} \left\| k(\cdot, \mathbf{x}_i) - \sum_{j \neq i} a_j k(\cdot, \mathbf{x}_j) \right\|^2 \quad (6)$$

From Equation (6), the kernel, to be pruned, is $i^* = \min_i \delta_i$.

The i^* -th kernel is to be replaced with a new kernel $k(\cdot, \mathbf{x}_t)$. Before the replacement, the i^* -th kernel is projected to the space spanned by the remaining $B - 1$ kernels. Here, f_t is given as follows:

$$f_t = f_{t-1-i^*} + \tau_{replace}^{opt} w_i P_{t-1-i^*} k(\cdot, \mathbf{x}_i^*) + y_t k(\cdot, \mathbf{x}_t) \quad (7)$$

where $P_{t-1-i^*} k(\cdot, \mathbf{x}_i^*)$ denotes the projected vector to the space spanned by the remaining kernels. $\tau_{replace}$ denotes the projection ratio, which is determined to minimize the loss caused by the replacement. However, if W_{prj} repeats this process, it will forget the previously learned results because the projection process modifies the parameter w_j . To prevent, this W_{prj} chooses one of two learning options by estimating and comparing

the loss caused by forgetting and the residual error of W_{prj} for each option. The two losses must be balanced.

The two options are as follows:

- Substitution: This is the above-mentioned algorithm.
- Modification: Modify existing kernel parameters w_i and R_i for $i = 1, 2, \dots, B$ to bring its output value as close to e_t as possible.

$$f_t = f_{t-1} + \tau_{\text{modify}}^{\text{opt}} e_t P_{t-1} k(\cdot, \mathbf{x}_t) \quad (8)$$

Evaluation functions for the learning options W_{prj} selects the learning option for which it calculates the lowest loss function value, represented as $L_{\text{modify}}(\tau, I_w)$ and, $L_{\text{replace}}(\tau)$. For example, if $L_{\text{replace}}(\tau_*)$, where $\tau_* = \arg \min_{\tau} L_{\text{replace}}(\tau)$, had the lowest value between the two options, W_{prj} would select the substitution option.

The loss functions consist of losses caused by the forgetting and residual errors for the new sample.

- **Modification:** The expected loss with this option is the sum of the loss due to projection and pruning. Hence,

$$\begin{aligned} L_{\text{modify}}(\tau, I_{w \text{ new}}) &\equiv I_{w \text{ new}} e_t^2 \|\tau q_1(\cdot) - k(\mathbf{x}_t, \cdot)\|^2 + \tau^2 e_t^2 \|q_1(\cdot)\|_w^2, \\ \tau_{\text{modify}}^{\text{opt}} &= \max\{\arg \min_{\tau} L_{\text{modify}}(\tau, I_{w \text{ new}}), 0\} \\ &= \max\left\{I_{w \text{ new}} \frac{\langle q_1(\cdot), k(\mathbf{x}_t, \cdot) \rangle}{\|q_1(\cdot)\|_w^2 + I_{w_i} \|q_1(\cdot)\|^2}, 0\right\} \end{aligned} \quad (9)$$

where $q_1(\cdot) \equiv \sum_i a_i k(\cdot, \mathbf{x}_i)$ and $\|q_1(\cdot)\|_w^2 = \|\sum_i a_i \sqrt{N_i} k(\cdot, \mathbf{x}_i)\|^2$, where N_i denotes the number of learned samples for the i -th kernel. N_i is set during the learning. The detailed algorithm for setting up the parameters is written in [6].

- **Substitution:**

$$\begin{aligned} L_{\text{replace}}(\tau) &\equiv N_{t-1} \|k(\mathbf{x}_i, \cdot) - \tau q_2(\cdot)\|^2 + \tau^2 W_i^2 \|q_2(\cdot)\|_w^2 \\ \tau_{\text{replace}}^{\text{opt}} &= \max\{\arg \min_{\tau} L_{\text{replace}}(\tau), 0\} \\ &= \max\left\{\frac{I_{w_i} \langle q_2(\cdot), k(\mathbf{x}_i, \cdot) \rangle}{\|q_2(\cdot)\|_w^2 + I_{w_i} \|q_2(\cdot)\|^2}, 0\right\} \end{aligned} \quad (10)$$

where $q_2(\cdot) \equiv \sum_{j \neq i^*} a_j k(\cdot, \mathbf{x}_j)$ and $\|q_2(\cdot)\|_w^2 \equiv \|\sum_{j \neq i^*} a_j \sqrt{N_j} k(\cdot, \mathbf{x}_j)\|^2$.

3.3 Optimization of hyperparameters

The hyperparameters of the kernels must be optimized to maximize accuracy. Normally, such hyper parameter optimization is performed gradually throughout the learning process. To adapt the hyperparameter optimization methods to one-pass learning, the method is executed without referencing past samples.

Kernel parameter optimization and speeding up Gaussian kernel variance is another important parameter that affects the generalization capability of W_{prj} . Given that the optimal variance is not known in advance, the W_{prj} must optimize the value based on its learned results. Yamauchi [16] proposed a method that optimizes σ using existing kernel centroids. However, that model does not address how to determine the timing of optimization of σ . Irfan et. al. [17] proposed a method to optimize σ using stored samples. This study applies an improved version of [16] that determines the timing of the optimization of σ . The suitability of variance is estimated using the leave-one-out method on existing kernels. Thus, W_{prj} calculates the output to the i -th kernel center using the following equation:

$$y_{t-i}(\mathbf{x}_i) = \langle f_{t-i}, K(\cdot, \mathbf{x}_i) \rangle, \tag{11}$$

where f_{t-i} mean f and in which the i -th kernel is removed. The error of $y_{t-i}(\mathbf{x}_i)$ is then estimated as follows:

$$e_i(\sigma) \equiv \{y_t(\mathbf{x}_i) - y_{t-i}(\mathbf{x}_i)\}^2. \tag{12}$$

Note that σ is used in f_{t-i} and g_{t-i} . $e_i(\sigma)$ for all kernels is averaged.

$$\hat{e}(\sigma) = \sum_i e_i(\sigma) / B \tag{13}$$

Algorithm 1 σ optimization

```

1: for  $\sigma = 0.001$  to  $2.0$ ; step =  $0.001$  do
2:    $Error_t \leftarrow \hat{e}(\sigma) : Eq13$ 
3:   if  $Error_t \leq Error_{t-1}$  then
4:      $Error_{t-1} \leftarrow Error_t$ 
5:   else if  $Error_t > Error_{t-1}$  then
6:      $BREAK$ 
7:   end if
8: end for

```

$\hat{e}(\sigma)$ is estimated for σ values, and σ^* that minimizing $\hat{e}(\sigma)$ is determined. In the experiment, σ was varied in the interval $[0.001\bar{\sigma}, 2.0\bar{\sigma}]$ where $\bar{\sigma}$ is a predetermined centroid value. The optimization process described above should be executed when part of a kernel center is changed (e.g., after exercising the replacement option). Furthermore, the kernel centers must also be uniformly distributed.

In this experiment, the optimization process was only run once, when the number of substitution with modification options exceeded a specified threshold. This threshold is when either the substitution or modification option is used 100 times. Note that optimizing σ must occur first. The error curve in this method takes the form shown in Figure 2.

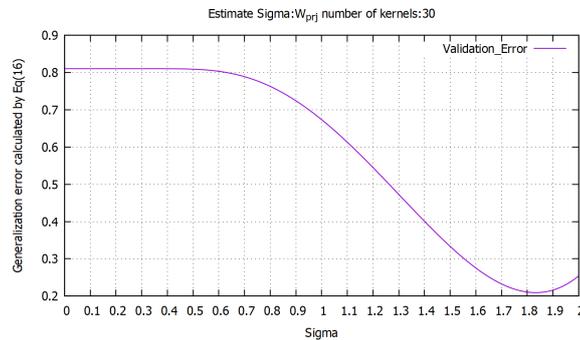


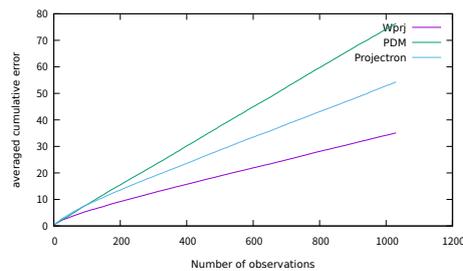
Fig. 2: Generalization error curve

The optimum value of σ in Figure 2 is 1.825. The generalization error continues decreasing monotonically to the optimum value, and the error monotonically increases when it exceeds the optimum value. Therefore, it is possible to eliminate the calculation of generalization error at the time when the error increases (See Algorithm 1).

4 Experiments, and discussion

4.1 Comparison of the weighted projection method and the others

Before examining the capabilities of a proposed method, we conducted the comparison between the weighted projection method and the other similar kernel perceptron learning methods [6]. We compared the projectron [13], and the perceptron with dynamic memory (PDM) [14] methods. Although PDM and projectron are designed to solve clustering problems, we applied these models to regression. To make a fair comparison, the label of each instance was replaced with the corresponding residual error of the kernel perceptron, similar to the weighted projection method.

Fig. 3: Cumulative errors for the concrete dataset (memory constraint : $B = 10$)

In the experiments, data from the servo, housing, cpu-performance, concrete and hearta1 datasets, were collected from the UCI machine learning repository ¹ and were

¹ <https://archive.isc.uci.edu/ml/index.php>

each input once to the system. Figure 3 shows the cumulative errors of the three methods for the concrete dataset as an example with a fixed memory constraint of $B = 10$. The cumulative error at the t -th round is

$$E_{cum}(t) = E_{cum}(t-1) + (y_t - \langle f_{t-1}, k(\cdot, \mathbf{x}_t) \rangle)^2 \quad (14)$$

Therefore, the squared residual error of f_{t-1} , which is the kernel perceptron before the learning of (\mathbf{x}_t, y_t) , is accumulated as E_{cum} . The experiments were repeated 50 times. The order of presenting data was changed and the results were averaged over the 50 trials. We can see from Figure 3 that the cumulative error of the proposed method (W_{prj} method) was the smallest.

Table 1: Averaged cumulative error for each dataset after learning (memory constraint $B = 10$)

dataset	W_{prj}	PDM	Projectron
servo	13.44 ± 0.2	13.67 ± 0.1	13.6 ± 0.1
housing	30.82 ± 1.1	48.7 ± 1.1	38.4 ± 1.6
cpu performance	2.81 ± 0.05	4.4 ± 0.1	3.2 ± 0.1
concrete	35.1 ± 0.9	76.4 ± 4.2	54.3 ± 2.3
heartal	119.7 ± 0.4	120.2 ± 0.3	119.6 ± 0.6

Table 1 lists the performances of the compared algorithms for the aforementioned datasets. We can see that the weighted projection methods showed the smallest cumulative error except for heartal.

4.2 Comparison of the used echo state network method with the original echo state network

We compared the proposed method, and the original ESN model. In the experiment, we used a data set with 1500 points in the first half as the Mackey glass time series and 1500 points as the Henon map (Figure 4).

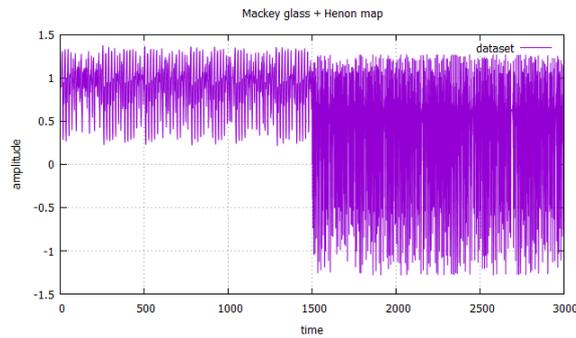


Fig. 4: Learning data set

After the 3000 points were presented, we presented the Mackey glass time series again without parameter modification, and measured the error, and confirmed whether additional learning could be performed without forgetting.

$RMSE = \sqrt{1/N \sum_{i=1}^N (y - \hat{y})^2}$ was used for error measurement. Here, N is the number of evaluation samples, y and \hat{y} are teacher signals and predicted values. The experiment was repeated 50 times. The reservoir connection rate and the spectral radius were 0.2 and 0.9, respectively. The following Figures 5a, 5b, 5c and 5d, show the RMSE's of various models, whose reservoir sizes and unit sizes were varied. Experiments also verified the effectiveness of using the σ optimization. Furthermore, in this experiment we assumed that the original ESN model modifies the output weight by using the back propagation with a learning rate of 0.1.

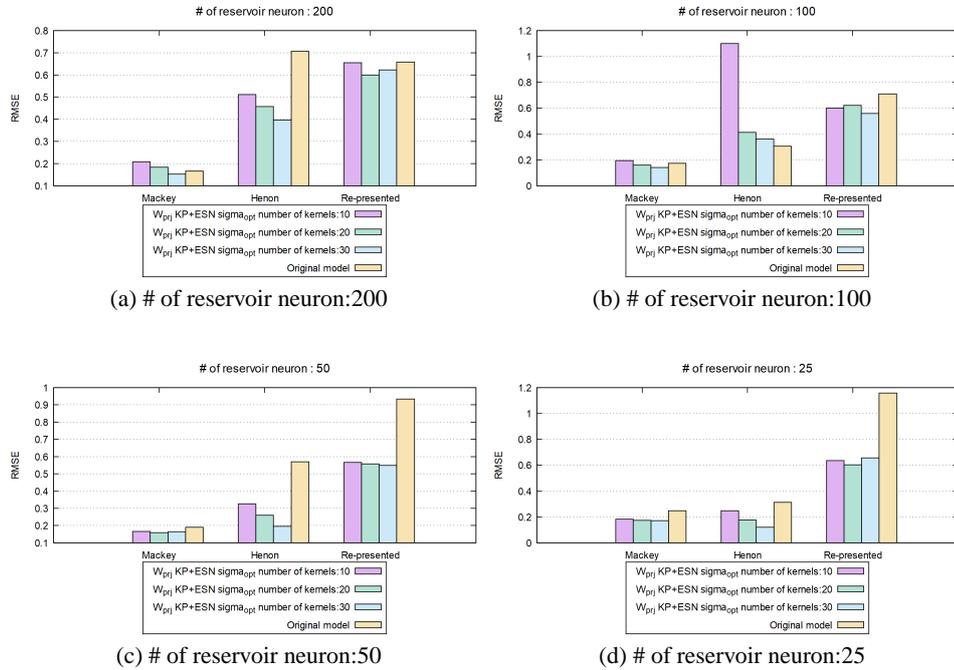


Fig. 5: RMSE with 50 times average histogram

From figures 5(a) and 5(b), we can see that, in the cases where the reservoir sizes were 200 and 100, the difference between original ESN and the proposed W_{prj} model's RMSE's were small. However, from figures 5(c) and 5(d), we can see that the performances of the original ESN are worse than those of the W_{prj} model. In particular, the RMSE of the original ESN when the Mackey glass data was presented again becomes large when the reservoir size is small. On the other hand, those of the proposed W_{prj} model are smaller. This means that if the number of dimensions of output is small, each output vector is not linearly dependent on the other output vectors. Even under such sit-

uation, the proposed model can support the learning because it converts the input vector to a high dimensional (infinite) vector.

5 Conclusion

In this paper, a one-pass learning scheme for time-series prediction was proposed. Preventing the time series data learned in the first half from being forgotten is an important task. To this end, the proposed learning scheme consists of the reservoir and applies the weighted projection method, which has been proposed in our previous study. Given that the reservoir has fixed connections, the one-pass learning ability depends only on the ability of the learner that receives inputs from the reservoir. According the results from our previous study [6], the incremental learning ability of the weighted projection method is stronger than the other similar online learning method for a fixed memory constraint. By using the kernel machine as the learner, we were able to reduce the size of the reservoir. The simulation results suggest that our proposed method is effective when the reservoir size is small. To realize accurate prediction, hyper parameter optimization is also required. To realize the optimization through the one-pass learning, a hyper parameter optimization method that uses the kernel centroids was developed. Experimental results suggest that the proposed method can learn time-series data using the one-pass method through write-once.

References

1. K. Yamauchi, N. Yamaguchi, and N. Ishii, "Incremental learning methods with retrieving interfered patterns," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 10, no. 6, pp. 1351–1365, November 1999.
2. R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 31, no. 4, pp. 497–508, November 2001.
3. S. Ozawa and K. Okamoto, "An incremental learning algorithm for resource allocating networks based on local linear regression," in *Neural Information Processing 16th International Conference on Neural Information Processing Bangkok, Thailand, December 1-5, 2009, Part I*, vol. LNCS5863, December 2009, pp. 562–569.
4. D. L. Wang and B. Yuwono, "Incremental learning of complex temporal patterns," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 7, no. 6, pp. 1465–1481, November 1996.
5. K. Yamauchi and M. Sato, "Incremental learning of spatio-temporal patterns with model selection," in *ICANN2007 International Conference on Artificial Neural Networks*, vol. LNCS4668. Springer-Verlag, September 2007, pp. 149–158.
6. K. Yamauchi, "An importance weighted projection method for incremental learning under unstationary environments," in *IJCNN2013: The International Joint Conference on Neural Networks 2013*. The Institute of Electrical and Electronics Engineers, Inc. New York, New York, August 2013, pp. 1–9.
7. S. Ruber, "Transfer learning-machine learning 's next frontier," <http://ruder.io/transfer-learning/>.
8. J. J. Steil, "Online reservoir adaptation by intrinsic plasticity for backpropagation?decorrelation and echo state learning," *Neural Networks*, vol. 20, pp. 353–364, 2007.

9. X. Deng and X. Wang, "Incremental learning of dynamic fuzzy neural networks for accurate system modeling," *FUZZY sets and systems*, vol. 160, pp. 972–987, 2009.
10. G. Feng, G.-B. Huang, Q. Lin, and R. Gay, "Error minimized extreme learning machine with growth of hidden nodes and incremental learning," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 20, no. 8, pp. 1352–1357, August 2009.
11. H. Soh and Y. Demiris, "Spatio-temporal learning with the online finite and infinite echo-state gaussian processes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 522–536, March 2015.
12. O. Dekel, S. Shalev-Shwartz, and Y. Singer, "The forgetron: A kernel-based perceptron on a fixed budget," <http://www.pascal-network.org/>, Tech. Rep., 2005.
13. F. Orabona, J. Keshet, and B. Caputo, "The projectron: a bounded kernel-based perceptron," in *ICML2008*, 2008, pp. 720–727.
14. W. He and S. Wu, "A kernel-based perceptron with dynamic memory," *Neural Networks*, vol. 25, pp. 105–113, 2012.
15. W. He and J. T. Kwok, "Simple randomized algorithms for online learning with kernels," *Neural Networks*, vol. 60, pp. 17–24, 2015.
16. K. Yamauchi, "Incremental learning on a budget and its application to quick maximum power point tracking of photovoltaic systems," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 18, no. 4, pp. 682–696, 2014.
17. M. Irfan, A. Koj, H. R. Thomas, and M. Sedighi, "Geographical general regression neural network (ggrrn) tool for geographically weighted regression analysis," in *GEOProcessing 2016 : The Eighth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 2016, pp. 154–159.

Nonlinear relationship detection using pseudocorrelation

Jozef Jakubík*

Slovak Academy of Sciences
Institute of Measurement Science
Dúbravská cesta 9
841 04 Bratislava 4
Slovakia
jozef.jakubik.jefo@gmail.com

Abstract. Quantifying relationships between time series is a common problem in many research fields, including e.g. finance or meteorology. There are many metrics (the Pearson correlation coefficient, Granger causality, ...) which capture some data relationships. Many of these, however, only consider linear relationships. We introduce a new measure of relationship strength between time series that captures a different aspect compared to previously proposed metrics. The ‘pseudocorrelation’ measures the degree to which a time series is a function of other time series.

Keywords: correlation, time series, neural networks

1 Introduction

Metrics such as the Pearson correlation coefficient, rank correlation, Granger causality [1], mutual information, joint entropy [2], etc., are popular for studying relationships between time series. Some of them are capable of capturing only specific kinds of relationships. For example, the Pearson correlation coefficient only considers linear relationships. Other metrics can inspect only unconditional probabilistic dependences, including e.g. causality detection methods in state spaces (Convergence cross mapping [3], Predictability improvement [4]). In this short paper, we introduce ‘pseudocorrelation’ which allows us to capture nonlinear relationships between one and more time series.

2 Pseudocorrelation

We denote a time series by $X = X(t) = [x(1), x(2), \dots, x(t-1), x(t)]$. Time series $X(t - \theta)$ shifted relative to $X(t)$ by θ time steps.

* The work was supported by the Scientific Research and Development Agency, grant APVV-15-0295, and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, grants VEGA 2/0047/15 and VEGA 2/0011/16.

Definition 1 (Pseudocorrelation). Consider two time series $X(t)$, $Y(t)$ and k time shifts $\delta_1, \dots, \delta_k$ in increasing order (δ_i can be positive or negative). Time series $Z(t)$ holds $z(t) = f(y(t - \delta_1), \dots, y(t - \delta_k))$. We use the notation $Z(t) = f([Y(t - \delta_1), \dots, Y(t - \delta_k)])$. We define the pseudocorrelation between $X(t)$ and $[Y(t - \delta_1), \dots, Y(t - \delta_k)]$ as:

$$\max_{f: \mathbb{R}^n \rightarrow \mathbb{R}} \rho_{X(t), Z(t)=f([Y(t-\delta_1), \dots, Y(t-\delta_k)])} \quad (1)$$

where $\rho_{X(t), Z(t)=f([Y(t-\delta_1), \dots, Y(t-\delta_k)])}$ is the Pearson correlation coefficient.

Next we describe some properties of the pseudocorrelation.

Observations 1 If the value of the pseudocorrelation is

- 0, there is no function $f()$ which can map $[Y(t - \delta_1), \dots, Y(t - \delta_k)]$ onto $X(t)$.
We say there is no relationship between $[Y(t - \delta_1), \dots, Y(t - \delta_k)]$ and $X(t)$.
- 1, it is possible to express $X(t)$ as a function of $[Y(t - \delta_1), \dots, Y(t - \delta_k)]$.

Estimation of pseudocorrelation. Given a candidate function f that is approximately optimal for equation 1, we propose to estimate the pseudocorrelation by the sample correlation,

$$\hat{\rho}_{X(t), Z(t)=f([Y(t-\delta_1), \dots, Y(t-\delta_k)])} = \frac{\sum_{i=\delta_1+1}^{T-\delta_k} (x(i) - \bar{X})(z(i) - \bar{Z})}{\sqrt{\sum_{i=\delta_1+1}^{T-\delta_k} (x(i) - \bar{X})^2} \sqrt{\sum_{i=\delta_1+1}^{T-\delta_k} (z(i) - \bar{Z})^2}} \quad (2)$$

where T is the length of the time series X and Y . We propose to split the available data into two sets, find a candidate function f^* by optimizing equation 2 in $f \in \mathcal{F}$ on the training set, and evaluating equation 2 for $f = f^*$ on the test set. The obtained value is necessarily a lower bound of the true value, due to constraining the optimization to a function class \mathcal{F} and due to only finding the training set optimizer rather than the population minimizer.

3 Data and Experimental setup

For illustration, we use a Hénon map with couplings C_x, C_y ; see equation 3. Table 1 shows pseudocorrelation values for different values of C_y . If $C_x = C_y = 0$, it is impossible to express $y(t)$ based on $X(t)$ and vice versa. If the constant is positive, $C_x, C_y > 0$, then it is possible to express $x(t)$, $y(t)$ (based on $[y(t + 1), y(t), y(t - 1)]$, $[x(t + 1), x(t), x(t - 1)]$ respectively) from equation 3. From definition 1 $k = 3$ and time shifts are $\{-1, 0, 1\}$.

We inspect pseudocorrelation for changing causality couplings C_x and C_y in the Hénon map:

$$\begin{aligned} x(t+1) &= 1.4 - (C_x x(t)y(t) + (1 - C_x)x^2(t)) + 0.3x(t-1) \\ y(t+1) &= 1.4 - (C_y x(t)y(t) + (1 - C_y)y^2(t)) + 0.3y(t-1) \end{aligned} \quad (3)$$

with the starting point $[0.7, 0, 0.91, 0.7]$.

We use the estimation procedure from section 2. We choose the function class \mathcal{F} to be fully-connected two-hidden-layer neural networks with 10 nodes on every hidden layer and a sigmoid activation function everywhere except the output neuron. We optimize the network for mean square error. The data is split 3 : 1 into the training and testing set. To reduce variance, we trained several neural networks and take the median of with training data and as estimate of pseudocorrelation we consider a median of pseudocorrelations on testing data.

4 Results

Bidirectional coupled Hénon map with $C_x = C_y = 0.1$. Generated data $X(t)$ and $Y(t)$ from equation 3 are uncorrelated $\rho_{X,Y} = 0.0714$, but as mentioned earlier, it is possible to express $x(t), y(t)$ based on $[y(t+1), y(t), y(t-1)], [x(t+1), x(t), x(t-1)]$ respectively. In contrast to the Pearson correlation coefficient, both pseudocorrelation coefficients, $PC_{Y(t),[X(t+1),X(t),X(t-1)]}, PC_{X(t),[Y(t+1),Y(t),Y(t-1)]}$, are close to one (0.9991 and 0.9984).

Unidirectional Hénon map with $C_x = 0$. In figure 1 we can see visual changes in the mapping of time series with changing coupling C_y and table 1 compares correlation and pseudocorrelation on training data. As we can see, pseudocorrelation $PC_{X(t),[Y(t+1),Y(t),Y(t-1)]}$ is zero if $C_y = 0$ and almost one if $C_y \neq 0$. Pseudocorrelation $PC_{Y(t),[X(t+1),X(t),X(t-1)]}$ is small almost for all C_y , except for the last three values, in which a phenomenon known as synchronization occurs, when the time series $Y(t)$ becomes like the time series $X(t)$.

C_y	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
$\rho_{X,Y}$	0.0114	0.0187	0.0057	0.0040	0.0622	0.1705	0.4512	0.9987
$PC_{X(t),[Y(t+1),Y(t),Y(t-1)]}$	0.0175	0.9973	0.9959	0.9988	0.9970	0.9989	0.9994	1
$PC_{Y(t),[X(t+1),X(t),X(t-1)]}$	0.0007	0.0088	0.0809	0.1595	0.1791	0.2715	0.5076	1

Table 1. Comparison of correlation and pseudocorrelation (PC) on a Hénon map for changing the coupling C_y with fixed coupling $C_x = 0$.

5 Conclusion

We introduce pseudocorrelation which allows capturing nonlinear relationships between time series. Also, it is allows us to study the relationship between one time series and a function of more time series. We are working on studying pseudocorrelation properties.

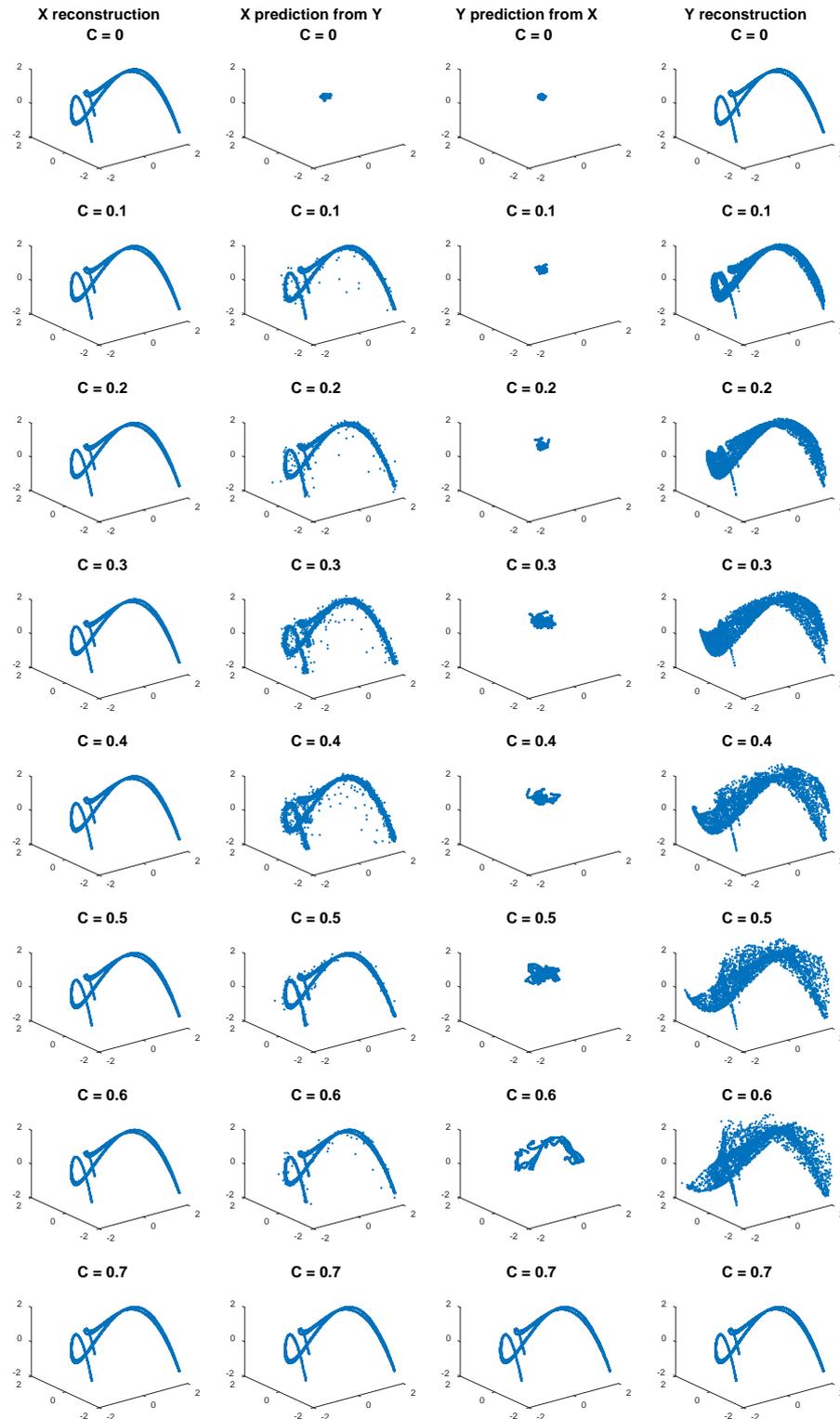


Fig. 1. Visualisation of reconstructed attractors. X, Y reconstruction is $[x(t), x(t-1), x(t-2)], [y(t), y(t-1), y(t-2)]$ respectively. X prediction from Y is $[\hat{x}(t), \hat{x}(t-1), \hat{x}(t-2)]$ where $\hat{x}(t)$ is estimation of $x(t)$ based on $[y(t+1), y(t), y(t-1)]$. Y prediction from X similarly.

References

1. Granger, Clive WJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424-438 (1969).
2. Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, (2012).
3. Sugihara, George, et al. Detecting causality in complex ecosystems. *science*, 1227079 (2012).
4. Krakovská, Anna, and Filip Hanzely. Testing for causality in reconstructed state spaces by an optimized mixed prediction method. *Physical Review E* 94.5, 052203 (2016).

Automatic detection of sleep disorders: Multi-class automatic classification algorithms based on Support Vector Machines

David López-García, María Ruz, Javier Ramírez Pérez de Inestrosa, and Juan Manuel Górriz Sáez

Signal Theory, Telematics and Communications Department (TSTC),
University of Granada, Spain
{dlopez,mruz,javierrp,gorriz}@ugr.es
<https://tstc.ugr.es>

Abstract. Currently, sleep disorders are a common problem within society. Sleep specialists diagnose these disorders by visual inspection of several biomedical signals, such as the electroencephalogram, obtained during the patient's sleep. This procedure is known as polysomnography. The classification, based on visual inspection, can be a tedious and time-consuming task. Consequently, automatic classification methods are constantly increasing in popularity. This study presents an effective method for multi-class classification of sleep stages and sleep apnea episodes using the MIT-BIH Polysomnographic Database. The analysis involved the use of a set of supervised learning algorithms known as Support Vector Machines (SVM). This learning technique provides a theoretically elegant, computationally efficient, and very effective solution for many practical pattern recognition problems. The results of this study showed that more than 80% of the test segments are successfully classified in a nine-class scenario.

Keywords: electroencephalography, sleep apnea, sleep stages, automatic classification, support vector machines, principal component analysis, partial least squares.

1 Introduction

Currently, sleep related breathing disorders, such as obstructive apnea, are a common problem within society. These diseases could lead to falling asleep while working or having severe headaches that are difficult to treat. In more severe cases, other complications may occur, such as hypertension, swelling of the legs or depression. In order to detect sleep disorders, sleep specialists perform a sleep study on patients, which is known as polysomnography. During the patient's sleep, neurophysiological and cardiorespiratory variables are collected continuously and simultaneously. Specifically, some of the following parameters are recorded and studied: electroencephalic activity (EEG), eye movements (EOG), face and leg muscle movements (EMG), air flow through the mouth and nose

(e.g. pneumotachography, thermistor), oxygen saturation (pulse oximetry), cardiocirculatory parameters such as heart rate or blood pressure and heart function (ECG).

After the recording stage, sleep specialists diagnose these disorders by visual inspection of all these biomedical signals, according to Rechtschaffen and Kales [1] scoring standard. This classification, based on visual inspection, can be a tedious and time-consuming task. Consequently, automatic sleep [2][3][4] and apnea [5][6] classification methods are constantly increasing in popularity. This article proposes an automatic classification system for sleep apnea episodes and sleep stages using different patient EEG feature extraction techniques and machine learning algorithms.

2 Materials

This study presents an effective method for multi-class classification of sleep stages and sleep apnea episodes. We used the MIT-BIH Polysomnographic Database[7][8], registered by the Boston's Beth Israel Hospital Sleep Laboratory to evaluate the chronic Obstructive Sleep Apnea (OSA) syndrome. The database consists of more than 80 hours of all-night recordings of multiple physiological signals and annotations of the sleep stages and apnea episodes registered during the patients' sleep (16 male subjects, 42 ± 7 years old, 118.85 ± 23.3 kg).¹

The database provides three types of files for each patient. The header file contains information regarding the patient's age, weight, and signal registration parameters (length of the recording, sampling rate, calibration constants, etc). The data file contains the patient's physiological signals in different channels, including the electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), electromyogram on the chin (EMG), the invasive blood pressure measured in the radial artery (BP), the respiratory signal obtained by means of a nasal thermistor (RESP), the respiratory effort signal derived from a respiratory inductance plethysmography (PLETH), the oxygen saturation measured in the ear lobe (SO₂), and the systolic volume signal (SV). Moreover, each data file has two annotation files associated with it, one for the heartbeat labeling and the other for the diagnosed sleep stages and sleep apnea episodes. Each label consists of several codes indicating the state of the patient in the following 30-second segment and any event occurred during this period. Table 1 shows the meaning of each annotation code.

Additionally, PhysioBank data manipulation requires specialized software. In order to read and convert the Physionet data files into a Matlab-friendly for-

¹ We excluded both spl41 and spl45 patients. For these patients there are no apnea annotations. Therefore, we will only find annotations related to sleep stages in the annotation file. Patients who are not fully labeled may induce error in the results.

Code	Label meaning	Code	Label meaning
W	Subject is awake	MT	Movement time
1	Sleep stage 1	H	Hypopnea
2	Sleep stage 2	HA	Hypopnea with arousal
3	Sleep stage 3	OA	Obstructive apnea
4	Sleep stage 4	X	Obstructive apnea with arousal
R	REM sleep	CA	Central apnea
L	Leg movements	CAA	Central apnea with arousal
LA	Leg movements with arousal	A	Unspecified arousal

Table 1. Annotation code and meaning

mat file we used the WFDB software package[9]. This is a set of subroutines for reading and writing files in the formats used by PhysioBank databases.

This study only analyzed single-channel EEG (sampled at 250Hz), recorded using two different dipolar configurations, C4/A1, O2/A1, and C3/O1 according to the International 10–20 system electrode placement[10], and depending on the patient.

3 Methods

3.1 Segmentation

The first part of this study consisted of reading, segmenting, labeling, and storing the EEG data of each patient. The database provides a 30-second trial segmentation following the labeling code showed in table 1. All these types of events were collapsed in nine main classes: AWAKE (patient is awake), S1 (sleep stage 1), S2 (sleep stage 2), S3 (sleep stage 3), S4 (sleep stage 4), R (REM sleep), H (hypopnea and hypopnea with arousal), O (obstructive apnea and obstructive apnea with arousal), and C (central apnea and central apnea with arousal). The leg movement (L), leg movement with arousal (LA), movement time (MT) and unspecified arousal (A) segments were rejected and not included in our study.

3.2 Feature extraction

One of the most important parts when designing an automatic classification system is to determine the features vector used to train and test the system. We proposed tree techniques in order to extract features from the EEG time series. We estimated the power spectral density (PSD) using parametric, non-parametric, and other techniques. For the PSD non-parametric estimation we used the following methods: i) periodogram, ii) modified periodogram, iii) Bartlett, iv) Welch and v) Blackman-Tuckey's methods. For the PSD parametric estimation we used i) autoregressive (AR), ii) Burg, and iii) Yule-Walker's methods. Thomson, Multiple Signal Classification (MUSIC), and eigenvalues methods were also applied

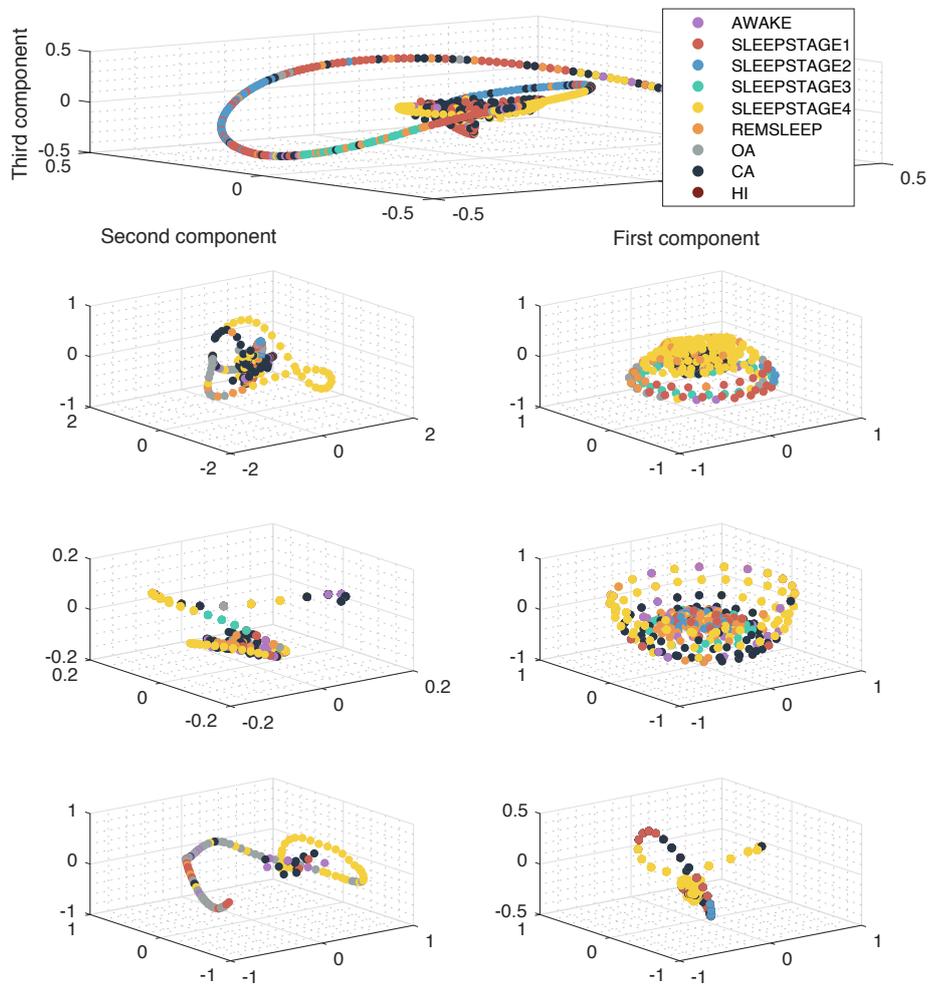


Fig. 1. 3D representations of the temporal evolution of the first three PCA scores. Each point corresponds to a 30-second segment of EEG for the patient a) SLP61, b) SLP59, c) SLP16, d) SLP60, e) SLP14, f) SLP67, g) SLP37.

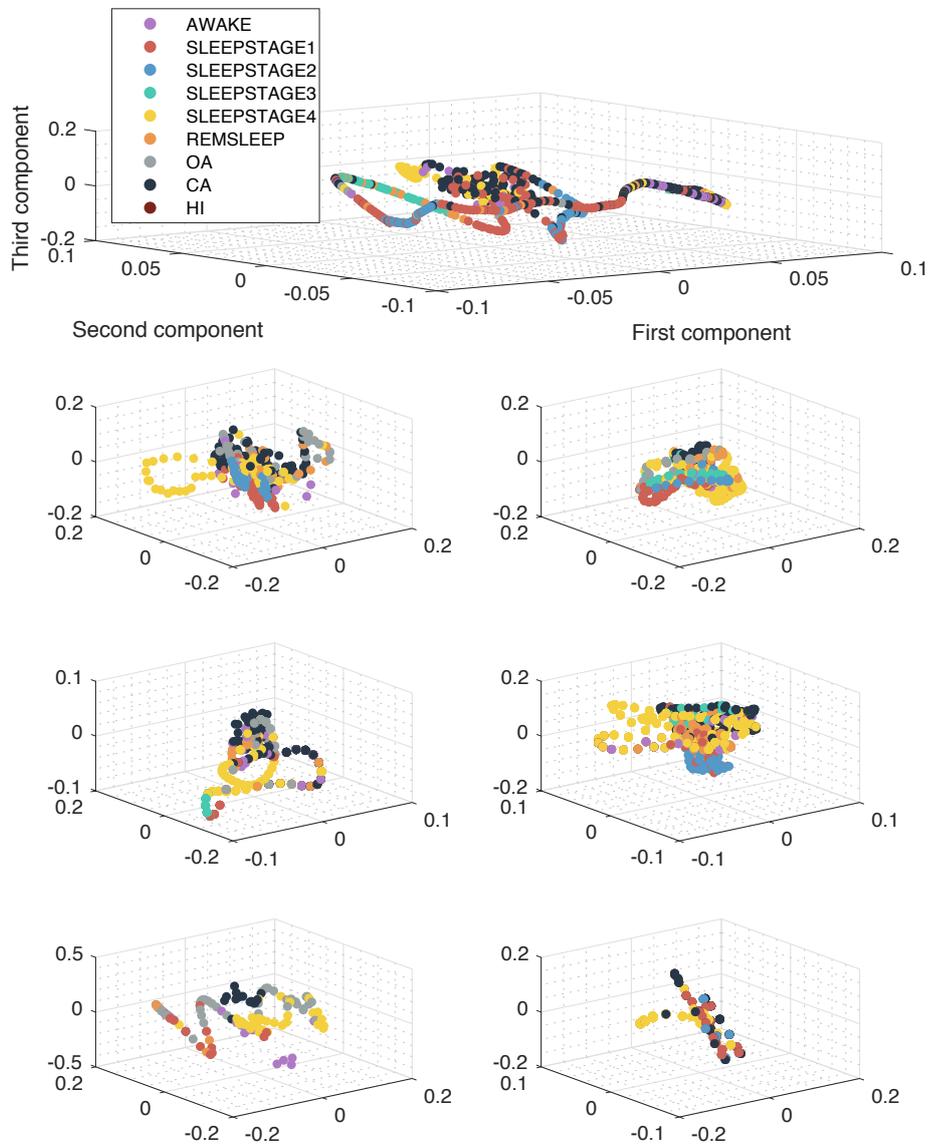


Fig. 2. 3D representations of the temporal evolution of the first three PLS scores. Each point corresponds to a 30-second segment of EEG for the patient a) SLP61, b) SLP59, c) SLP16, d) SLP60, e) SLP14, f) SLP67, g) SLP37.

to estimate the PSD. All of these techniques were computed in Matlab using the Signal Processing Toolbox and the default parameters for the PSD estimation.

In addition, we computed Principal Component Analysis (PCA) and Partial Least Squares (PLS) directly from the EEG time series, which decompose the data into two sets of variables named scores and loadings. The model structures of PLS and PCA are similar. The data are first transformed into a set of intermediate linear latent variables (components) and these new variables are taken into account. While PCA creates orthogonal weight vectors by maximizing the covariance between the variables, PLS not only considers the variance of the samples but also the class labels[11]. These techniques are frequently used to look for patterns in multidimensional datasets[12]. More specifically, both PCA and PLS have been applied in the design of several Computer-Aided Diagnosis (CAD) systems for early diagnosis of alzheimer's disease[13][14][15]. Here we characterized each 30 second EEG segment (the different sleep stages and apnea episodes) in a multidimensional PCA (fig.1) and PLS (fig.2) space, showing the temporal evolution of the sleep state of each patient. This 3D representations are possible due to the dimension reduction performed by PCA/PLS algorithms.

3.3 Classification

The classification block involved the use of a set of supervised learning algorithms known as Support Vector Machines (SVM) [Vapnik, 1982] for different kernel functions (linear, polynomial, and radial basis function) and different training-classification algorithms (one-versus-one and one-versus-all). This learning technique provides a theoretically elegant, computationally efficient, and very effective solution for many practical pattern recognition problems. Specifically, SVMs minimize the Vapnik Chervonenkis (VC) dimension, which is a robust solution in classification learning[16], minimizing the separation margin between the binary classes by constructing an HLT decision function $F(\alpha, \mathbf{x})$ whose norm is minimum[16]:

$$\|\alpha\|^2 + C \sum_{i=1}^l \xi_i \quad (1)$$

subject to:

$$\omega_i(\alpha \cdot \mathbf{x}_i) \geq 1 - \xi_i; \quad \xi_i \geq 0; \quad i = 1, \dots, l$$

where C is a constant which modulates the trade-off between the training error and the complexity of the model, ξ_i are slack variables, and the decision rule is defined as $F(\mathbf{x}, \alpha)$. The solution is computed using $\alpha = \sum_{i=1}^l a_i y_i \mathbf{x}_i$. The multipliers $0 \leq a_i \leq C$ were derived using the Sequential Minimal Optimization (SMO) algorithm[17] of a dual Lagrangian problem in equation 1.

Given that SVM algorithms solve binary problems, we needed slightly different approaches for solving the multi-class scenario. To do so, two algorithms

(one versus one and one versus all) were implemented. On the one hand, for the one versus all approach, nine binary classifiers were trained, with each individual class facing all the rest together. To solve the ambiguities we used the a posteriori probabilities for each individual segment. On the other hand, for the one versus one approach, we trained as many classifiers as the number of non-repeated combinations for all the individual classes (thirty-two binary classifiers for a nine-class scenario).

Finally, to ensure that the results are independent of the training and test data partition, cross-validation techniques (leave-one-out and k-fold) were used during the validation period.

4 Results and discussion

In order to obtain the optimal configuration, the classifier's performance was tested for each kernel function, each classification algorithm, and each features extraction technique. For both PCA and PLS techniques, we also studied the optimal number of dimensions (number of principal components) required to obtain the best classification performance. Furthermore, we also assessed which electrode position configuration provides a better classification performance.

Preliminary results showed that, for the optimal configuration, more than 80% of the test segments were successfully classified into a nine-class scenario. The classifier performed better when the signal was recorded at C4/A1 electrode position configuration, regardless of the features extraction technique.

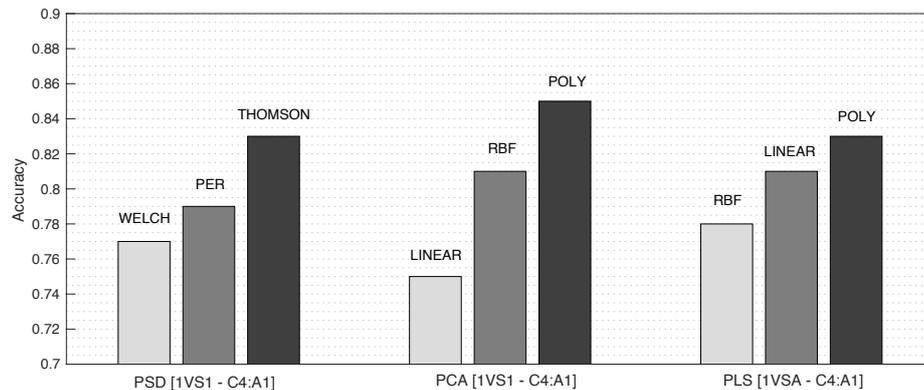


Fig. 3. Overall classification performance for the three feature extraction methods: a) PSD (for the best three estimation methods), b) PCA (for each SVM kernel function) and c) PLS (for each SVM kernel function).

However, different multi-class classification algorithms (one-vs-one and one-vs-all) did not show significant differences in the classification performance. Thomson and Welch's methods generally provide a better classification performance than other PSD estimation techniques (fig.5) regardless of the kernel function used for the classification. Finally, for PCA and PLS, we found that the polynomial and RBF kernel functions performed better than the linear one (fig.3), and the optimal number of components used for the features vector was between 10 and 20, with a 90-95% of the variance explained. See fig. 6 and 7 for more details.

The temporal evolution of PCA and PLS scores for each patient shown in figure 1 and 2 support these classification results. For each patient, different classes are clearly separable from each other using non-linear classifiers, which means that sleep stages and apnea episodes are well characterized in both PCA and PLS spaces. Additionally, figure 4 shows three PLS loadings vectors for the patient slp61 and three examples of different EEG classes. The components of these three loadings are shown in figure 2(a).

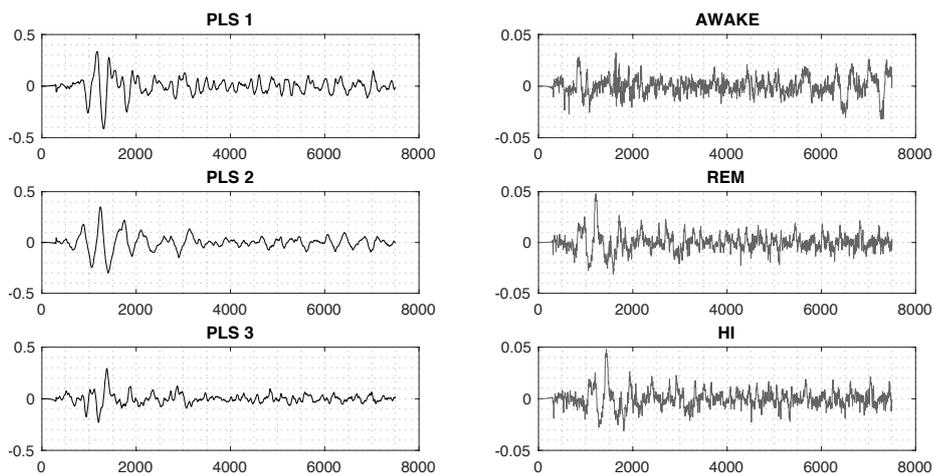


Fig. 4. Three PLS loadings vector representation for patient slp61 (left) and three examples of different EEG classes (right).

Table 2 shows three confusion matrices for different configurations and fifteen samples per class. These matrices represent the performance of the classifier for each class individually. Classes S1, S2 and S3 tend to be misclassified more frequently than the others, while S4, for example, is perfectly classified in the three cases.

Confusion Matrix (a)										Confusion Matrix (b)										Confusion Matrix (c)																																																																					
Overall accuracy: 83%																														Overall accuracy: 80%																														Overall accuracy: 78%																													
S1	S2	S3	S4	O	H	R	W	S1	S2	S3	S4	O	H	R	W	S1	S2	S3	S4	O	H	R	W																																																																		
S1	10	1	0	0	0	0	3	2	S1	9	0	0	0	0	0	0	2	S1	6	0	0	0	0	0	0	2																																																															
S2	0	13	3	0	0	3	0	0	S2	0	10	2	0	0	3	0	0	S2	0	12	1	0	0	3	0	0																																																															
S3	0	1	11	0	0	0	0	1	S3	0	4	13	0	0	0	0	2	S3	0	1	7	0	0	0	0	1																																																															
S4	0	0	0	15	0	0	0	0	S4	0	0	0	15	0	0	0	0	S4	0	0	0	15	0	0	0	0																																																															
O	0	0	0	0	15	0	0	0	O	0	0	0	0	13	0	0	0	O	0	1	5	0	15	0	0	0																																																															
H	0	0	0	0	0	12	0	0	H	0	1	0	0	0	12	0	0	H	0	1	0	0	0	12	0	0																																																															
R	4	0	0	0	0	0	12	0	R	3	0	0	0	0	0	15	0	R	6	0	0	0	0	0	15	0																																																															
W	1	0	1	0	0	0	0	12	W	3	0	0	0	2	0	0	11	W	3	1	2	0	2	0	0	12																																																															

Table 2. Confusion matrices for different configurations and fifteen samples per class.

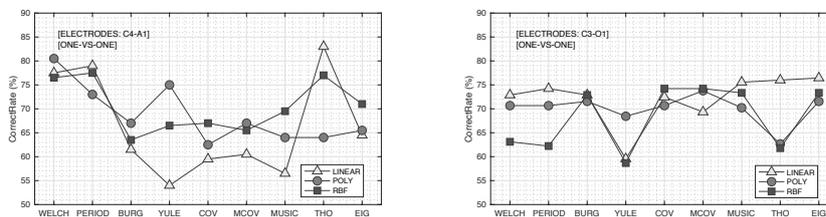


Fig. 5. Classifier performance for C4-A1 (left) and C3-O1 (right) electrodes and for each PSD estimation technique using the one-versus-one classification algorithm.

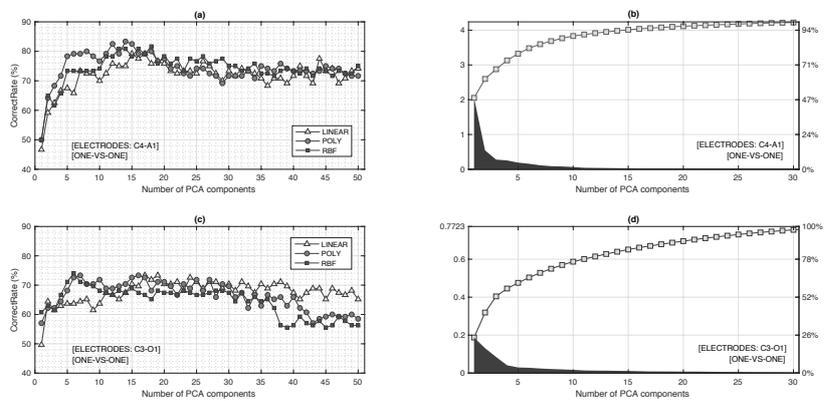


Fig. 6. Classifier performance for (a) C4-A1 and (c) C3-O1 electrodes based on the number of PCA scores using the one-versus-one classification algorithm. Each line represents the kernel function used. Figures (b) and (d) show the percentage of variance explained.

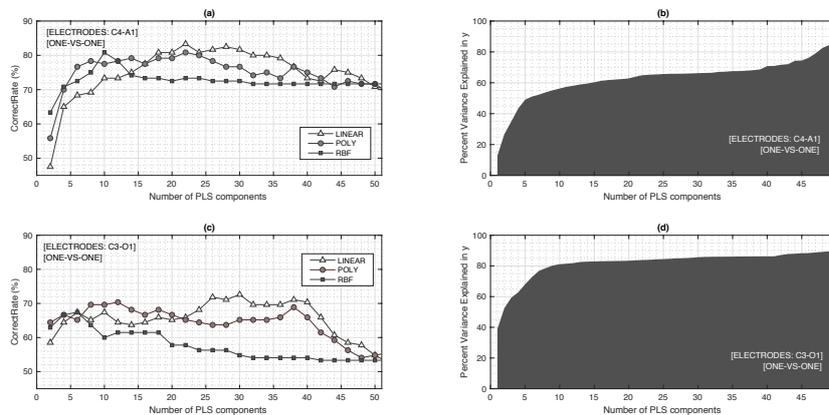


Fig. 7. Classifier performance for (a) C4-A1 and (c) C3-O1 electrodes based on the number of PLS scores using the one-versus-one classification algorithm. Each line represents the kernel function used. Figures (b) and (d) show the percentage of variance explained.

5 Conclusions

This study has designed and tested the performance of an automatic classification system for sleep apnea episodes and sleep stages, yielding an 85% accuracy for the optimal configuration in a nine-class scenario. The PSD estimation, PCA and PLS scores were used as features vectors for the system. Moreover, several configurations of binary SVMs were tested in combination with both one-versus-one and one-versus-all multi-class classification algorithms. Nevertheless, it would be of interest to improve the system using not only SVM but other different classification approaches.

Acknowledgments. This research was supported by the Spanish Ministry of Economy and Business under the TEC2015-64718-R and PSI2016-78236-P projects. The first author of this work is supported by a grant from the Spanish Ministry of Economy and Business (BES-2017-079769).

References

1. Anthony Kales and Allan Rechtschaffen. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. US Department of Health, Education and Welfare, Public Health Service, National Institutes of Health, National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968.
2. Khalid AI Aboalayon, Helen T Ocbagabir, and Miad Faezipour. Efficient sleep stage classification based on eeg signals. In *Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island*, pages 1–6. IEEE, 2014.

3. Yili Li, Kon Max Wong, et al. Eeg signal classification based on a riemannian distance measure. In *Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, pages 268–273. IEEE, 2009.
4. Cuneyt Yucelbas, Seral Ozsen, Salih Gunes, and Sebnem Yosunkaya. Effect of some power spectral density estimation methods on automatic sleep stage scoring using artificial neural networks. *IADIS International Journal on Computer Science & Information Systems*, 8(2), 2013.
5. Robert Lin, Ren-Guey Lee, Chwan-Lu Tseng, Heng-Kuan Zhou, Chih-Feng Chao, and Joe-Air Jiang. A new approach for identifying sleep apnea syndrome using wavelet transform and neural networks. *Biomedical Engineering: Applications, Basis and Communications*, 18(03):138–143, 2006.
6. Derong Liu, Zhongyu Pang, and Stephen R Lloyd. A neural network method for detection of obstructive sleep apnea and narcolepsy based on pupil size and eeg. *IEEE Transactions on Neural Networks*, 19(2):308–318, 2008.
7. Yuhei Ichimaru and GB Moody. Development of the polysomnographic database on cd-rom. *Psychiatry and clinical neurosciences*, 53(2):175–177, 1999.
8. M Je, M Gb, and P Ck. Goldberger al, amaral lan, glass l, hausdorff jm, ivanov pch, mark rg, mietus je, moody gb, peng ck, stanley he. physiobank, physiokit, and physionet: Components of a new research resource for complex physiologic signals. *vol*, 101:220.
9. Ikaro Silva and George B Moody. An open-source toolbox for analysing and processing physionet databases in matlab and octave. *Journal of open research software*, 2(1), 2014.
10. Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382, 1987.
11. Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.
12. Juan Manuel Górriz, Javier Ramirez, John Suckling, Ignacio Alvarez Illán, Andrés Ortiz, Francisco Jesús Martínez-Murcia, Fermín Segovia, Diego Salas-Gonzalez, and Shuihua Wang. Case-based statistical learning: a non-parametric implementation with a conditional-error rate svm. 2017.
13. Laila Khedher, Javier Ramírez, Juan Manuel Górriz, Abdelbasset Brahim, Fermín Segovia, Alzheimer’s Disease Neuroimaging Initiative, et al. Early diagnosis of alzheimer’s disease based on partial least squares, principal component analysis and support vector machine using segmented mri images. *Neurocomputing*, 151:139–150, 2015.
14. Fermín Segovia, JM Górriz, Javier Ramírez, Diego Salas-Gonzalez, and Ignacio Álvarez. Early diagnosis of alzheimer’s disease based on partial least squares and support vector machine. *Expert Systems with Applications*, 40(2):677–683, 2013.
15. Fermín Segovia, JM Górriz, Javier Ramírez, Diego Salas-Gonzalez, Ignacio Álvarez, Míriam López, Rosa Chaves, Alzheimer’s Disease Neuroimaging Initiative, et al. A comparative study of feature extraction methods for the diagnosis of alzheimer’s disease using the adni database. *Neurocomputing*, 75(1):64–71, 2012.
16. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
17. John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

Relevance of Filter-Banked Features using Multiple Kernel Learning for Brain Computer Interfaces

1 **Abstract.** Brain-Computer Interfaces directly communicate the human brain and
2 machines through the analysis of sensorimotor activity, relying on the Motor
3 Imagery paradigm of cognitive neuroscience. Conventional BCI systems use electroencephalographic
4 signals due to its high temporal resolution, portability, and easiness to implement,
5 for which the filter-banked analysis works as the characterization baseline. Due
6 to such analysis yields to highly dimensional representation spaces leading to
7 overtrained systems, we propose to combine the multiple spectral bands into
8 a single representation space through the maximization of the centered kernel
9 alignment criterion. As a result, the similarity between the measured EEG data
10 and the available label sets is maximized, with the additional benefit of enhancing
11 the spectral interpretation of the subject performance. The proposed κ -FB is
12 evaluated in the dataset IIa of the BCI competition IV for a binary classification
13 task. Attained accuracy proves that κ -FB outperforms other filter-banked representations
14 without compromising the system confidence.

15 **Keywords:** Brain Computer Interfaces; Common Spatial Patterns; Multiple Kernel
16 Learning

17 1 Introduction

18 Brain Computer Interfaces (BCI) take advantage of the extracted information from EEG
19 signals to establish a direct communication between the human brain and the machine.
20 BCI is used to help people with disability by means of the analysis of the human
21 sensorimotor functions, which are based on the paradigm in cognitive neuroscience,
22 named as Motor Imagery (MI). Among the non-invasive techniques for measuring brain
23 activity, like functional magnetic resonance imaging and magnetoencephalography, electroencephalography
24 (EEG) is preferred due to its portability, easiness to implement, and higher temporal
25 resolution.

26 However, BCI systems are heavily dependent on an effective feature extraction
27 from EEG signals. Common spatial patterns (CSP) approach is one of the most popular
28 techniques to extract features and it is commonly used in MI classification. CSP technique
29 constructs spatial filters that both maximizes the variance in a specific class and minimizes
30 the variance in the other one. Since achieving high classification accuracy depends on a
31 pre-specified frequency band, election of the filter is a hard task to perform due to each
32 subject has a different spectral behavior and it is determined in a manual way. Filter
33 bank common spatial patterns (FBCSP) approach is proposed to overcome the filter
34 band selection problem at the cost of providing larger feature sets that may incur in the
35 curse of the dimensionality.

36 To solve above-mentioned issue, [12] proposed a discriminant FBCSP using Fisher
37 ratio to select subject-specific filter bands instead of fixed ones aiming to enhance

38 accuracy of the FBCSP. Then, the sliding window discriminative CSP was proposed
 39 to select a discriminative feature set by exploiting affinity propagation [11]. Another
 40 approach estimates CSP features on multiple bands that are filtered from raw EEG
 41 data by a set of overlapping filter bands [14]. The Sparse Filter Band Common Spatial
 42 Pattern (SFBCSP) approach implements a lasso regression to select the significant
 43 CSP features according to the provided labels. Such approach was further improved
 44 by including a sparse Bayesian learning scheme to implement selection of significant
 45 features with a linear discriminant criterion for classification [13]. Also, the recursive
 46 band elimination was proposed to select the most discriminative frequency filters without
 47 hampering the training performance [6]. More recently, the Weighted Overlap-Add
 48 (WOLA) approach introduced the dynamic filtering using the event-related desynchronization/synchronization
 49 to yield significant CSP-based characteristics according to the classification task [3].
 50 Nevertheless, most of these feature selection methods are computationally expensive,
 51 and lack of a measure relating the frequency band with its discriminative contribution [5].
 52 Moreover, in most of the cases, there is not a suitable validation framework that allows
 53 ensuring a generalized performance, leading to overtrained systems [2].

54 Here, we introduce a filter band selection method based on CSP that computes the
 55 discriminative relevance of filter band features, reflecting their contribution to discriminate
 56 the considered motor imagery task. Instead of selecting one specific kernel function
 57 to encode whole information, the proposed Multiple kernel learning (κ -FB) combines
 58 several single kernels (i.e., one kernel per filter band) through a weighted sum, attempting
 59 to maximize the similarity between measured EEG data and available set of labels (i.e.,
 60 prior knowledge) using the Centered Kernel Alignment (CKA) cost function. Then,
 61 computed weights are interpreted as the contribution of the corresponding kernels to
 62 improve the classification performance, that is, the relevance of a filter band feature set
 63 to discriminate the considered MI classes. As an additional benefit, weights from κ -FB
 64 provide interpretability on the performed accuracy for each subject.

65 The agenda of this paper is organized in five sections: Section 2 describes the
 66 mathematical background on common spatial pattern for feature extraction and kernel
 67 learning gathering multiple representations. The developed experiments are described
 68 in Section 3. Performance results are showed and discussed in Section 4. Finally, Section 5
 69 presents the conclusions and future research directions.

70 **2 Materials and methods**

71 **2.1 Feature extraction using Common Spatial Patterns**

72 In binary classification tasks, CSP finds a spatial filter matrix $\mathbf{W} \in \mathbb{R}^{C \times 2K}$ to linearly
 73 map the EEG channel signals $\mathbf{X} \in \mathbb{R}^{C \times T}$ onto a space $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$, so that variance of
 74 the mapped signal is maximized for one class while the variance of another class
 75 is minimized. The spatial filters $\mathbf{w}^* \in \mathbb{R}^C$ to extract MI features are the solution of
 76 maximizing the Rayleigh quotient:

$$\mathbf{w}^* = \max_{\mathbf{w}} \frac{\mathbf{w}^\top \boldsymbol{\Sigma}^- \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma}^+ \mathbf{w}}, \text{ s.t.: } \|\mathbf{w}\|_2 = I_C \quad (1)$$

77 where \mathbf{I}_C is the identity matrix sizing $C \times C$, and the spatial covariance matrix of the
 78 class $l \in \{-, +\}$ is estimated as $\hat{\Sigma} = \mathbb{E} \{ \mathbf{X}_n \mathbf{X}_n^\top : n \in N_l \}$, being N_l the number of trials in
 79 class l . Notations $\|\cdot\|_2$ and $\mathbb{E} \{ \cdot \}$ stand for ℓ_2 -norm and expectation operator, respectively.

80 The optimization framework in Eq. (1) can be equivalently transformed into the
 81 generalized eigenvalue problem $\Sigma^- \mathbf{w}^* = \lambda \Sigma^+ \mathbf{w}^*$ with $\lambda \in \mathbb{R}^+$. Thus, a set of spatial
 82 filters $\mathbf{W}^* = [\mathbf{w}_1^* \dots \mathbf{w}_{2K}^*]$ can be obtained by collecting eigenvectors corresponding to
 83 the K largest and smallest eigenvalues of the generalized eigenvalue problem. To account
 84 further for brain activity at different spectral bandwidths, each n -trial of input MI data
 85 is bandpass filtered within a set of F frequency bands $\{\mathbf{X}_{n,f} : f \in F\}$. The CSP feature
 86 vector is then formed as $\xi = [\xi_k : k \in 2K]$ with $\xi \in \mathbb{R}^{2FK}$, having entries $\xi_{n,k}^f = \ln(\text{var}\{\mathbf{w}_k^{*\top} \mathbf{X}_{n,f}\})$,
 87 where $\text{var}\{\cdot\}$ stands for the variance operator.

88 2.2 Multiple kernel learning for filter-bank representation

To measure the pairwise proximity between trials at frequency band f , we perform
 the dot product-based kernel function as a measure of similarity between functionals,
 resulting in a set of kernels $\{\kappa(\xi_n^f, \xi_m^f) \in \mathbb{R}^+ : f \in F\}$ that must be properly combined to
 work out a unique similarity value for each pairwise trial comparison. We address this
 issue through a Multiple Kernel Learning (MKL) framework by the following weighted
 sum [8]:

$$\kappa(\xi_n, \xi_m; \boldsymbol{\mu}) = \sum_{f \in F} \mu_f \kappa(\xi_n^f, \xi_m^f), \quad (2a)$$

$$\text{s.t.} : \sum_{f \in F} \mu_f = 1, \mu_f \geq 0 \quad (2b)$$

89 where vector $\boldsymbol{\mu} \in \mathbb{R}^F$ holds the mixture weights that must be optimized to improve
 90 discrimination between classes across the label set. Under the constraints in Eq. (2b)
 91 to reproduce a convex function in Eq. (2a), optimization of $\boldsymbol{\mu}$ is carried out by the
 92 maximization of Centered Kernel Alignment, so that each CSP feature set is weighted
 93 according to its discrimination capability: the higher the weight, the more significant its
 94 contribution to the classifier performance.

95 As a result, the proposed approach using Multiple kernel learning for filter-bank
 96 representation in Eqs. (2a) and (2b) (noted as κ -FB) allows selecting the CSP feature
 97 sets extracted from the most relevant frequency bands for discriminating a motor imagery
 98 task at hand.

99 3 Experimental set-up

100 3.1 EEG dataset and preprocessing

101 The κ -FB approach is evaluated on a BCI competition IV dataset IIa¹, holding EEG
 102 data recorded by 22-electrode montage. Nine subjects were instructed to perform four
 103 MI tasks (“left hand”, “right hand”, “both feet”, and “tongue”). On different days, each

¹ available at <http://www.bbc.de/competition/iv/>

104 subject completed two sessions, each one including six runs to perform 48 trials (i.e.,
 105 12 trials per class) and resulting in 288 trials per sitting. In order to compare our results
 106 with CSP, *s*-FB and WOLA, we consider just the bi-class classification task: “left”
 107 versus “right hand”.

108 In the preprocessing stage, all EEG recordings are band-pass filtered between [0.5–
 109 100] *Hz* and sampled at 250 *Hz*, followed by a fifth order Butterworth band-pass filter to
 110 remove noises over 40 *Hz* and slow baseline signal under 4 *Hz*. Sub-band decomposition
 111 of each channel is further performed, for which the linearly distributed filters have been
 112 heuristically adjusted to 7 *Hz* bandwidth and 90% overlap. In total, we calculate 42
 113 fifth-order Butterworth band-pass filters as to cover the whole bandwidth from 4 to
 114 40 *Hz*. Focusing on the learning part of the MI task, each trial recording is temporarily
 115 segmented, extracting for analysis the interval that ranges from 2.5 *s* to 4.5 *s*. Lastly,
 116 CSP-based feature extraction is accomplished from the preprocessed recordings, yielding
 117 42 feature vectors of dimension 22 per trial. However, we only take the first and the last
 118 three vectors [4].

119 3.2 Evaluation scheme and performance assessment

120 Due to the provided universal approximating property, the κ -FB approach is implemented
 121 calculating one Gaussian kernel for each filter, and their bandwidth parameters are
 122 computed as in [1]. For the sake of comparison, κ -FB is contrasted with CSP and
 123 two CSP-based approaches: The sparse version of filter bank common spatial pattern
 124 (*s*-FB) that implements a sparse regression of filter-bank CSP features to predict the trial
 125 labels [14]; and (WOLA) that generate different filter bands increasing the discrimination
 126 between classes “left” and “right hand” [3].

127 The classifier is evaluated on a kernelized *k*-nearest neighbors (*Kk*-NN) machine,
 128 for which the number of neighbors is tuned via an exhaustive search, obtaining $N = 3$
 129 as the best parameter performance, besides, we estimate the average accuracy (a_c) by a
 130 five-fold cross-validation scheme as a measure of classifier performance.

131 4 Results and discussion

132 By handling a convex combination of introduced functional kernels, here, the feature
 133 selection of Filter Bank Common Spatial Patterns is performed to improve discrimination
 134 capability in Motor Imagery tasks. During validation of κ -FB approach, following
 135 findings come out from the experiments:

136 Multiple kernel learning is used to improve class discrimination in MI tasks by
 137 optimizing the mixture weights of each kernel. Optimization is performed by the maximization
 138 of Centered Kernel Alignment, that learn a mapping matrix matching the provided
 139 binary-class label set with all kernel frequency bands. This optimization highlights the
 140 most discriminant frequency bands. As seen in Fig. 1, the proposed κ -FB approach
 141 outperforms the classifier accuracy reached by other compared methods in most of the
 142 subjects. Moreover, the average accuracy across all subjects is as follows: CSP= 77.1%,
 143 *s*-FB= 82.7%, WOLA= 78.9%, and κ -FB= 84.8%.

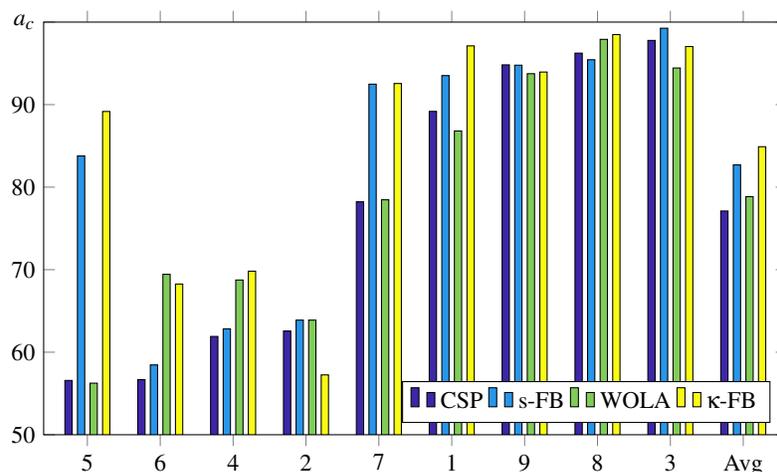


Fig. 1. Accuracy results per subject on the test set for all considered CSP-based approaches for motor imagery classification task.

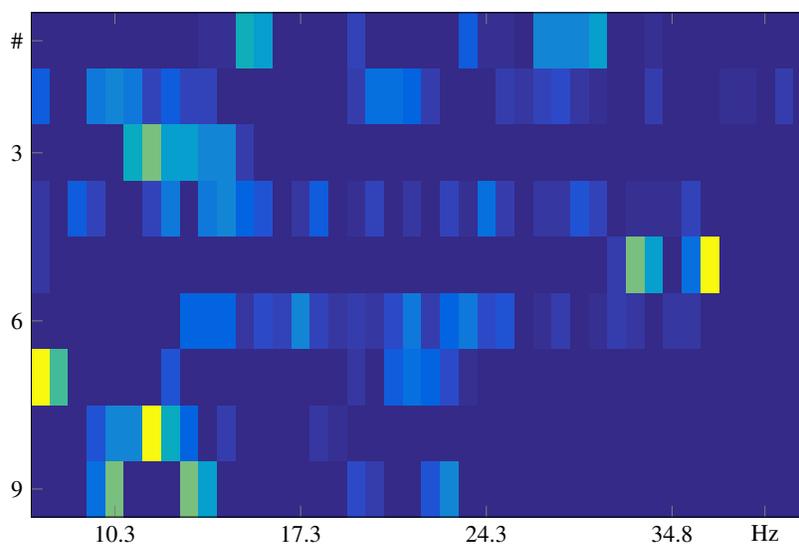


Fig. 2. MKL weights along the frequency bands computed per subject.

144 As seen in Fig. 1, filter-bank methods reach a marginal difference in comparison
 145 with CSP accuracy among the best subjects for CSP (#9, #8, #3) due to CSP reach a
 146 robust feature discrimination. On the contrary, performance of subjects with the worse
 147 accuracy for CSP (#5, #6, #4) are improved suggesting those subjects have their own
 148 specific frequency range to perform the MI task. In subject #5, it is interesting to analyze
 149 that WOLA has the lowest accuracy despite it use frequency partitions suggesting that

150 the obtaining dynamic filters could not taking the most active components containing in
 151 the events related to left and right hand movements, in fact, those filters are misleading
 152 the classification stage. Besides, our method has the lowest performance in subject
 153 #2 indicating that the specific bands used for this task are not taking into account
 154 due to the bandwidth and overlap used in the experiment. In turn, κ -FB improves
 155 both, performance and interpretability, since we estimate the relevance directly from
 156 the weights corresponding to specific frequency bands. For the sake of illustration,
 157 Fig. 2(a) depicts the relevance of each filter band per subject, where subjects with
 158 widespread weights (#2, #6, #4) reach the lowest accuracy rates. On the contrary, the
 159 most sparse-distributed weights correspond to the best performing subjects (#8, #1, #3).
 160 At the same time, subject #5 presents a particular distribution with highly relevant bands
 161 over 30 Hz suggesting that he manages MI tasks different than the others do.

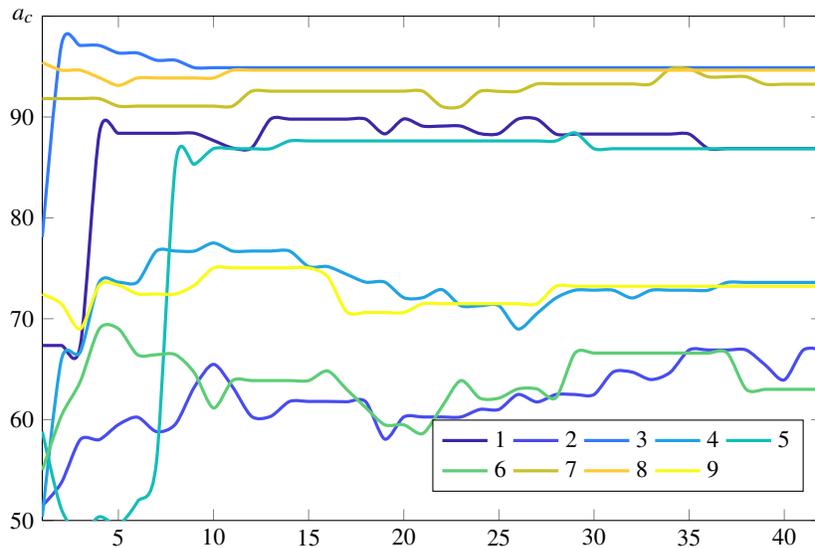


Fig. 3. Validation accuracy curve according to the descending sorted MKL weights.

162 Fig. 3(a) displays classification performance when successively including filter bands
 163 to the convex combination from the most to the least relevant. The attained curves show
 164 that the test accuracy grows quickly for the first sub-bands. Any other attached filter
 165 varies the performance around the best accuracy meaning that the remaining kernels
 166 are noisy or have redundant information that do not contribute to the discrimination
 167 task. Particularly, in subjects 1, 3, and 5 including the fourth, second, and seventh
 168 bands considerably enhances the accuracy in $\sim 20\%$, $\sim 19\%$ and $\sim 30\%$, respectively.
 169 The lowest performances are obtained by the subjects 2 and 6, where the incremental
 170 learning curves reflect high variability in the accuracy as more kernels are combined.
 171 Such a behavior is the consequence of the noisy information in some filter bands.

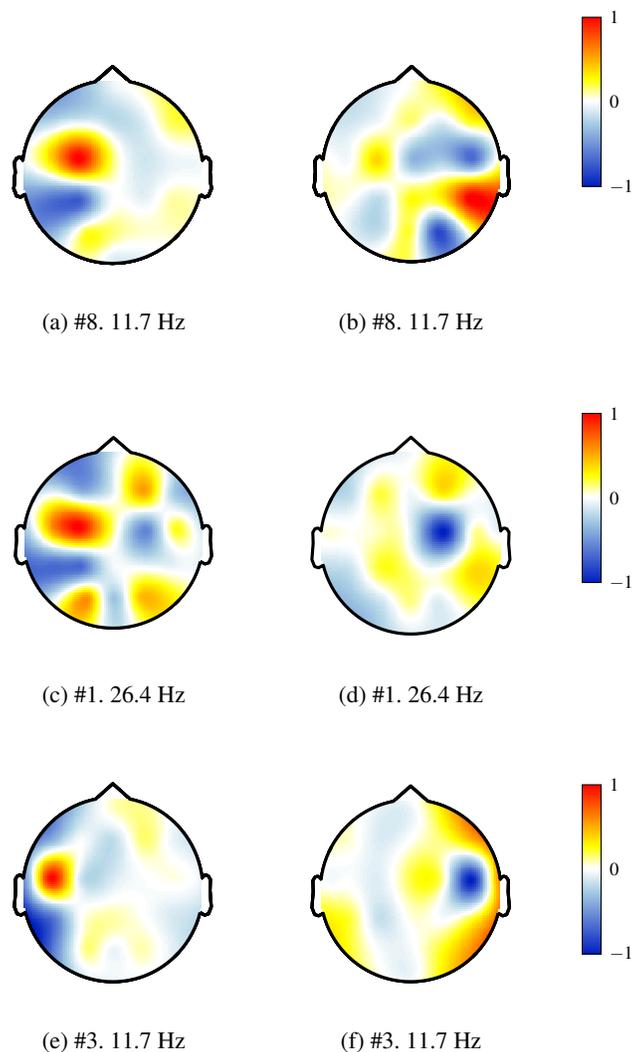


Fig. 4. Estimated spatial filters for the best performing subjects (#8, #1, #3) in the most weighted frequency band. Left: First filter ($k=1$). Right: Last filter ($k=2K$).

172 As seen from the topoplots in Fig. 4, the proposed κ -FB method highlights more
 173 discriminative information contained in the μ band and localized around of the central
 174 cortex; this brain region has been associated with planning and executing of voluntary
 175 movements [10]. Besides, the activation surrounding the frontoparietal area ($FC3$, $FC4$,
 176 $CP3$, and $CP4$) is related to the subject thinking strategy during MI tasks [9]. On the
 177 other hand, Fig. 5 reveals that the spatial filters associated with the largest-weighted
 178 frequency band involve most of the brain areas for the worst performing subjects.

179 This fact is due to the lack of discriminant information in the motor cortex, which
 180 may be related to disconnected electrodes, signal artifacts, subject tiredness, or lack of
 181 attention [7].

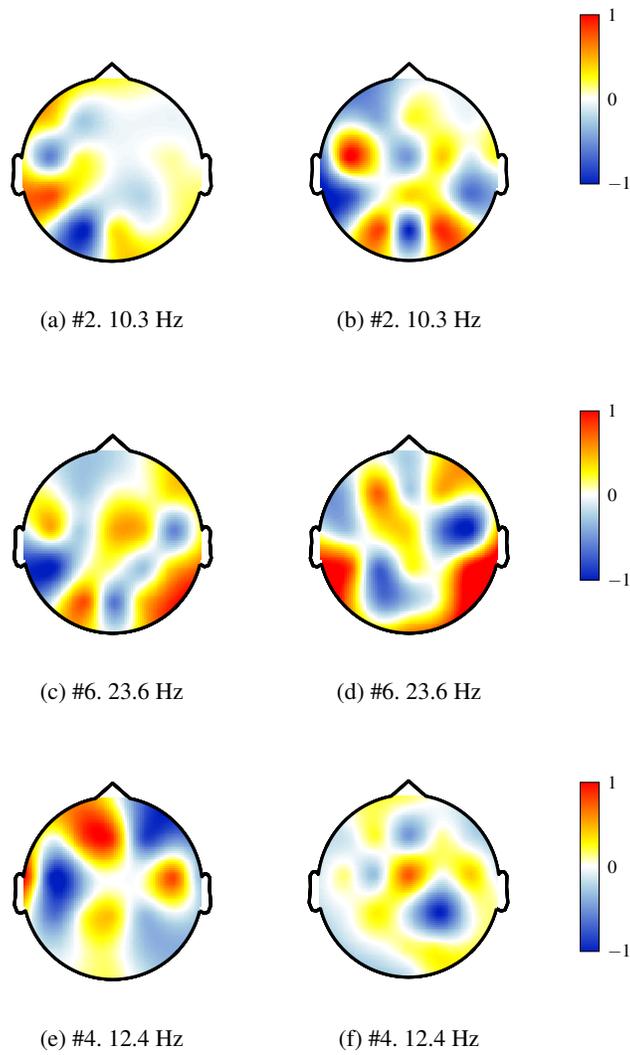


Fig. 5. Estimated spatial filters for the worst performing subjects (#2, #6, #4) in the most weighted frequency band. Left: First filter ($k=1$). Right: Last filter ($k=2K$).

182 5 Conclusion and future work

183 From the attained results, we conclude that filter-banked approaches generally enhance
 184 the BCI system performance. However, properly choosing frequency bands is a hard
 185 task due to each subject has a specific spatio-spectral behavior to perform the MI task.
 186 In this regard, the proposed κ -FB highlights the contribution of the frequency bands
 187 with the spatial information better discriminating the classes. Our approach improve the
 188 single band CSP that cannot unravel the relevant information, and the SFB that forces
 189 the lasso regressor to detect a few features from a large set. Then, matching data and
 190 label spaces using the centered kernel alignment results in a set of relevance weights
 191 that provide enhanced interpretability of performed accuracy by each subject.

192 For the future work, we plan to extend the kernel learning to other augmented
 193 representations, for instance, including time sliding windowing to decode information
 194 from time-varying event-related potentials and improving the discrimination of MI
 195 tasks. From the machine learning point of view, other kernel functions, as the polynomial
 196 and the Stein, and kernel combination strategies, as the tensor-based, must be explored
 197 to interpret discriminating information extracted from subjects with a complicated execution
 198 of MI tasks. Besides, we plan to use dynamic filtering to find the specific frequency band
 199 used for each subject in this kind of tasks enhanced both performance and interpretability.

200 References

- 201 1. Álvarez-Meza, a.M., Cárdenas-Peña, D., Castellanos-Dominguez, G.: Unsupervised Kernel
 202 Function Building Using Maximization of Information Potential Variability. *Progress in*
 203 *Pattern Recognition, Image Analysis, Computer Vision, and Applications SE - 41* **8827**,
 204 335–342 (2014)
- 205 2. Alvarez-Meza, A.M., Velasquez-Martinez, L.F., Castellanos-Dominguez, G.: Time-series
 206 discrimination using feature relevance analysis in motor imagery classification.
 207 *Neurocomputing* **151**(P1), 122–129 (2015)
- 208 3. Belwafi, K., Romain, O., Gannouni, S., Ghaffari, F., Djemal, R., Ouni, B.: An embedded
 209 implementation based on adaptive filter bank for brain–computer interface systems. *Journal*
 210 *of neuroscience methods* (2018)
- 211 4. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.R.: Optimizing Spatial
 212 Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine* (January
 213 2008), 41–56 (2008)
- 214 5. Das, A.K., Suresh, S., Sundararajan, N.: A discriminative subject-specific spatio-spectral
 215 filter selection approach for EEG based motor-imagery task classification. *Expert Systems*
 216 *with Applications* **64**, 375–384 (2016)
- 217 6. Das, A.K., Suresh, S., Sundararajan, N., Subramanian, K.: A subject-specific frequency band
 218 selection for efficient BCI- an interval type-2 fuzzy inference system approach. 2015 IEEE
 219 International Conference on Fuzzy Systems (FUZZ-IEEE) pp. 1–8 (2015)
- 220 7. Dornhege, G.: *Toward brain-computer interfacing*. MIT press (2007)
- 221 8. Gönen, M., Alpaydın, E.: Multiple Kernel Learning Algorithms. *Journal of Machine*
 222 *Learning Research* **12**, 2211–2268 (2011)
- 223 9. Hanakawa, T., Immisch, I., Toma, K., Dimyan, M.A., Van Gelderen, P., Hallett, M.:
 224 Functional properties of brain areas associated with motor execution and imagery. *Journal of*
 225 *neurophysiology* **89**(2), 989–1002 (2003)

- 226 10. Jeannerod, M.: Fundamentals of cognitive neuroscience. The cognitive neuroscience of
227 action. Malden,: Blackwell Publishing (1997)
- 228 11. Sun, G., Hu, J., Wu, G.: A novel frequency band selection method for Common Spatial
229 Pattern in Motor Imagery based Brain Computer Interface. In: The 2010 International Joint
230 Conference on Neural Networks (IJCNN). pp. 1–6. IEEE (jul 2010)
- 231 12. Thomas, K.P., Guan, C., Lau, C.T., Vinod, A.P., Ang, K.K.: A New Discriminative
232 Common Spatial Pattern Method for Motor Imagery Brain 2013;Computer Interfaces. IEEE
233 Transactions on Biomedical Engineering **56**(11), 2730–2733 (nov 2009)
- 234 13. Zhang, W., Sun, F., Tan, C., Liu, S.: Low-Rank Linear Dynamical Systems for Motor
235 Imagery EEG. Computational Intelligence and Neuroscience (2016)
- 236 14. Zhang, Y., Zhou, G., Jin, J., Wang, X., Cichocki, A.: Optimizing spatial patterns with sparse
237 filter bands for motor-imagery based brain-computer interface. Journal of Neuroscience
238 Methods **255**, 85–91 (2015)

Multiple Instance Learning Selecting Time-Frequency Features for Brain Computing Interfaces

No Author Given

No Institute Given

Abstract. Brain-Computer Interface is a technology which uses measures of brain activity to help people with motor disabilities. BCI applications based on Electroencephalography commonly rely on Motor Imagery paradigm. However, the estimation of motor brain patterns is affected by both variations in the signal properties over time (i.e. non-stationarity) and differences between frequency bands activations. Generally, Common Spatial Patterns is used as feature extraction. Nevertheless, its performance depends on the filter band selection and the time when the brain activity is associated with the task. A new method of time-frequency segmentation based on multi-instance learning is proposed. The spatial filters are built taking to account the obtained frequency-temporal segments where an instance selection based on Sparse Representation Classification method is developed together with a feature selection stage. The experiments are developed using a well-known dataset BCI competition IV dataset Ila that contains EEG records of nine subjects recorded from 22-electrodes mesh. The results evidencing that significant features appear at the end of MI interval and the found spatial patterns are consistent with MI neurophysiology. Furthermore, the proposed method outperforms the average classification accuracy of CSP, SFTOFCRC and TSGSP for 8.21%, 1.23% and 2.21% respectively without deteriorating classification accuracy with statistical significance for subjects that present high accuracy with the compared methods.

Keywords: Electroencephalography· Motor Imagery· Multi-Instance Learning· Feature selection· Instance selection.

1 Introduction

Nowadays, Brain-Computer Interface (BCI) is an important technology which uses measures of brain activity to help people with motor disabilities allowing them to an efficient communication with the world. Usually, BCI applications are based on electroencephalography (EEG) due to its non-invasive brain signal measurement method. A common EEG-based BCI system relies on Motor Imagery (MI) paradigm that is the imagination of a given motor action without executing it. Decoding MI activity can be checked by the changes in the power bands [14]. However, estimating these changes are affected by i) Variations in the signal properties over time known as non-stationarity due to the natural variability of neural responses even for the same condition. Additionally, these variations are associated to artifacts like loose electrodes, muscle movements,

blinking, sudden shifts of attention, and effects of tiredness [7]. ii) Differences between subjects brain activations, i.e. differences in frequency bands which are related to developed brain network for MI activations [1].

Consequently, the discrimination of MI activations requires a feature representation, feature extraction, and classification methodologies that must be suitably developed to reveal the main brain patterns from EEG. Common Spatial Pattern (CSP) is an algorithm proposed in [8] as a strategy of MI feature extraction. It searches optimal spatial filters to maximize variance for a class while minimizing variance for another class. However, CSP performance depends on the filter band selection [5]. In this regard, several methods are proposed as in [18] Optimum Spatio-Spectral Filtering Network (OSSFN) and in [20] Sparse Filter Band Common Spatial Pattern (SFBCSP). OSSFN proposed a CSP feature selection from multiple frequency bands. Similarly, SFBCSP method introduced by exploiting sparse regression for automatic frequency band selection. Nevertheless, the time during which the subject performs the MI task is known, the time when the brain activity is associated with the task is unknown. Despite, the significance of temporal features is exposed in [14, 6, 19] the extracted temporal features are ponderated without taking into account the variability between trials of the neural response.

A new method of time-frequency segmentation based on multi-instance learning defined as MILCSP is proposed. Our proposal build the spatial filters taking to account the obtained frequency-temporal segments. Each time-frequency segment CSP features is considered as an instance belonging to a bag. A relevance analysis allows analyzing of each time-frequency segment depending on the form of cerebral activation presented in each trial, without being tied to the search for a time window or a general frequency band for each subject. Multi-Instance Learning (MIL) framework used in [9, 13, 11] proposes an instance selection in order to choose the most relevant information for the classification stage. In summary, MILCSP in order to use only the significant time-frequency information, the instance selection is developed based on Sparse Representation Classification (SRC) [16] method which is proposed together with a feature selection stage as in [14].

2 Materials and methods

2.1 Instance selection

A multiple-instance learning problem holds a set of samples (bags) formed by several feature vectors (instances) [10]. For binary discrimination of MI tasks, we define the i -th EEG trial as a bag $\mathbf{B}_i \in \mathbb{R}^{M \times p}$ with M feature vectors extracted from time windows sliding along frequency bands, termed time-frequency instances, $\mathbf{x}_{ij} \in \mathbb{R}^p$, and a label $l_i \in \{-1, +1\}$. Based on the sparse representation classifier (SRC) [16], bag instances are ranked according to their results for regressing instances of the same class as follows: $\hat{\boldsymbol{\beta}}_i = \arg \min_{\boldsymbol{\beta}_i} \left(\frac{1}{2} \|\mathbf{A}_i \boldsymbol{\beta}_i - \mathbf{x}_{ij}\| + \lambda_1 \|\boldsymbol{\beta}_i\|_1 \right)$, where $\|\cdot\|_1$ and $\|\cdot\|$ denote the L1 and euclidean norm, respectively, $\boldsymbol{\beta}_i \in \mathbb{R}^M$ represents the vector of scalar coefficients for the linear regression, λ_1 is a positive regularization parameter which controls the sparsity of $\boldsymbol{\beta}_i$, and the dictionary \mathbf{A}_i holds the instances of other bags belonging to the class of \mathbf{B}_i ,

i.e., $\mathbf{A}_i = \{\mathbf{x}_{kj} : \forall l_k = l_k; k \neq i\}$. As a result, an instance of the i -th bag is sparsely reconstructed as the linear combination of the elements within the i -th dictionary $\mathbf{y}_{ij} = \mathbf{A}_i \boldsymbol{\beta}_i$. Further, the residual criterion selects the instances with reconstruction error smaller than a threshold as [17]:

$$\tilde{\mathbf{B}}_i = \{\tilde{\mathbf{x}}_{ij} : \|\mathbf{x}_{ij} - \mathbf{y}_{ij}\| < \mathbb{E} \{\|\mathbf{x}_{ik} - \mathbf{y}_{ik}\|\}; k=1, \dots, M\} \quad (1)$$

2.2 Supervised discriminant representation

Under the assumption that discriminant time-frequency instances vary for each trial, a sparse regression of the labels allows selecting the instances better classifying a bag in a supervised framework: $\mathbf{z} = \arg \min_{\mathbf{z}} (\frac{1}{2} \|\mathbf{S}\mathbf{z} - \mathbf{l}\| + \lambda_2 \|\mathbf{z}\|_1)$. Where the vector $\mathbf{l} \in \{-1, +1\}^N$ contains the labels of the N training bags, $\mathbf{z} \in \mathbb{R}^q$ is the sparse vector to be learned, $\lambda_2 \in \mathbb{R}^+$ is the regularization parameter. Matrix $\mathbf{S} \in \mathbb{R}^{q \times N}$ contains the dissimilarity representation of the training bags:

$$\mathbf{S} = \begin{bmatrix} s(\tilde{\mathbf{x}}_1, \mathbf{B}_1) & \cdots & s(\tilde{\mathbf{x}}_1, \mathbf{B}_N) \\ s(\tilde{\mathbf{x}}_2, \mathbf{B}_1) & \cdots & s(\tilde{\mathbf{x}}_2, \mathbf{B}_N) \\ \vdots & \ddots & \vdots \\ s(\tilde{\mathbf{x}}_q, \mathbf{B}_1) & \cdots & s(\tilde{\mathbf{x}}_q, \mathbf{B}_N) \end{bmatrix}. \quad (2)$$

being $s(\tilde{\mathbf{x}}_q, \mathbf{B}_i)$ the dissimilarity function between the bag \mathbf{B}_i and the selected instance $\tilde{\mathbf{x}}_q$ [9]: $s(\tilde{\mathbf{x}}, \mathbf{B}_i) = \max_j \exp\left(-\frac{\|\mathbf{x}_{ij} - \tilde{\mathbf{x}}\|^2}{\sigma^2}\right)$. With $\sigma \in \mathbb{R}^+$ is the bandwidth of a radial basis function. As a result, the instances with $|z_q| > 0$ compose a set of the most discriminant features that correspond to the time-windows and frequency-bands that better reflect the MI task, so that the performance of classification machines is improved.

3 Experimental Set-Up

3.1 BCI competition IV dataset IIa

This database contains EEG records of nine subjects recorded from 22-electrodes mesh¹. The subjects are instructed to perform four types of MI classes (left, hand, right hand, both feet, and tongue) in two sessions. Each session comprehends 6 runs with 48 trials (i.e., 12 trials per class and 288 trials per session). At the beginning of each trial, a cross of fixation is shown on black screen. After 2 seconds a cue in the form of an arrow pointing to the left, right, down or up (corresponding to the four classes respectively) appears by 1.25 seconds, where the subjects are asked to carry out the MI task until the fixation cross disappeared from the screen after 4 seconds. Finally, a short break is made where the screen is black. The EEG signals are sampled at $F_s = 250\text{Hz}$ and bandpass-filtered between 0.5Hz and 100Hz. In this analysis, all the trials marked with artifact are excluded and the EEG signals between 0.5 and 2.5 seconds after the visual cue onset are extracted for training as in[2].

In order to compare the proposed MILCSP method with the other state-of-the-art approaches, the first two classes are selected.

¹ <http://www.bbci.de/competition/iv/>

3.2 Parameter tuning

CSP method is applied to a set of 48 frequency-temporal segments obtained from both frequency filter bands and time windows for each trial. To obtain the frequency-temporal segments, each trial is filtered using 16 sliding filters of 4Hz between 6–40Hz with 2Hz overlap. Then, each slide filtered trial is segmented using sliding one-second time windows with 50% overlap[14]. Thereafter, a bag of instances is generated, which contains all the frequency-temporal components of each trial. Additionally, a dictionary is created for each bag and the instances are selected using a sparse regression that uses the regularization parameter λ_1 . After the instance-based feature mapping to the dissimilarity space, a sparse regression (with regularization parameter λ_2) is performed in order to select features. The parameters λ_1 and λ_2 are tuned according to the maximum classification accuracy for each subject through a thorough search. Finally, the classification stage is developed using a KSVM through 5-fold cross-validation scheme. A Gaussian kernel is employed to estimate the similarity between \mathbf{s}_i and \mathbf{s}_t as $K(\mathbf{s}_i, \mathbf{s}_t) = \exp(-\|\mathbf{s}_i - \mathbf{s}_t\|_2^2/2\theta^2)$ where the kernel bandwidth parameter θ is tuned as proposed in [4]. Figure 1 presents chosen parameters λ_1 and λ_2 for the maximum accuracy value (represented by the red dot). For illustrative purposes, these results are presented for the subjects 2, 3, and 5.

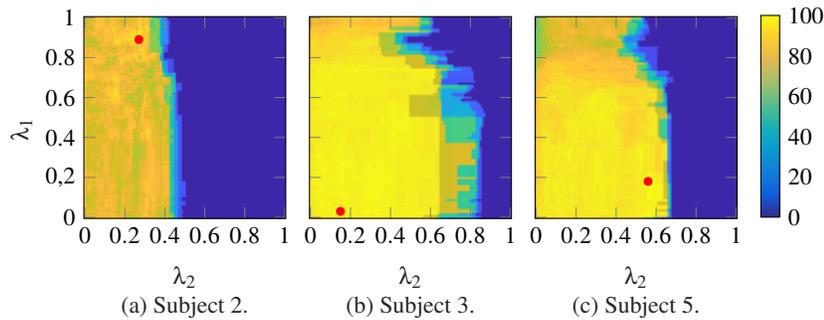


Fig. 1: Parameter tuning of features and instances selection.

4 Results and Discussion

This paper proposes a new approach to the analysis of EEG mainly based on both instance and feature selection in a multiple instance learning framework. The importance of the two mentioned stages can be highlighted by analyzing the behavior of the subjects shown in the figures 1 and 2. Accordingly, for the subject 2 (1-a), as is shown for the selected parameters values $\lambda_1=0.89$ and $\lambda_2=0.27$, the regularization generates a smaller instance selection matrix. On the other hand, the subject 5(1-c) presents values $\lambda_1=0.18$ and $\lambda_2=0.56$, i.e., that not all the selected instances serve as discriminating features for this subject, so, the greatest contribution of the algorithm in this particular case is in the feature selection. This finding is in accordance to [3], who reported that

subjects with the lower performance show fewer prominent features(or instances in our case) than those who perform better. Finally, subject 3(1-b) shows little regularization in the two stages($\lambda_1=0.03$ and $\lambda_2=0.15$), which shows that the proposed approach select effectively the most relevant regularization for each subject.

The instance selection and the feature selection with and without the instance selection are presented in Figure 2. The analysis shows a grid of 16 frequency bands for 3 time windows where the lighter colors represent the most often selected components in both instance selection and feature selection stages. The subject 5 shows activity during the two final segments of time in frequencies corresponding to high β band that is present in some subjects as is presented in [1]. These results evidencing that significant features appear at the end of MI interval and the importance of the time-frequency segmentation and instance-features selection presented. According to the subject 3, while the obtained classification accuracy by CSP, SFTOFSRC, TSGSP and the proposed MILCSP method (see Figure 4) is high, the selected features shows activity in the mu(μ) which is mainly linked to the motor cortex activity.

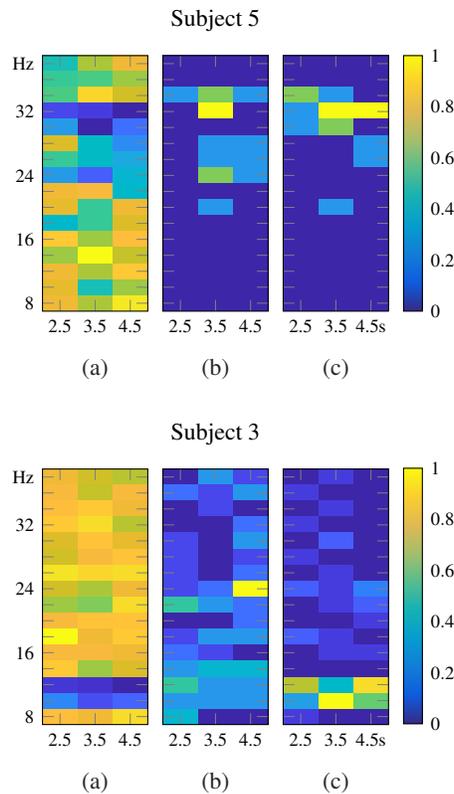


Fig. 2: Relative frequency of the time-frequency components for instance selection (a), feature selection without (b) and with (c) instance selection in the subjects 3 and 5.

Figure 3 shows the computed spatial filters using the proposed MILCSP for both the two subjects with the worst and the two subjects with the best classification performing (i.e. the subjects 5, 2, 8, and 3). The found CSP spatial patterns include the discriminative information between all time-frequency extracted windows. For all the subjects the proposed method MILCSP found spatial patterns over the primary motor cortex (M1) and others secondary motor cortices as the posterior parietal cortex (PP) and supplementary motor area (SMA) which is consistent with MI neurophysiology [15, 12]. Movement preparation links the M1 region which is important in sensory processing, the PP that translate visual information into motor commands and generating mental movement representations, and SMA which is important for planning and coordinating tasks. However, for the subjects 5 and 2 also show spatial patterns in frontal and prefrontal regions. This activation is related to higher mental functions as concentration or emotional expressions, that as is shown in [1], psychological and physiological states generate a low performance, as well as, reported the MI illiteracy having a less-developed brain network that is incapable of motor imagery.

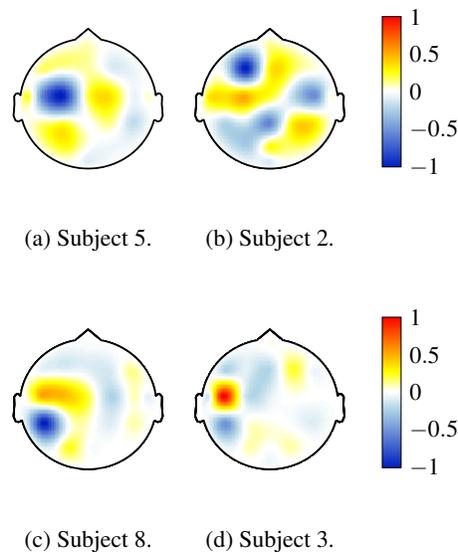


Fig. 3: spatial pattern during MI of left and right hand using the proposed MILCSP.

Figure 4 contains classification accuracies for the nine subjects using CSP [8], SFTOFSRC [14], TSGSP [19], and the proposed MILCSP. The average classification accuracy for the methods above are 76.50%, 83.48%, 82.5%, and 84.71% respectively. In this way, the results of the proposed method are both significantly better than CSP and competitive compared to the others two methods improving the accuracy especially in subjects 2, 6 and 7 with respect to SFTOSRC and subjects 1, 6 and 7 with respect to

TSGSP, without deteriorate classification accuracy with statistical significance for other subjects.

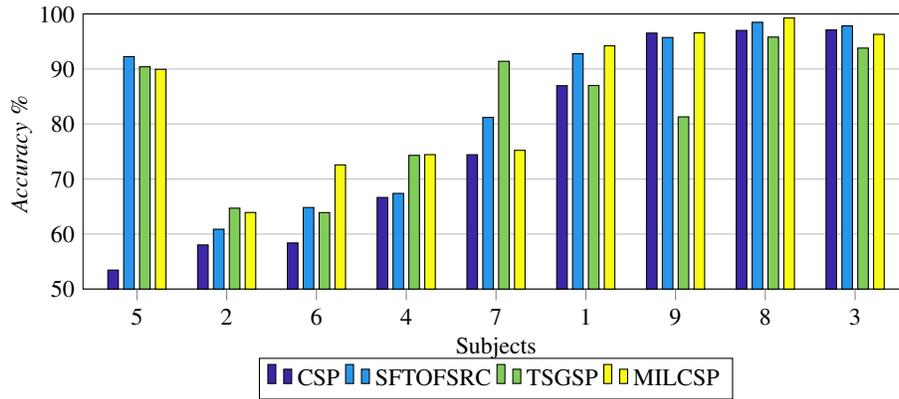


Fig. 4: Comparison of classification accuracies obtained by CSP, SFTOFSRC, TSGSP and the proposed method, MILCSP.

5 Conclusions and future work

This work proposed a new method of time-frequency segmentation based on multi-instance learning. The multiple-instance learning techniques together with the selection of instances and features allow an effective selection of the frequency-temporal segments where the cognitive task is carried out, allowing to use the relevant information in each domain. The results evidencing that significant features appear at the end of MI interval and the found spatial patterns are consistent with MI neurophysiology. Furthermore, the proposed method outperforms the average classification accuracy of CSP, SFTOFCRC and TSGSP for 8.21%, 1.23% and 2.21% respectively without deteriorating classification accuracy with statistical significance for subjects that present high accuracy with the compared methods.

From the attained findings of the work, we propose two future research directions for MIL: Firstly, we plan to improve the dictionary construction using multiple bags aiming to reduce the computational cost. Secondly, we intend to combine multiple similarity measures and features extraction procedures, highlighting discriminant properties of trials from subjects with low performance.

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. Alimardani, F., Boostani, R., Blankertz, B.: Weighted spatial based geometric scheme as an efficient algorithm for analyzing single-trial EEGs to improve cue-based BCI classification. *Neural Networks* **92**, 69–76 (2017). <https://doi.org/10.1016/j.neunet.2017.02.014>, <http://dx.doi.org/10.1016/j.neunet.2017.02.014>
3. Allison, B.Z., Neuper, C.: Could anyone use a bci? In: *Brain-computer interfaces*, pp. 35–54. Springer (2010)
4. Álvarez-Meza, A.M., Cárdenas-Peña, D., Castellanos-Dominguez, G.: Unsupervised Kernel Function Building Using Maximization of Information Potential Variability. In: Bayro-Corrochano, E., Hancock, E. (eds.) *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. pp. 335–342. Springer International Publishing, Cham (2014)
5. Ang, K.K., Chin, Z.Y., Wang, C., Guan, C., Zhang, H.: Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience* **6**, 39 (2012)
6. Balzi, A., Yger, F., Sugiyama, M.: Importance-weighted covariance estimation for robust common spatial pattern. *Pattern Recognition Letters* **68**, 139–145 (2015)
7. Bian, Y., Qi, H., Zhao, L., Ming, D., Guo, T., Fu, X.: Improvements in event-related desynchronization and classification performance of motor imagery using instructive dynamic guidance and complex tasks. *Computers in Biology and Medicine* **96**, 266 – 273 (2018). <https://doi.org/https://doi.org/10.1016/j.compbiomed.2018.03.018>, <http://www.sciencedirect.com/science/article/pii/S0010482518300751>
8. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.R.: Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine* **25**(1), 41–56 (2008). <https://doi.org/10.1109/MSP.2008.4408441>
9. Chen, Y., Bi, J., Wang, J.Z., Member, S.: MILES : Multiple-Instance Learning via Embedded Instance Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(12), 1–17 (2006)
10. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* **89**(1-2), 31–71 (1997). [https://doi.org/10.1016/S0004-3702\(96\)00034-3](https://doi.org/10.1016/S0004-3702(96)00034-3), <http://linkinghub.elsevier.com/retrieve/pii/S0004370296000343>
11. Fu, Z., Robles-Kelly, A., Zhou, J.: MILIS: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(5), 958–977 (2011). <https://doi.org/10.1109/TPAMI.2010.155>
12. Hanakawa, T., Immisch, I., Toma, K., Dimyan, M.A., Van Gelderen, P., Hallett, M.: Functional properties of brain areas associated with motor execution and imagery. *Journal of neurophysiology* **89**(2), 989–1002 (2003)
13. Li, W.J., Yeung, D.Y.: MILD: Multiple-instance learning via disambiguation. *IEEE Transactions on Knowledge and Data Engineering* **22**(1), 76–89 (2010). <https://doi.org/10.1109/TKDE.2009.58>
14. Miao, M., Wang, A., Liu, F.: A spatial-frequency-temporal optimized feature sparse representation-based classification method for motor imagery EEG pattern recognition. *Medical and Biological Engineering and Computing* **55**(9), 1589–1603 (2017). <https://doi.org/10.1007/s11517-017-1622-1>
15. Saiote, C., Tacchino, A., Bricchetto, 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)
16. Shin, Y., Lee, S., Lee, J., Lee, H.N.: Sparse representation-based classification scheme for motor imagery-based brain-computer interface systems. *Journal of Neural Engineering* **9**(5) (2012). <https://doi.org/10.1088/1741-2560/9/5/056002>
17. Wright, J., Yang, a.Y., Ganesh, a., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and ma-*

- chine intelligence **31**(2), 210–227 (2009). <https://doi.org/10.1109/TPAMI.2008.79>, <http://www.ncbi.nlm.nih.gov/pubmed/21646680>
18. Zhang, H., Chin, Z.Y., Ang, K.K., Guan, C., Wang, C.: Optimum spatio-spectral filtering network for brain–computer interface. *IEEE Transactions on Neural Networks* **22**(1), 52–63 (2011)
 19. Zhang, Y., Nam, C.S., Zhou, G., Jin, J., Wang, X., Cichocki, A.: Temporally constrained sparse group spatial patterns for motor imagery bci. *IEEE Transactions on Cybernetics* (99), 1–11 (2018)
 20. 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)

Event Study in Tehran Stock Exchange: Central Bank Intervention and Market Impact Reaction

GholamReza Keshavarz-Haddad^{*1}, Hadi Heidari²

Abstract

Market impact is one of the most crucial components of transaction cost in a stock market which arises from information leakage or liquidity demand. In this paper we investigate the effect of the Jan 2015 Central Bank of Iran (CBI) policy measure on transaction cost as an economic event. We use 5-minuts intra-day microstructure data for 497 companies from the Tehran Stock Exchange (TSE) market. The key contribution of this study is the use of Dynamic Diff-in-Diff and propensity score matching (PSM) to identify the CBI intervention effect on the transaction cost in the TSE. Our results for core model with filtered data from investment portfolios of Iranian banks reveal that the issuance of CBI policy has significantly decreased the transaction cost, and in turn, the market impact. The findings remain robust to the use of alternative proxy for the transaction cost like price return as the outcome of interest, and a wide range of scenarios for sample restrictions over time: 2 months, six months and rolling windows 4 through 15 days to depletion of policy effect.

Keywords: Market Impact; Event study; Dynamic Diff-in-Diff; PSM Model

JEL Classification: G14, D23 and C31

1. Introduction

Admati and Pfleiderer, 1988, recognize the information and liquidity as two key motivations for trade in financial markets. In economics, transaction cost includes the expenses which are paid by buyer but not received by seller, or vice versa. In finance, transaction cost is a premium above the current market price which is required to attract more sellers into the market, and/or is a negative margin to the current market price that is required to attract additional buyers into the market. In a pioneer paper by Ronald Coase (1937) transaction cost is described as an

*Corresponding author. Tel.: +98(21) 66049195 and 6 ext.149; E-mail: G.K.Haddad@sharif.edu

1 Associated professor at Economics Sharif University, and visiting scholar at Brandeis University, M.A. US, Gkhaddad@brandeis.edu

2 Institute for monetary research, Central Bank of Iran, E-mail: H_Heidari@gsme.sharif.edu

unavoidable cost of doing business. A hypothetical pyramid of transaction cost in financial markets includes nine components: broker commission fees, exchange fees, taxes, bid-ask spread, investment delay, price appreciation, market impact, timing risk, and opportunity cost. Due to the fact that some of its components are unobservable, the pyramid is described as an iceberg. Based on the evidence in financial markets, the market impact is considered as an important part of transaction cost which usually is unobservable. The market impact is defined as movements in stock price that is caused by a certain trade or order. As an ex-post outcome the market impact is not measurable and in the financial literature it is referred to as the "Holy Grail" of transaction cost.

Events in the financial markets affect sell and buy orders, and the price of stocks as well. Therefore the policy effect study is an appropriate methodology in financial markets to examine the price and volume sensitivity, and in turn the market impact to policy changes.

Various events may change the transaction cost. We categorize them in two groups: firm specific and economy wide events. Firm specific events include internal changes in firms such as, mergers and acquisitions, earnings announcements, issuance of new debt or equity, and announcements of macroeconomic variables, e.g., the trade deficit and interest rates. Economy wide events are the news or change in legal or regulatory environment which may impact the stock market.

Dolly (1933) is one of the seminal works that investigates effect of stocks split on the stock price. Dolly's evidence is used as a base to develop new tests in the event studies. Using Heckman's identification strategy for event study, Acharya (1988) finds that the decision of companies to call an outstanding convertible bond causes volatilities in the stock price. The study's empirical findings are consistent with the Harris and Raviv (1985) and Prabhala (1997). Nayak and Prabhala (2001) evaluate the individual effect of two, separated but simultaneously issued announcements, stocks split and dividend yield increase,

in terms of their observed announcement effects. In their working sample about 80% of investigated shares are affected by both of the firms' policy changes³.

Eckbo et al (1990) studied the effect of companies take over rumor on the market gains by using truncated regressions.

Hubbard and Palia (1995) find out a relationship between merger announcement effects and managerial ownership levels. Li and McNally (2006) applied the EMW method to open market share repurchases in Canada and find supporting evidence from a signaling interpretation of repurchase announcement effects.

Villalonga (2004) uses the propensity score matching (PSM) identification method to investigate the effect of portfolio diversification on firms' profit.

Their selected sample includes 167 firms in which the control group was the firms by non-diversified portfolio. The results clarifies that the change in the value of firms with diversified portfolio in treatment group is not significantly different from those of control group.

Most of the above outlined studies are using the price returns as outcome variable, but a new generation of event studies in stock markets have chosen the market impact as the outcome of interest, Almgren and Chriss (1997) and Kissell and Roberto Malamut (1998). While the second group are using a top-down cost allocation based methodology, Kissell and Malamut (1998) I-Star, model among others. In the former approach stock price movements are used as a measure of market impact, Madhavan (2000, 2002). The papers review transaction cost analysis in stock market from a microstructure perspective to develop a trade algorithm. There is a large body of studies in the literature which design a algorithmic framework for trading strategy in financial markets, for instance, Wagner (1991); Kissell and Glantz (2003); and Domowitz and Yegerman (2006, 2011) ; Gatheral (2010, 2012).

Reviewing key findings of the literature (by Loebb (1983); Holtausen, etal(1987); Chan and Lakonishok (1993); Plexus Group (2000)) reveals that the transaction

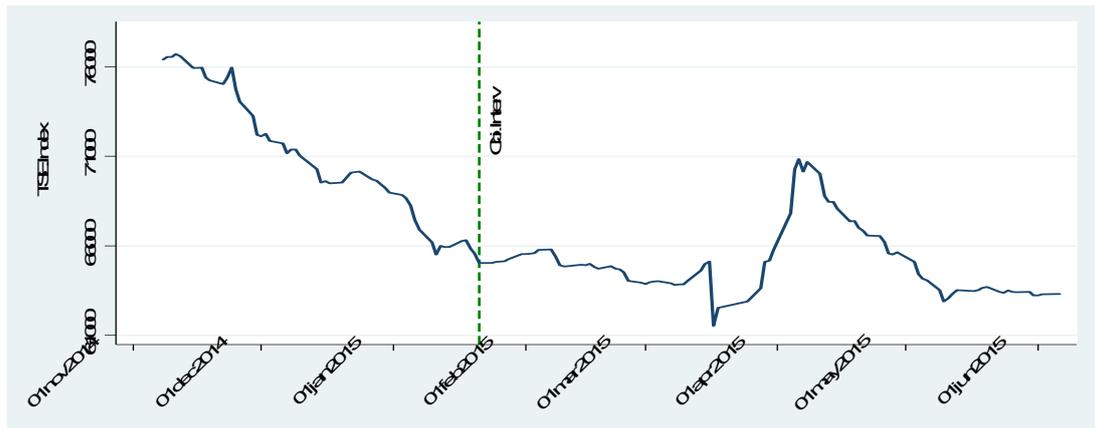
³ - They use a Bivariate Probit estimation technique for the two announcements. The residuals of the estimated models, which represent the private information components not included in the specification of the models, are significantly correlated. Implying that the estimation method selection is correct and the two events are not independent decisions.

costs are determined by market microstructure factors including: trade volume, total volume and buying or unsettled selling in the market. However, some empirical evidence provided by Stoll (1978); Amidhud and Mendelson (1980); Madhavan and Sofianos (1998); Chan and Lakonishok (1995); Keim and Madhavan (1997); and Breen et al (2002) show that transaction cost affected by market volatilities and total trading volume. Also, Beebower and Priest (1980); Wagner and Edwards (1993); Perold and Sirri (1993) and Breen et al (2002) argue that national wide economic and political events like macroeconomic announcement or change in regulation and supervision rules of the financial markets, may affect transaction cost.

Severe crunch with over 30% decrease in TSE index in 2012 resulted in a jump of transaction cost and market impact in TSE. To protect TSE further excess falls in the stock prices, CBI issued an ordinance in terms of which the commercial banks and their investment subsidiaries were permitted to buy more shares in the stock market. Figure (1) shows the decreasing trend in TSE index over a 6-month interval by June 2015. As is indicated in this graph after CBI ordinances in January 23, 2015 to encourage the commercial banks for making additional investments in TSE, it turned out to be relatively stable. However, implementation of the CBI ordinances was not mandatory for banks, but it encouraged the banks to actively step in and take long positions in the stock market.

In fact, the CBI ordinance appears to be a determinant event for major players in the financial markets, and basically is regarded as a tangible reduction in the strictness of the bank supervisory rules. Because, prior to the issuance of the ordinance a typical bank was permitted to spend at most 40% of its core capital in the TSE, while under the new regulatory rule in some cases banks' investment in the TSE amounted to over 50% of their core capital. Therefore, one can derive a preliminary conclusion that the new ordinance may have caused a flow of the banks' liquidity to the TSE.

Figure 1- TSE index before and after CBI ordinance intervention



Note: TIPEX trend in symmetric interval time, before and after issue the CBI ordinance . Vertical line is the issued date of CBI ordinance .

Based on the revealed news in public medias following the time of CBI's decision, the Ministry of Economic Affairs and Finance and the banking system agreed to invest about 54000 billion Rials into TSE which is a quite remarkable value when is compared with average of daily trading volumes (1900 billion Rials) in TSE. Also, alongside a tangible decrease in the average number of traders in TSE, the ratio of trading volume to number of dealers increased by 29% , up to 16830 shares per dealer. This indicates a rise in the number of Investment Companies as nondiscretionary trader, like banks or their subsidiaries in the market. Due to the fact that the major players in the TSE were nondiscretionary traders, the CBI intervention supplied additional liquidities for the traders. The liquidity in turn decreases market impact as a key component of transaction cost. This article addresses the question whether the regulatory policy of the Central Bank has reduced the market impact for those stocks which were already in investment portfolio of commercial banks.

With the documented background that we have presented so far, we outline the contributions of our paper as follows. First, while Fama et al (1969), Brown and Warner (1980 and 1985), Acharya (1988), Harris and Raviv (1985), Probhala (1997), Eckbo (2008), have used the daily price returns of shares as the outcome of interest to investigate effect of individual the events on the stock market, we apply 5- minutes intra-day micro structural data of TSE to reproduce the market impact. Second, whereas Schowert (1981), McQueen and Roley (1993), and Michel and Nitre (1995) have studied the effects of economic events and legal change for ownership, on the stock market, our center of interest is a supervisory policy effects study of the liquidity inflow on the market impact. Third, our estimation and inference methodologies are causal identification techniques,

dynamic Difference-in-differences and propensity score matching model which evaluate the policy effect.

Using the two identification strategies, we find statistically strong evidence which confirms that the CBI ordinance significantly reduced market impact after the announcement date. Significance of the effect remains robust for several samples in terms of the time ranges, rolling windows and the shares which are included in the banks' portfolio.

Rest of the paper continues as follows. In section 2 we document time profile of economic and political events which may affect stock market. Section 3 introduces market impact definition and its practical calculation methodology. Next in section 4 we present the institutional background of Iran's stock market and statistical properties of our working sample. Section 5 introduces the estimation methodology, Dynamic DiD and PSM for our framework. Furthermore a subsection of this part addresses robustness check results. Finally, section 6 summarizes and concludes.

1-2. Events and Banks Investment portfolio

Table (1) presents the major economic and political events in Iran business space which presumably affect the TSE. Among the five events we concentrate our period of study to the Central Bank ordinance intervention. In order to avoid the confounding and interaction effect of the other events we limit the time range of our working sample symmetrically to 3 months before and 3 months after the January 24, 2014, which is the day that Central Bank officially issued the ordinance.

Table (1)- Economic and political events in Iran

Events	Event date
Geneva agreement	Novmber 24,2014
Geneva agreement extension	Jun 24, 2014
Central Bank ordinance	January 24,2014
JCPOA ⁴	July 24,2014
JCPOA implimentation	January 24, 2014

Note: Political and economic events calendar. According to this schedule, time distance between Geneva agreement extension and Central Bank ordinance is about ????? and after????

Table 2 shows the bank's total investment portfolios in TSE. Total financial investment of the private owned banks in TSE have experienced a large positive growth of 35 percent between

⁴ Joint Comprehensive Plan of Action

2014 and 2015. If we include the **short and long terms investments of private banks in their portfolios**, then the positive change of investments will amount to 35%. In addition, this table confirms over the period, dispartate the reduction in total number of traders in TSE, the average of trading volume per number of traders has increased 29 percent. This evidence implies the existence of non-discretionary traders (the banks and financial companies which are owned by the banks) in the TSE.

Table 2- Bank's and bank holding companies' total portfolios invested in TSE (million Rials)

Investment value	Before event	After event	Growth
Private Banks	174,699,899	236,098,939	35.2
Bank holding Companies	296,165,685	375,872,414	27
Average of traders	62,606	55,867	-11
Average of volume to traders	13,032	16,830	29

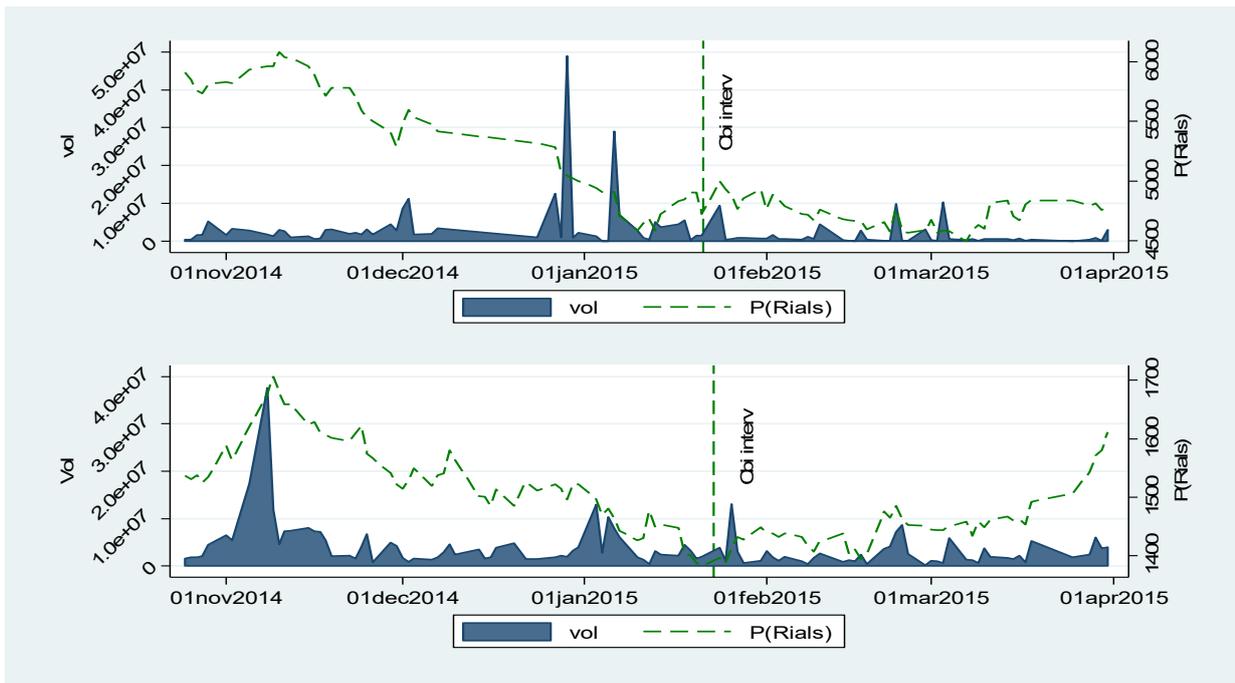
Source: The information have been extracted from published financial balance sheets in financial years 2014 and 2015.

Note: Value of Banks' portfolio in TSE. Calculations are based on internal Accounting Standard No. 15 and 18. Simple averages for the trading volume and number of traders are calculated by daily data.

Figure (2) shows the modified prices and daily trading volumes for shares of two large companies over Nov 01, 2014 to Apr 01, 2015. Share of these two companies in terms of Market Values is more than 24% of TSE in April, 2017⁵. Vertical dash line indicates the CBI intervention announcement day. The trends in the dual scale graph are showing more stability of the shares' price and larger traded volumes around and after the January 23, 2015. The large volumes of traded shares are important hinder for price crunch of shares.

Figure 2- Trading volume and price of "Fars" and "Akhaber" shares and CBI intervention time

⁵ - Modified prices are calculated when companies have capital increase or dividends distribution. An effective adjustment factor usually are calculated based on the amount of distributed dividends or capital increase by the TSE rules, www.old.tse.ir .



Note: Left vertical axis represents the daily traded volumes and right vertical axis indicates modified prices (Rials) for two large companies' shares in TSE. Upper panel is Iran Telecommunication Company (Akhabr) and the lower one is the holding of Persian Gulf Petrochemical Company (Fars). Reaction of these to shares to the CBI intervention is clearly observed in January 23, 2015.

Influence of time-series extraction on binge drinking interpretability using functional connectivity analysis

J.I. Padilla- Buriticá^{1,3}, J.D. Martínez-Vargas², J.M. Ferrández³, E.Pereda⁴, A.Correa⁴,
and G. Castellanos-Domínguez¹

¹ Signal Processing and Recognition Group, Universidad Nacional de Colombia

² Instituto Tecnológico Metropolitano, Medellin, Colombia

³ Diseño Electrónico y Técnicas de Tratamiento de Señal, Universidad Politécnica de
Cartagena, Cartagena, Spain

⁴ Universidad de la Laguna

jipadilla@unal.edu.co

Abstract. Brain connectivity analysis has recently gained considerable importance in different cognitive tasks and detection of pathological conditions. Despite the latest advances in connectivity analysis, several problems that directly influence its accuracy remain, being a proper extraction of the time-series to characterize the regions of interest (ROI) one of the challenges. In this work, we examine the influence of the time-varying mean estimation on the brain connectivity analysis for control and binge drinker subjects. The obtained results show that the performance of brain connectivity improves using the eigenvalue-based averaging since it may face better the nonstationarity behavior and inter-trial variability of MEG activity.

Keywords: Connectivity analysis, MEG Inverse problem, Lagged phase synchronization.

1 INTRODUCTION

Connectivity analysis has recently gained considerable importance in different cognitive tasks and detection of pathological conditions as a way to study the brain neural activity. This analysis can be devised either in the MEG-channel or source space. In the first case, the MEG electrodes only detect summed activities of a large number of neurons, which are affected by the field spread effects, tend to bias the estimated neural activity and therefore the connectivity analysis [11]. Moreover, it is difficult to associate a physiological meaning with estimated connections since the measured signals are not located in the same spatial proximity to the underlying sources. To overcome this issue, source reconstruction methods can take into account the cortical activity propagation across the scalp, increasing their performance on space and time domains. Besides, source solutions may support a better interpretation of the calculated interactions by their more direct association with the brain processes of integration and segregation [10].

Despite the latest advances in connectivity analysis, several problems that directly influence its accuracy remain, being one of the challenges a proper time-series ex-

traction to characterize the regions of interest (ROI) [9]. Mostly, the time-series extraction method focus on producing a single averaged time-course of MEG activity, relying on different estimates of time-varying means. However, each estimation is performed under diverse statistical assumptions, without taking into account the effects that can generate the method to obtain each time series [3]

In this work, we examine the influence of the time-varying mean estimation on the brain connectivity analysis for control and binge drinker subjects. The investigated averaging methods, which are commonly-used in literature, are the following: time-varying average and eigenvalue-based mean estimation. The comparison comprises three stages: i) The activity in the source space is estimated through Empirical Bayesian Beamformer (BMF) [1]. ii) Some regions of interest are selected, which in this case have been previously defined according to similar studies, taking into account the Brodmann's regions [12, 2]. iii) A connectivity brain measure is employed to quantify the changes in the information flow over the selected regions of interest. The obtained results show that the performance of brain connectivity improves using the eigenvalue-based averaging since it may face better the nonstationarity behavior and inter-trial variability of MEG activity.

2 METHODS

2.1 Estimation of brain source activity

With the aim of estimating a brain activity measured from M-EGG recordings, we will consider the distributed inverse solution $\mathbf{Y}=\mathbf{L}\mathbf{J}+\mathbf{\Xi}$, where $\mathbf{Y}\in\mathbb{R}^{C\times T}$ is the scalp M-EGG data measured by $C\in\mathbb{N}$ sensors at $T\in\mathbb{N}$ time samples, $\mathbf{J}\in\mathbb{R}^{D\times T}$ is the amplitude of $D\in\mathbb{N}$ current dipoles, which are placed in each three-dimensional dimension and distributed through cortical surface. Also, the lead field matrix $\mathbf{L}\in\mathbb{R}^{C\times D}$ holds the relationship between sources and M-MEG measurements, which can be assumed zero-mean Gaussian noise $\mathbf{\Xi}\in\mathbb{R}^{C\times T}$, having matrix covariance $\mathbf{Q}_{\mathbf{\Xi}}=\sigma_{\mathbf{\Xi}}^2\mathbf{I}_C$, where $\mathbf{I}_C\in\mathbb{R}^{C\times C}$ is an identity matrix, and $\sigma_{\mathbf{\Xi}}^2$ is the noise variance. Under these constraints, the measured brain source activity can be estimated as $\hat{\mathbf{J}}=\mathbf{Q}\mathbf{L}^T(\mathbf{Q}_{\mathbf{\Xi}}+\mathbf{L}\mathbf{Q}\mathbf{L}^T)^{-1}\mathbf{Y}$, being $\mathbf{Q}\in\mathbb{R}^{D\times D}$ the source covariance matrix.

As a rule, the source mapping approaches need spatial prior knowledge (priors) upon \mathbf{Q} to include information derived from multiple modalities and/or subjects. In this regard, Empirical Bayesian Beamformer (BMF) relies on a prior that assumes a covariance matrix with elements q_{dd} in the main diagonal as follows [7]:

$$q_{dd}=(\mathbf{l}_d^T(\mathbf{Y}\mathbf{Y}^T)\mathbf{l}_d)^{-1}/\delta_d, \quad \forall d=1,\dots,D,$$

where $\mathbf{l}_d\in\mathbb{R}^{C\times 1}$ is d -th column of \mathbf{L} , and $\delta_d=1/\mathbf{l}_d^T\mathbf{l}_d$ is a normalization parameter.

2.2 Time-series extraction from measured MEG data

To perform the functional connectivity analysis, two tasks must be performed previously: selection of regions of interest (ROI), and extraction of time series from the selected ROI set [6]. In this study, we use l tags (with $l\in\mathbb{N}^{ROI}$) to designate the regions

of interest associated with the Brodmann's areas Table 1; which have been chosen since they are affected by alcohol consumption [2]. We represent $\hat{\mathbf{J}}_l$ as the set of time series estimated for the dipoles labeled l . Further, all time series extracted from the dipoles, belonging to each Brodmann's area, are encoded into matrix $\hat{\mathbf{J}}_r = [\hat{\mathbf{J}}_1, \hat{\mathbf{J}}_2, \dots, \hat{\mathbf{J}}_l]$. Nonetheless, we carry out the clustering of active brain sources, which are spatially adjacent and correlated to the studied phenomena, estimating the time-courses that properly describe the temporal patterns emerged in each region. Hence, an additional step must be accomplished to provide a single time series, aiming to characterize each ROI as a whole. To this end, a couple of reduction approaches are widely used:

- Time-varying mean value averaged across all dipoles at each ROI, that is, $\mathbf{x}_r = \mathbb{E} \{ \hat{\mathbf{J}}_r : \forall i \in n \}$, being $\mathbb{E} \{ \cdot \}$ notation for expectation operator.
- Time-course of the eigenvectors associated with the maximal non-zero eigenvalues computed for the covariance matrix of $\hat{\mathbf{J}}_r$ at each ROI as explained in [8].

2.3 Measure of brain connectivity

The lagged phase synchronization (LPS), noted as $\phi \in \mathbb{R}^+$, measures the relationship between a couple of time series, \mathbf{x} and \mathbf{y} , as follows:

$$\phi = \sqrt{1 - |\mathcal{S}(\omega)| / |\Re(\mathcal{S}(\omega))|} \quad (1)$$

whith $\mathcal{S} \in \mathbb{C}$, the cross-spectral density matrix, defined as:

$$\mathcal{S}(\omega) = \begin{bmatrix} \mathcal{S}_{xx}(\omega) & \mathcal{S}_{xy}(\omega) \\ \mathcal{S}_{yx}(\omega) & \mathcal{S}_{yy}(\omega) \end{bmatrix}$$

Where the diagonal elements of $\mathcal{S}(\omega)$ reflect the power estimates of signals \mathbf{x} and \mathbf{y} , and the off-diagonal elements reflect the averaged crossspectral density terms. Notation \Re : stands for the real part of the matrix \mathcal{S} .

3 EXPERIMENTAL SET-UP

Fig. 1 shows the main schema for which we have studied the influence of the different forms of time series extraction, belonging to different ROI anatomically associated with the Brodmann's areas, using lagged phase synchronization (LPS) for connectivity analysis. We propose a methodology appraising the following stages: *i*) MEG brain activity mapping, *ii*) selection of regions of interest (ROI), *iii*) Extraction of time series belonging to the ROI set and *iv*) connectivity analysis.

3.1 MEG database and preprocessing

Four minutes of MEG signal were acquired (1000 Hz sampling rate and an online band-pass filter at 0.1 – 330 Hz) during eyes-closed resting state using a 306-channel (102 magnetometers and 204 gradiometers) system (Elekta[©], VectorView). In this study, only magnetometers (102 channels) information was submitted to source and statistical

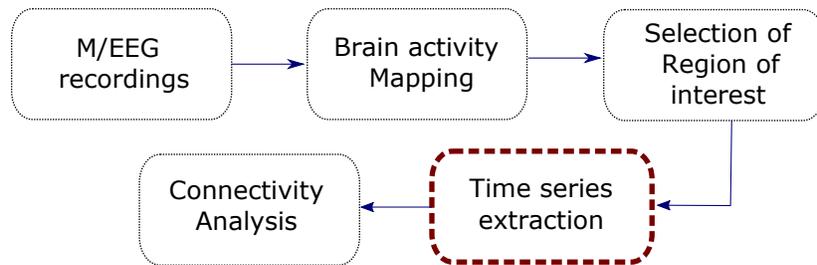


Fig. 1. Scheme of comparison between different methods for time series extraction in connectivity analysis. The block marked in dashed line is the subject of present study. The details of each step are described below.

analyses. The system was housed in a magnetically shielded room (VacuumSchmelze GmbH, Hanua, Germany). The head movement was monitored by means of four head-position indicator coils attached to the scalp. Ocular movements were tracked with two bipolar electrodes [4].

For the preprocessing, the raw recording data were at first submitted to Maxfilter software (v2.2, Elekta Neuromag) to remove external noise with the temporal extension of the signal space separation method with movement compensation. In this study, we used only magnetometers data in order to avoid mixing MEG sensors with different sensitivities or resorting to scaling. Accordingly, all of the magnetometer's resting state signals were automatically scanned for ocular, muscle and jump artifacts with Fieldtrip package (available online) and were visually confirmed by a MEG expert. The artifact-free data were segmented in continuous 4 seconds fragments (trials). At least 15 clean trials were obtained from all participants and preserved for further analyses. The number of surviving trials did not differ significantly between groups. To calculate the source reconstruction, the time series were filtered using a wavelet packet-based algorithm for the extraction of neural rhythms: θ : $4 - 7$ Hz; β : $14 - 30$ Hz and γ : $30 - 50$ Hz [4].

3.2 Brain activity mapping

A source reconstruction was obtained using 8196 dipoles distributed through the brain for each of the subjects. Each dipole corresponds to a likely source location (based on single dipole models) and the separation between them is 5mm. The leadfields are calculated using a single-sphere volume conductor model.

Concerning the choice of the prior set, \mathbf{Q} in Section 2.1 is a diagonal matrix formed from a direct projection of the data into the source space. BMF uses a single, global functional-anatomical prior (functional because it is based on assumptions about source covariance and anatomical because it is constrained to the cortical manifold) provides just one estimated covariance component at the sensor level. This method was selected because of its satisfactory spatial and temporal resolution in brain activity mapping for different levels of SNR [1].

3.3 ROI selection and time-series extraction

Each source on the brain grid is anatomically associated with a Brodmann's area. Thus, 12 areas are selected previously, which are shown in Table 1, giving a total of 24 ROIs [6].

Anatomical regions	Brodman area
middle temporal area (MT) ●	37
frontal eye fields (FEF) ●	6
superior parietal lobule (SPL) ●	7
anterior prefrontal cortex (aPFC) ●	10
Left dorsolateral prefrontal cortex (dlPFC) ●	9
Anterior cingulate cortex (aCC) ●	32
anterior inferior parietal lobule (aiPL) ●	40
anterior insula (aINS) ●	47
Posterior cingulate cortex (pCC) ●	23
posterior inferior parietal lobule (piPL) ●	39
Visual fields (Vis) ●	18
auditory fields (Aud) ●	41

Table 1. Selected Brodmann's areas for connectivity analysis.

For the selected ROI, two forms of time series extraction are used, which are based on the mean and the decomposition in singular values. In different studies of functional connectivity, it is necessary for the choice of representative time series in the region of interest (ROI). The usual approach is to compute the mean value across the first eigenvector of ROI dipoles, which is further employed to calculate the measure of brain connectivity [5].

3.4 Connectivity analysis

The significant differences between control subjects and binge subjects are shown in Figs. 2 and 3. We compared the results obtained with LPS (Eq. (1)), from the extracted time series with the average and the first eigenvector, in the complete set of regions of interest.

We obtain the connectivity results for the bands β and γ , afterwards we carried out a paired t-test, in which the null hypothesis is that there is greater connectivity in the control subjects than in the binge subjects ($\rho = 0.05$), and the results obtained are as follows:

In Figs. 2 and 3, the upper part shows the results obtained by means of the average and the lower part the results obtained with the first eigenvector. Also, the red color highlights the most representative connections for binge subjects that control subjects (with $\rho = 0.05$), the blue color shows the connections where the opposite happens.

There is a fronto-parietal coupling in the β rhythm as can be seen in Fig. 2, which is related to phase synchronizations during concentration and attention processes. It may

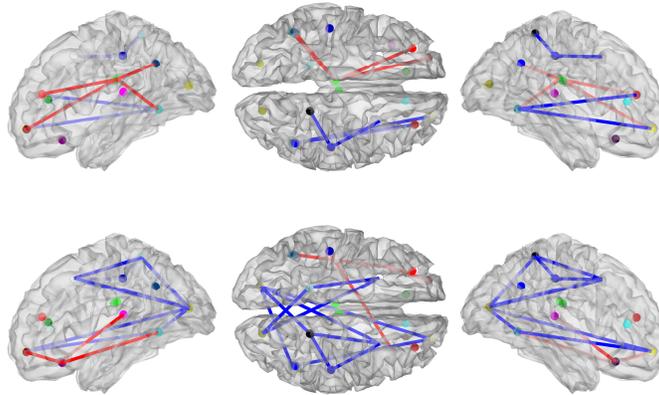


Fig. 2. Meaningful differences between binge (red line) and control (blue line) group in the connectivity analysis with the time series obtained with average (top) and the first eigenvalue (bottom) for β rhythm, with (LPS) and significant difference ($\rho = 0.05$)

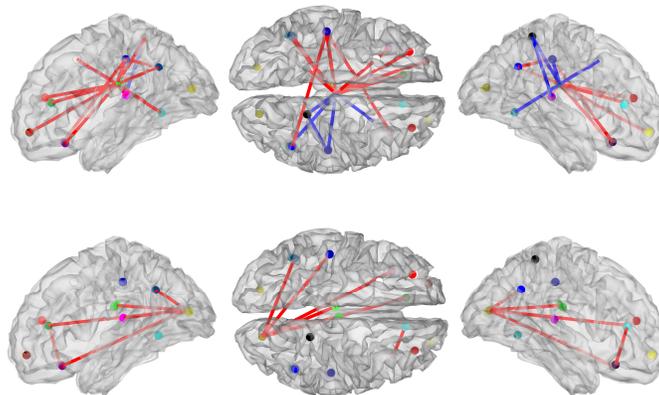


Fig. 3. Meaningful differences between binge (red line) and control (blue line) group in the connectivity analysis with the time series obtained with average (top) and the first eigenvalue (bottom) for γ rhythm, with (LPS) and significant difference ($\rho = 0.05$)

also be noted that greater activity is maintained when the analysis is performed by the first eigenvector, and the connections on the brain decrease for the band β when the analysis is performed by the average. This aspect is very important, since we are analyzing the same frequency band and the representative connections in the brain cortex are different.

In Fig. 3 it can be seen that there is a less number of connections than for the band β , in both control subjects and binge subjects, on the other hand, the representative

connections for the control largest eigenvalue). In addition, the connections over the parietal lobe decrease in the γ band. In general, for the rhythms β and γ it is possible to observe that the connections are different and change considerably depending on the method of time-series extraction, in spite of analyzing the same task.

4 DISCUSSION AND CONCLUDING REMARKS

We study the influence of the time series extraction method from different ROI on the brain connectivity analysis taking into account two methods for extracting the time series, based on the following steps:

1. Selection of brain mapping method, in this case we select BMF.
2. We apply two time series extraction from selected ROIs: we compute the average and the first eigenvector from predefined ROIs.
3. We estimate the connectivity measure using LPS between the extracted time-series.

The first task is related to the selected brain mapping method for imaging MEG activity. The tested method of brain mapping was BMF, the use of BMF improves identification of the source signals from MEG measurements [1], nevertheless, BMF tend to estimate several spurious activated areas, misleading the connectivity analysis. This effect appears to be directly related to the estimation complexity of the source covariance matrix [9].

Generally speaking, a challenging issue relating to brain connectivity analysis is how to extract the time series from the ROIs, since from these series, the analysis of functional connectivity is carried out, in which the different statistical relationships between the ROI are studied. Therefore, it is necessary to choose the method that best describes the dynamics of the time series, which are describing the behavior of all dipoles associated with an ROI. In the case of the average, there is a risk of eliminating components associated with phase change, while in the case of decomposition into eigenvectors, the assumptions are made about the probability distributions of the data, which leads to restrictions in the measure of connectivity.

Another aspect of consideration is the involved measure of connectivity analysis. Here, we have compared the changes of the LPS over the rhythms β and γ for the two methods for the time series extraction, it can be clearly noticed that the time series extraction method clearly influences the connectivity analysis. For instance, a poor time series extraction could lead to different physiological interpretations for the same task in the brain. As a result, the method for time series extraction must be chosen carefully in all studies conducted in brain connectivity.

As future work, the authors plan to test the introduced approach over diverse paradigms, clustering, and connectivity measures. Furthermore, an online extension of the brain connectivity analysis can be proposed to include the temporal variations of the inter-channel relationships directly.

ACKNOWLEDGEMENTS

This work was supported by the research project 111077757982 founded by COLCIENCIAS, and Programa Nacional de Becas de Doctorado, convocatoria 647(2014).

References

1. Belardinelli, P., Ortiz, E., Barnes, G., Noppeney, U., Preissl, H.: Source reconstruction accuracy of MEG and EEG Bayesian inversion approaches. *PloS one* **7**(12), 51985 (2012)
2. Carbia, C., Cadaveira, F., Lopez-Caneda, E., Caamaño-Isorna, F., Holguín, S.R., Corral, M.: Working memory over a six-year period in young binge drinkers. *Alcohol* **61**, 17–23 (2017)
3. Cho, J.H., Vorwerk, J., Wolters, C.H., Knösche, T.R.: Influence of the head model on EEG and MEG source connectivity analyses. *Neuroimage* **110**, 60–77 (2015)
4. Correas, A., Cuesta, P., López-Caneda, E., Holguín, S.R., García-Moreno, L., Pineda-Pardo, J., Cadaveira, F., Maestú, F.: Functional and structural brain connectivity of young binge drinkers: a follow-up study. *Scientific Reports* **6** (2016)
5. Gajdoš, M., Mračková, M., Elfmarková, N., Rektorová, I., Mikl, M.: 50. Comparison of canonical correlation analysis and pearson correlation in resting state fMRI in patients with parkinson's disease. *Clinical Neurophysiology* **126**(3), 47–48 (2015)
6. Hata, M., Kazui, H., Tanaka, T., Ishii, R., Canuet, L., Pascual-Marqui, R.D., Aoki, Y., Ikeda, S., Kanemoto, H., Yoshiyama, K., et al.: Functional connectivity assessed by resting state EEG correlates with cognitive decline of Alzheimer's disease—An eLORETA study. *Clinical Neurophysiology* **127**(2), 1269–1278 (2016)
7. Henson, R.N., Flandin, G., Friston, K.J., Mattout, J.: A parametric empirical bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Human brain mapping* **31**(10), 1512–1531 (2010)
8. Martinez-Vargas, J.D., Strobbe, G., Vonck, K., van Mierlo, P., Castellanos-Dominguez, G.: Improved localization of seizure onset zones using spatiotemporal constraints and time-varying source connectivity. *Frontiers in neuroscience* **11**, 156 (2017)
9. Padilla-Buriticá, J.I., Martinez-Vargas, J.D., Castellanos-Dominguez, G.: Emotion discrimination using spatially compact regions of interest extracted from imaging EEG activity. *Frontiers in computational neuroscience* **10** (2016)
10. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**(3), 1059–1069 (2010)
11. Schoffelen, J.M., Gross, J.: Source connectivity analysis with MEG and EEG. *Human brain mapping* **30**(6), 1857–1865 (2009)
12. Spear, L.P.: Effects of adolescent alcohol consumption on the brain and behaviour. *Nature Reviews Neuroscience* (2018)

MoCap multichannel time series representation and relevance analysis by kernel adaptive filtering and multikernel learning oriented to action recognition tasks

J. D. Pulgarin-Giraldo^{1,2}(✉), A. M. Alvarez-Meza³, S. Van Vaerenbergh⁴,
I. Santamaría⁴, and G. Castellanos-Dominguez²

¹ G-BIO Research Group, Universidad Autónoma de Occidente, Cali, Colombia.

² Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales, Colombia.

`jdpulgarin@unal.edu.co`

³ Faculty of Engineering, Universidad Tecnológica de Pereira, Pereira, Colombia.

⁴ Dept. of Communications Engineering, University of Cantabria, Santander, Spain.

Abstract. A framework based in kernel adaptive filters and multikernel learning for MoCap multichannel data is presented. In this sense, kernel adaptive filters are used to encode the dynamic of each channel. Then, a model for each time series is constructed with a codebook and latent functions estimated by KRLS tracker algorithm. These independent channel representations assemble similarity between multiple realizations in a RKHS thanks to Maximum Mean Discrepancy criterion. Later on, a kernel alignment algorithm is used to assemble multiple channels in a unique kernel that relates all realizations. Supervised classification over this kernel shows a good assembling for different actions realizations. Moreover, relevance estimated by the kernel alignment highlights the most significant channels in action realizations. Results show that our methodology easily constructs a good representation for MoCap multiple channel data, and results agree with the findings made in the biomechanical analysis for our kind of data: tennis stroke's records.

Keywords: Multichannel data, kernel adaptive filters, maximum mean discrepancy, center kernel alignment.

1 Introduction

Human action recognition from Motion Capture(MoCap) data is a well-established area in pattern recognition [1, 2]. To date, the main efforts are directed at creating a sufficiently robust dynamic model of human movement accomplished under a priori given actions. All of these models have been validated in predicting the movement and/or action recognition through a certain (di)similarity measure, often in a Reproducing Kernel Hilbert Space (RKHS) due to the nonlinear dynamics in biomechanic action generation. However, the models are mostly oriented to classify with a high accuracy the executed action rather than to observe

the relevance of each channel. Relevance studies in human action recognition are oriented to highlight the most useful features in feature extraction, but the body segments and articulations where the sensors are placed are not studied. For this reason, in addition to predict and classify, these techniques must define in some way the relevance of each channel in the task.

Kernel adaptive filters (KAFs) are nonlinear adaptive filters based on the framework of kernel methods [3]. They are frequently used in problems of nonlinear time series prediction. Besides, KAFs not only are a well-established methodology in time series prediction but also provide compact *dictionary* or *codebook*. This is an interesting property since at the end the stored elements will be the most representative of the time series for the current task, which is one-step ahead prediction. This property avoids segmentation stages to obtain a suitable data analysis and overpass well-established methods of feature extraction in human action recognition [4, 5].

On the other hand, Hilbert space embeddings are a recent trend in kernel methods that map distributions into infinite-dimensional feature spaces using kernels, such that comparisons and manipulations of these distributions can be performed using standard feature space operations like inner products or projections [6]. Moreover, an existing framework for analyzing distributions and comparing distributions called Maximum Mean Discrepancy (MMD) [7] allows us to evaluate expectations over functions in the unit ball of a RKHS. However, multiple probability distributions analysis or even joint probability distributions analysis based on Hilbert space embeddings are not explored.

Regarding the combination of multiple dynamic models by kernel methods, the most widely studied have been those built by convex combinations of a finite set of base kernels [8, 9, 1]. In classification and regression tasks, the classical uniform combination solution has been improved thanks to the Centered Kernel Alignment (CKA) algorithm proposed by Cortes [10]. This one uses a similarity measure between kernels or kernels matrices to measure the similarity of each base kernel with a target kernel obtained from output labels. CKA is efficient and easy to implement and, additionally, the weights of the combinations provide the relevance of each base kernel. If the base kernels are constructed independently from each channel in a MoCap multichannel time series, these weights somehow reveal the most important channels involved in an action execution.

Here, a methodology for MoCap multichannel data representation is presented without data segmentation. This methodology is oriented to assemble appropriately all channels in a Reproducing Kernel Hilbert Space. This embedding allows aligning kernels with a target kernel with label classes in supervised classification. This kernel alignment procedure not only provides good classification accuracy but also reveals the most significant channels in the action execution associated with the kernel target.

2 Theoretical framework

We assume a scenario in which a set of J time series $\mathbf{x}_j[t]$ are obtained from sensor measurements, with $j = 1, \dots, J$. For each time series, T time steps are available, i.e. $t = 1, \dots, T$. We collect the entire set of measurements in the matrix $\mathbf{X} \in \mathbb{R}^{J \times T}$, which contains the J time series as its rows,

$$\mathbf{X} = \begin{bmatrix} x_1[1] & x_1[2] & \dots & x_1[T] \\ x_2[1] & x_2[2] & \dots & x_2[T] \\ \vdots & \vdots & \ddots & \vdots \\ x_J[1] & x_J[2] & \dots & x_J[T] \end{bmatrix} \quad (1)$$

We further assume that multiple such sets are available. The n -th set is represented as \mathbf{X}^n , with $n = 1, \dots, N$, and to indicate that a time series belongs to a particular set n we use the notation $\mathbf{x}_j^n[t]$.

Our goal is to develop a similarity measure between different such multichannel time series and to perform different types of analyses with this measure. To this end, we will first represent each individual time series $\mathbf{x}_j^n[t]$ as a compact model \mathcal{M}_j^n , and then we will define a similarity measure between sets of these models.

2.1 Dynamical channel model encoded by kernel adaptive filtering

With the aim of properly modeling each individual time series $\mathbf{x}_j^n[t]$, we will represent its dynamic behavior through kernel adaptive filters (KAFs) [3]. For the sake of clarity, in the following we will omit the superscript n until section Section 2.3.

In kernel methods, the Representer Theorem allows us to express the nonlinearities of a wide range of problems as a kernel expansion in terms of the training data

$$f(\mathbf{x}_j) = \sum_{m=1}^M \alpha_m \kappa(\mathbf{x}_j[m], \mathbf{x}_j), \quad (2)$$

where $\mathbf{x}_j[m]$ represents the m -th training input, and $\kappa(\cdot, \cdot)$ is a positive semidefinite *kernel* function. When the desired outputs $y_j[m]$ are available for the training data, these can be used to estimate the optimal expansion coefficients α_m . Given a time series $\mathbf{x}_j[t]$, for $t = 1, \dots, T$, the problem of one-step ahead prediction can be formulated as a regression problem with input-output pairs $\{\mathbf{x}_j[t], y_j[t]\}$ in which the input data is taken as the time-embedded version of the series with L lags, $\mathbf{x}_j[t] = [x_j[t], x_j[t-1], \dots, x_j[t-L+1]]$, and the desired output is the next sample, $y_j[t] = x_j[t+1]$.

We will employ kernel adaptive filtering techniques to solve the described regression problem. These algorithms estimate a model of the form (2) by minimizing the least-squares error between the true labels $y_j[m]$ and their predictions $f(\mathbf{x}_j[m])$. Furthermore, they do so in an online fashion, i.e. by performing one or more passes over the data, which is preferred if not all data fits in memory.

KAF algorithms limit the amount of stored training data $\mathbf{x}_j[m]$ by constructing a compact *dictionary* or *codebook*.

Among KAF algorithms, we are interested in those that only store a fixed amount R of training data in the codebook, which is referred to as the *budget*. In particular, we will use the Kernel Recursive Least-Squares Tracker (KRLST) algorithm [11], which represents the state of the art in kernel adaptive filtering. KRLST allows to maintain a fixed budget during operation, and it obtains high accuracy in a wide range of regression and prediction tasks. The prediction of KRLST represents an estimate of the latent noiseless time series, which is unobserved. KRLST is based on a probabilistic input-output model, which provides some additional advantages compared to other KAFs. A Matlab implementation of KRLST and other KAF algorithms is available in [3].

The steps followed to obtain the dynamical channel model \mathcal{M}_j for a time series $\mathbf{x}_j[t]$ can be summarized as follows:

1. Construct input-output pairs $\{\mathbf{x}_j[t], y_j[t]\}$ with time-embedding L for the inputs.
2. Run a training pass over all $\{\mathbf{x}_j[t], y_j[t]\}$ using KRLST.

The final model for each time series consists in the codebook data selected by KRLST out of the observed inputs, which we will refer to as the *centers* or *centroids* $\mathbf{c}_j[r]$, and their corresponding estimated latent function outputs, or *desired values* $d_j[r]$:

$$\mathcal{M}_j = \{\mathbf{c}_j[r], d_j[r]\}, \quad r = 1, \dots, R. \quad (3)$$

2.2 Similarity measure between models

Let us consider two different models $\mathcal{P} = \{\mathbf{p}_r\}_{r=1}^P$ and $\mathcal{Q} = \{\mathbf{q}_r\}_{r=1}^Q$. Each model represents the dynamic behavior of a given univariate time series and it is composed of a sequence of ordered pairs of codebook elements and output latent functions obtained by the KRLST algorithm; that is, $\mathbf{p}_r = (\mathbf{c}_p[r], d_p[r])$ and $\mathbf{q}_r = (\mathbf{c}_q[r], d_q[r])$. For generality purposes, we assume in this section that each model can have a different number of elements (complexity).

The elements of each model or model samples, as given by the KRLST, are not ordered; therefore, any permutation or reordering of the elements represents the same model. Bearing this in mind, we interpret each model as a cluster of points in the input space. We now define a mapping from the set of models \mathcal{Z} to a RKHS \mathcal{H} as follows

$$\begin{aligned} \Phi: \mathcal{Z} &\longrightarrow \mathcal{H}, \\ \{\mathbf{p}_r\}_{r=1}^P &\longmapsto \{\Phi(\mathbf{p}_r)\}_{r=1}^P \end{aligned}$$

which maps each model in the input space to a model in the feature space. A model can be interpreted as a distribution function from which P realizations are available. Then, to define a distance between models we resort to the Maximum Mean Discrepancy (MMD) defined by Gretton in [7]. Given two models \mathcal{P} and

\mathcal{Q} , the MMD criterion computes the distance between them, $\mathfrak{d}^2(\mathcal{P}, \mathcal{Q})$, as the squared Euclidean distance between the sample means of the two distributions, i.e.,

$$\begin{aligned} \mathfrak{d}^2(\mathcal{P}, \mathcal{Q}) &= \left\| \frac{1}{P} \sum_{r=1}^P \Phi(\mathbf{p}_r) - \frac{1}{Q} \sum_{r=1}^Q \Phi(\mathbf{q}_r) \right\|_2^2 \\ &= \frac{1}{P^2} \sum_{r=1}^P \sum_{r'=1}^P \kappa(\mathbf{p}_r, \mathbf{p}_{r'}) + \frac{1}{Q^2} \sum_{r=1}^Q \sum_{r'=1}^Q \kappa(\mathbf{q}_r, \mathbf{q}_{r'}) + \\ &\quad - \frac{2}{PQ} \sum_{r=1}^P \sum_{r'=1}^Q \kappa(\mathbf{p}_r, \mathbf{q}_{r'}), \end{aligned} \quad (4)$$

where $\kappa(\mathbf{p}_r, \mathbf{p}_{r'})$ is the kernel function between two model samples.

Without loss of generality and to simplify notation, let us denote two model samples as $\mathbf{p} = (\mathbf{c}_p, d_p)$ and $\mathbf{q} = (\mathbf{c}_q, d_q)$. Assuming a separable model that decouples the influence of the input and the output [12], the kernel function between two model samples is,

$$\kappa(\mathbf{p}, \mathbf{q}) = \kappa(\mathbf{c}_p, \mathbf{c}_q) \kappa(d_p, d_q).$$

Assuming a linear kernel for the output and the usual Gaussian kernel for the input, the proposed input-output kernel is finally defined as

$$\kappa(\mathbf{p}, \mathbf{q}) = \exp\left(-\frac{\|\mathbf{c}_p - \mathbf{c}_q\|^2}{2\sigma_c^2}\right) d_p d_q. \quad (5)$$

Using this separable kernel, the distance between models in Eq. (4) can be rewritten more compactly in terms of kernel matrices as

$$\mathfrak{d}^2(\mathcal{P}, \mathcal{Q}) = \frac{1}{P^2} (\mathbf{d}_p^T \mathbf{K}_{pp} \mathbf{d}_p) + \frac{1}{Q^2} (\mathbf{d}_q^T \mathbf{K}_{qq} \mathbf{d}_q) - \frac{2}{PQ} (\mathbf{d}_p^T \mathbf{K}_{pq} \mathbf{d}_q), \quad (6)$$

where $\mathbf{K}_{pq}(r, r') = \exp(-\|\mathbf{c}_p[r] - \mathbf{c}_q[r']\|^2 / 2\sigma_c^2)$.

2.3 Multikernel learning for relevance assessment

In the previous subsection we have introduced a similarity measure between pairs of models each one extracted from a different real-valued time series. Here we extend the procedure to deal with multichannel time series.

Assume that we have a labeled set of J -dimensional time series that we denote as $\mathbf{X}^n \in \mathbb{R}^{J \times T}$, $n = 1, \dots, N$. The j -th row of \mathbf{X}^n is a real-valued time series acquired by the j -th sensor. We first extract a model for each channel using the KRLST algorithm. Then, for the n -th multichannel time series we have a collection of J models that we denote as $\{\mathcal{M}_j[n]\}_{j=1}^J$. With some abuse of notation, let us denote as \mathbf{K}_j the $N \times N$ kernel matrix that measures the

(di)similarities for the j -th channel between the N time series in the training data set. The element (n, m) of this kernel matrix is given by

$$\mathbf{K}_j(n, m) = \exp - \left(\frac{\mathfrak{d}^2(\mathcal{M}_j[n], \mathcal{M}_j[m])}{2\sigma_0^2} \right), \quad (7)$$

where $\mathfrak{d}^2(\mathcal{M}_j[n], \mathcal{M}_j[m])$ is the pairwise distance between models described in Section 2.2 (Eq. (6)).

To combine the information from the J channels we propose to use a multi-kernel constructed as follows

$$\hat{\mathbf{K}} = \sum_{j=1}^J \alpha_j \mathbf{K}_j, \quad (8)$$

where the weights α_j $j = 1, \dots, J$ are yet to be determined. A simple solution would be to choose $\alpha_j = \frac{1}{J}$, but this would overlook differences in relevance or discriminative power between the channels. To find more informative weights that allow us to quantify the relevance of individual channels, we propose to use a centered kernel alignment procedure [10]. The basic idea is to find the optimal α_j^* maximizing the alignment between the multikernel matrix \mathbf{K} and the target kernel matrix $\mathbf{K}_l = \kappa(\mathbf{l}, \mathbf{l}') = \mathbf{U}^T$, which is calculated from the known label classes $\mathbf{l} = \{l[i]\}_{i=1}^N$. For a given set of weights α_j , the centered correlation or alignment between matrix kernels \mathbf{K} and \mathbf{K}_l is given by

$$\rho(\mathbf{K}, \mathbf{K}_l; \alpha) = \frac{\langle \mathbf{H}\mathbf{K}\mathbf{H}, \mathbf{H}\mathbf{K}_l\mathbf{H} \rangle}{\|\mathbf{H}\mathbf{K}\mathbf{H}\|_F \|\mathbf{H}\mathbf{K}_l\mathbf{H}\|_F}, \quad \rho \in [0, 1] \quad (9)$$

where $\mathbf{H} = \mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^T$ is a centering matrix, $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, $\mathbf{1} \in \mathbb{R}^N$ is an all-ones vector, and notations $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_F$ stand for the inner product and the Frobenius norm, respectively.

Then, the optimal relevance weights are $\alpha^* = \operatorname{argmax} \rho(\mathbf{K}, \mathbf{K}_l, \alpha)$ subject to the constraint $\|\alpha^*\| = 1$. This problem is solved by the Centered Kernel Alignment (CKA) algorithm [10].

3 Experimental setup

3.1 Database description

The data were collected from 17 high-performance tennis players of the Caldas-Colombia tennis league. The players had the following anthropometric parameters: age 18.9 ± 2.7 , mass 64 ± 14.9 kg, height 168.8 ± 8.4 cm and all players are right-handed. The employed motion capture protocol was Biovision Hierarchy (BVH) with a full body skeleton of 23 channels. Optitrack Flex V100 (100 Hz) infrared videography was collected from six cameras to acquire sagittal, frontal, and lateral planes and skeleton and multichannel time series were estimated in Optitrack Arena[®]. All subjects were encouraged to hit the ball with the same velocity and action just as they would in a match. They were instructed to hit one series continuously by 30 seconds of each indicated stroke. The strokes indicated in each record were: forehand, backhand, volley, and backhand volley.

3.2 MoCap data

Let $U \in \mathbb{R}^{T \times (J \times D)}$ be a multi-channel input matrix that holds T frames and $J \times D$ channels. Each $U_i = \{\mathbf{u}_{ij} \in \mathbb{R}^D : j \in J\}$ gathers the skeletal posture at the i -th frame with J D -dimensional body-joints (see Fig. 1(a)). Meanwhile, each $U_j = \{\mathbf{u}_{ij} \in \mathbb{R}^D : i \in T\}$ assembles time behavior of D -dimensional body-joint j (see Fig. 1(b)).

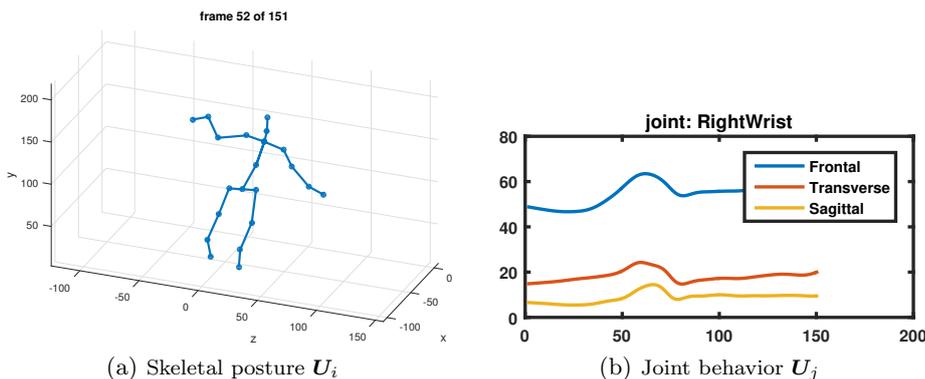


Fig. 1. MoCap data. Tennis serve example

3.3 PCA Preprocessing

Initially, all channels are centered respect to the limb center. Then, to describe the time behavior of the j -th body-joint from U_j , we perform a dimensional reduction stage from $\mathbb{R}^D \rightarrow \mathbb{R}$ to obtain a compact representation of its time behavior. In this case, from the covariance matrix $W \in \mathbb{R}^{D \times D}$ we consider only the first principal component \mathbf{w}_1 , obtained from the first eigenvector of the covariance matrix. Then, we obtain the linear projection $\mathbf{x}_j = U_j \mathbf{w}_1$, where $\mathbf{w}_1 \in \mathbb{R}^{D \times 1}$ (see Fig. 2).

3.4 Channels dynamic models estimated by KRLS tracker

Before the encoding, each \mathbf{x}_j is decimated by a factor of 5 in order to reduce the computational complexity of the KAF. We compute each model \mathcal{M}_j with KRLST parameters set as follows: forgetting factor 1, time embedding $L = 6$, codebook size $R = 50$, noise to signal ratio $\lambda = 10^{-6}$, a Gaussian kernel with σ calculated as the median value of channel \mathbf{x}_j and the initial codebooks are built directly from the input time series $\mathbf{x}_j \in \mathbb{R}^{T \times 1}$. Each model is validated doing a simple task: predict $x(t+1)$ from data available up to time t .

Fig. 3 shows the mean prediction error in each channel j for all sets of multichannel data, in this case, $N=68$. Although the number of outliers looks high,

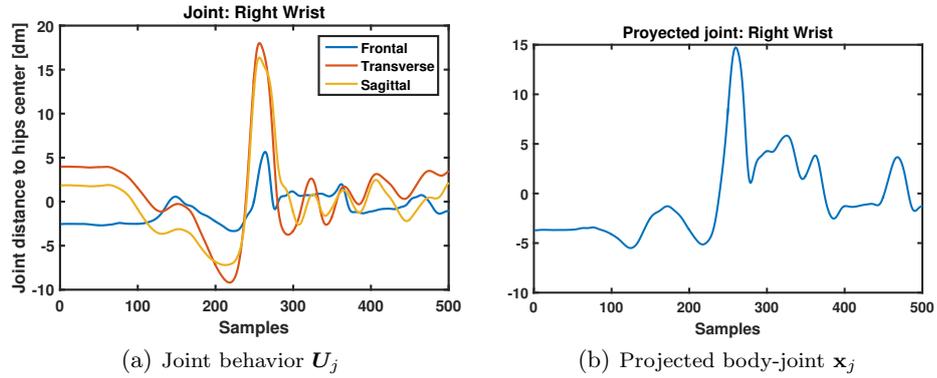


Fig. 2. Body-joint time behavior projected $\mathbb{R}^{T \times 3} \rightarrow \mathbb{R}^{T \times 1}$. Example on Right wrist joint in a serve record

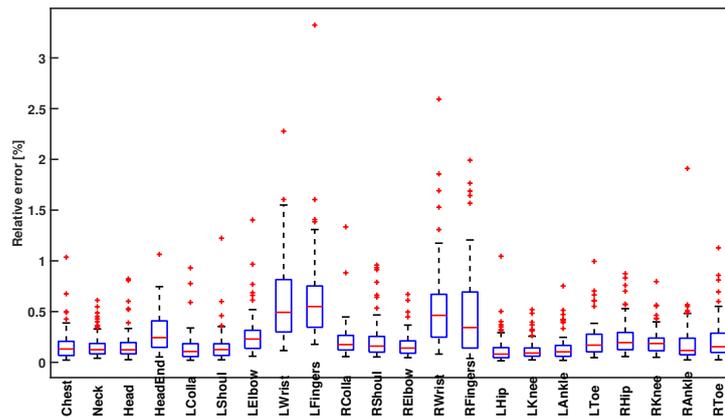


Fig. 3. Relative error results for each joint model \mathcal{M}_j^n estimated over N records with four different classes

it shows a low and regular mean error, which is significant due to the high variability of both: inter-subject and inter-class variability. Besides, our approach works with the 30 seconds full-long one take videos where several and continuous actions were recorded. There are approximately 12 to 16 strokes in each individual record. It is worth saying that segmentation and selection of actions are not required in our modeling process.

3.5 Similarity between models

Our proposed functional \mathfrak{D}^2 allows us to construct a kernel similarity measure $\kappa(\mathcal{M}_j[n], \mathcal{M}_j[m])$ which highlights each group of actions without previous information about the classes. In Fig. 4(a) we can see the block diagonal structure of the Gram matrix \mathbf{K} constructed over records of the right wrist joint. In fact, KPCA 2D-embedding in Fig. 4(b) shows the separability between groups of records that are colored according to its true label.

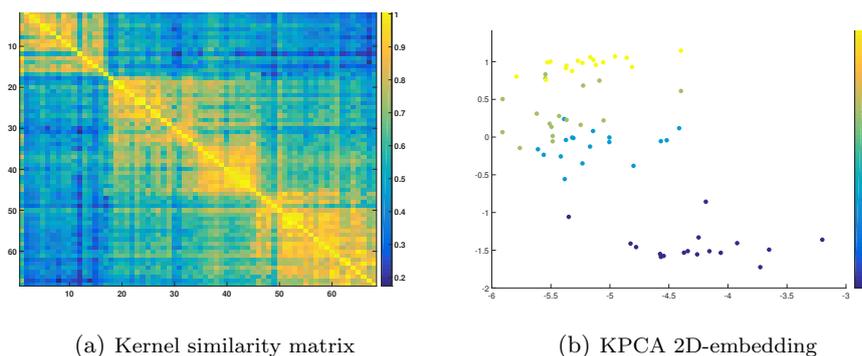


Fig. 4. Model similarity comparison for right wrist body-joint over 68 records. In both plots, the four classes of 17 strokes records are distinguishable

4 Relevance and classification results

The goal of multikernel learning described in Section 2.3 was to develop a supervised classifier for multichannel time series which, at the same time, allows us to assess the relevance of each channel for the given classification problem. Once the multikernel $\hat{\mathbf{K}}$ is constructed it allows compare multichannel data, so that we can apply any kernel-based classifier. In this work, we use a kernel nearest neighbor (KNN). The KNN classifier finds the k samples in the training dataset closest to test data (with maximum similarity) and carries out majority vote. Classification performance and relevance are computed using a cross-validation validation scheme.

Fig. 5 shows the attained α values into boxplot depending on the channel relevance approach. Particularly, the body joints at the end of the limbs are the most relevant. These channels highlight the difference between the four classes of action executed. Nonetheless, the variability observed in the most relevant channels implies a strong dependency in the execution, namely, the angle of the racket in the hit moment varies with the wrist and fingers channels relation.

Regarding to the classification results, as can be seen in Fig. 6(a), accuracies over 90% are attained for almost provided nearest neighbors. In Fig. 6(b), the

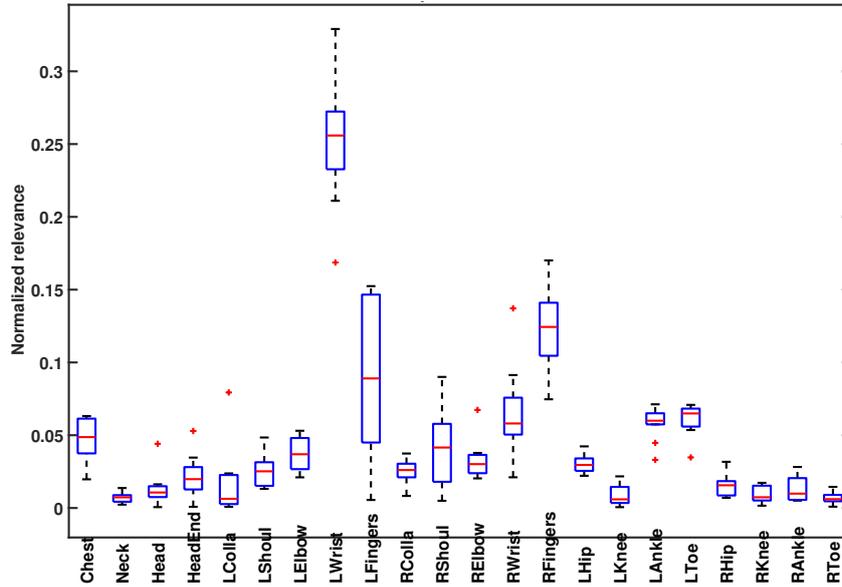
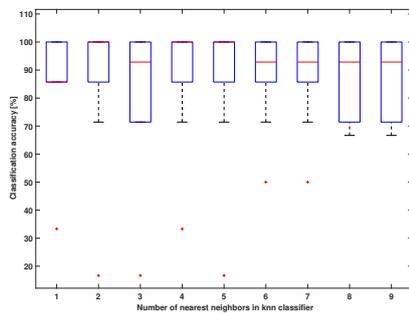


Fig. 5. Relevance body joint analysis in four activities. 10 folds in cross-validation were used over 68 records



	Forehand	Backhand	Volley	BH volley
Forehand	100	0	0	0
Backhand	0	85	10	5
Volley	5	0	75	20
BH volley	0	0	10	90

(a) Classification performance versus number of nearest neighbors in KNN classifier (b) Confussion matrix with three nearest neighbors. Accuracy results in %

Fig. 6. Classification results in four activities. 10 folds in cross-validation were used over 68 records

lowest results must be analyzed in confrontation with the action, where backhand presents low ball speeds after the impact and it were closer to speeds obtained in volley strokes executions. Nevertheless, each record classified contains 12 to 16 continuously stroke executions without segmentation, so the confused actions depend of execution’s speed after 30 seconds.

4.1 Discussion and concluding remarks

The proposed framework for MoCap multichannel analysis presents a methodology that first: obtains an appropriate and individual representation of the dynamic of each channel; and second: this channel representation based on KAFs allows us to combine similarity between several realizations. In fact, this framework easily matches with a multikernel algorithm as CKA, which merges multiple channels into just one kernel that can be used in classification tasks. It can be seen that CKA reveals the most significant channels in a set of actions, and these results are congruent with biomechanic theory in tennis actions execution [13].

This framework should be expanded to analyze the ideal optimal number and placement of sensors in human action recognition tasks, no matters its source; optical markers, inertial sensors or depth cameras. Besides, human motion action involves an interaction between all body segments: every action has a biomechanical chain that produces it, so relevance of channels must give information about the most relevant body segments involved across the time. The results encourage us to develop an algorithm for biomechanical chain generation without kinetic information, just from skeleton representations of actions.

As future work, this framework must be validated in larger action datasets, as well as must be evaluated in assessment motor disorders in order that relevance shows alterations in specific body segments or articulations.

Acknowledgments. This work is supported by the project 111077757982 and the program “Doctorados Nacionales 2014” number 647 funded by COLCIENCIAS, by “Convocatoria nacional para el apoyo a la movilidad internacional 2017-2018” funded by Universidad Nacional de Colombia, as well as PhD financial support from Universidad Autónoma de Occidente.

Bibliography

- [1] Salah Althloothi, Mohammad H. Mahoor, Xiao Zhang, and Richard M. Voyles. Human activity recognition using multi-features and multiple kernel learning. *Pattern Recognition*, 47(5):1800 – 1812, 2014.
- [2] F. Ofi, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *J. Visual Communication and Image Representation*, 25(1):24–38, 2014.
- [3] Steven Van Vaerenbergh and Ignacio Santamaría. A comparative study of kernel adaptive filtering algorithms. In *2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, pages 181–186, August 2013. Software available at <https://github.com/steven2358/kafbox/>.
- [4] S. García-Vega, A. M. Álvarez-Meza, and G. Castellanos-Dominguez. *Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings*, chapter

- Time-Series Prediction Based on Kernel Adaptive Filtering with Cyclostationary Codebooks, pages 354–361. Springer International Publishing, Cham, 2015.
- [5] J. D. Pulgarin-Giraldo, A. M. Alvarez-Meza, L. G. Melo-Betancourt, S. Ramos-Bermudez, and G. Castellanos-Dominguez. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - CIARP 2016, Proceedings*, chapter A Similarity Indicator for Differentiating Kinematic Performance Between Qualified Tennis Players, pages 309–317. Springer International Publishing, 2016.
- [6] Edgar A. Valencia and Mauricio A. Álvarez. Short-term time series prediction using hilbert space embeddings of autoregressive processes. *Neurocomputing*, 266:595 – 605, 2017.
- [7] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012.
- [8] Donghui Wu, Zhelong Wang, Ye Chen, and Hongyu Zhao. Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing*, 190:35 – 49, 2016.
- [9] Fabio Aioli and Michele Donini. EasyMKL: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215 – 224, 2015.
- [10] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.*, 13(1):795–828, March 2012.
- [11] Steven Van Vaerenbergh, Miguel Lazaro-Gredilla, and Ignacio Santamaria. Kernel recursive least-squares tracker for time-varying regression. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1313–1326, Aug 2012.
- [12] Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, March 2012.
- [13] J. Landlinger, S. Lindinger, T. Stoggl, H. Wagner, and E. Muller. Key factors and timing patterns in the tennis forehand of different skill levels. *Journal of sports science & medicine*, 9:643–651, 2010.

Forecast Model for Current, Wave and Wind Climate at the Danish Test Site for Wave Energy, DanWEC

Amélie Tetu^{1*}, Jens Peter Kofoed¹, Flemming Schlütter², and Poul Hammer²

¹ Department of Civil Engineering, Aalborg University,
Thomas Manns Vej 23, 9220 Aalborg Ø, Denmark

² DHI, Agern Allé 5, DK-2970 Hørsholm, Denmark

Abstract. This paper presents the forecast model developed for the Danish test site for wave energy. It includes results from the project *Resource Assessment, Forecasts and wave energy converters' operation and maintenance (O&M) strategies at DanWEC and beyond* which has been initiated to deliver detailed information on the environmental conditions at DanWEC and to review implementation of O&M procedures, which will ultimately improve wave energy converters' operation and reduce their costs. The forecast model is based on an hindcast wave model which has been validated against data from wave measuring buoys and it provides six days wave and current prognostic for the test site. The forecast model is coupled with an O&M tool that helps planning operation and reduce costs for users and operators of the test site.

Keywords: Wave energy, Wave climate, Resource assessment, Wave energy converters, Forecast model, Spectral wave model, Weather windows, Operations & Maintenance planning, Probability assessment, Forecasting service

1 Introduction

In the process of setting up the Danish test site for wave energy Aalborg University and DHI have collaborated with DanWEC [1] to provide dedicated descriptions of the wave climate at the site by hindcast wave modelling and setting up a forecast system.

The detailed assessment of the wave climate both in terms of wave energy resource and extreme conditions are imperative for developers in order to be able to accurately design the wave energy converters (WECs) to the test site. The characterization of the wave climate was done based on 35 years hindcast modelling applying DHI's model MIKE 21 SW [2] and is available in [3].

The dedicated MIKE 21 wave model was also applied as the basis for setting up a forecast service. The forecast updates a range of parameters related to the wave conditions, wind conditions and current speed throughout the modelling

* Correspondence: at@civil.aau.dk; Tel.: +45 9940 2924

area twice every 24 hours [4]. The forecast model provide a 5 days prognostic of the conditions at the test site and the model forcing comprises input from regional DHI models and wind fields.

The forecast is a valuable tool to plan operations at the test site. An O&M decision support system is being build as assistance to the operators.

This paper will first give an introduction to the test site and its sensor network. The hindcast model which is used as base for the forecast model will be introduced together with validation of the model against data from wave measuring buoys. The forecast system and the ability of the forecast to accurately predict the wave and current conditions will then be presented. A brief introduction to the O&M system will be given. The tool will be able to assess the risk of downtime by probability, in relation to operation types and the criteria set up for each type of operation. The catalogue of operations types, vessel types and associated criteria is developed based on DanWEC experience.

2 Description of the DanWEC test site

The DanWEC test site is situated on the North-West coast of the Danish peninsula Jutland, at Hanstholm, facing the Danish part of the North Sea. The data acquisition network of the test site comprises three buoys, as shown in Fig. 1. It consists of one Datawell Mark II non directional buoy, placed outside Hantsholm's harbour, and two Datawell DWR4 directional buoys including current measurements. The non-directional buoy *Buoy I* has been installed in 1998 and has provided almost 20 years of data [5]. Before 1998 a similar older version of a wave rider buoy was placed outside the harbour and paper records of wave data over the period 1979–1988 was analysed in relation to the first Wave Power experiments by Danish Wave Power Aps in 1989 [5].

Table 1. DanWEC wave measurement instrumentation

	Coordinates (Lat [°], Lon [°])	Water depth [m]	Model
Buoy I	(57.1315, 8.5821)	17.5	Datawell Mark II non directional
Buoy II	(57.1112, 8.5457)	14.5	Datawell DRW4 directional wave and current
Buoy III	(57.1171, 8.5173)	24.6	Datawell DRW4 directional wave and current

The two DanWEC directional buoys were installed in March 2015 and have been providing new information on the wave climate at this location, including insight on the directionality of the waves, the wave spectra and current characteristics. The two directional buoys are situated at a distance of approximately 3 km from the shore and are equipped with accelerometers providing displacements over time after proper filtering and double integration. The accelerometer measuring the vertical displacement is placed on a gravity-stabilized platform, decoupling the movement of the buoy from the measurement of the wave

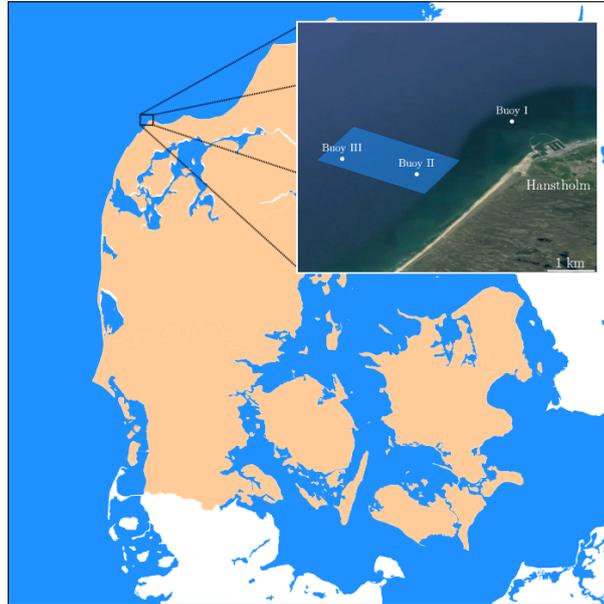


Fig. 1. DanWEC network sensor situated on the north-west coast of mainland Denmark.

height through vertical acceleration. The directional buoys are also equipped with three acoustic current transducers placed 120° laterally apart. They measure the Doppler shift of reflected 2 MHz pings at roughly 1 m water depth. All directions are measured relative to the north magnetic pole as both systems are equipped with a magnetic compass.

The directional buoys measure the north, west and vertical displacements at a rate of 2.56 Hz and the raw data is transferred to a computer onshore through a radio link signal. The current measurement is taken every 10 minutes and is sent by radio link signal to the same computer onshore. The raw data is processed with Datawell Waves4 software suite [6]. Fourier analysis is used to obtain the spectral parameters from the horizontal and vertical displacements over a period of 30 minutes.

Frequency-domain parameters are available for Buoy I, II, and III. Table 2 lists the parameters provided by Buoy I, while Tab. 3 lists the parameters for Buoy II and III.

The water depth in the test site varies from 15 meter closest to the coast to about 25 meter at the deepest. In general the seabed is covered with sand and silt, however at some locations this cover is washed away and the chalk is exposed. DanWEC has carried out a geotechnical survey of the test area which defines the water depth variation as well as the typical variation of the sediments. This information is made available for developers that enter a testing agreement with DanWEC.

Table 2. Frequency-domain parameters for Buoy I

Symbol	Unit	Title
H_{m0}	[m]	Significant wave height
T_e	[s]	Energy wave period
$T_{0,1}$	[s]	Mean wave period
T_z	[s]	Zero-crossing wave period
T_p	[s]	Peak wave period
ϵ	[-]	Spectral bandwidth
L_p	[m]	Peak wavelength
P_w	[W/m]	Wave power

Table 3. Frequency-domain parameters for Buoy II and III

Symbol	Unit	Title
H_{m0}	[m]	Significant wave height
T_e	[s]	Energy wave period
$T_{0,1}$	[s]	Mean wave period
T_z	[s]	Zero-crossing wave period
T_c	[s]	Crest wave period
T_p	[s]	Peak wave period
θ_p	[rad]	Mean wave direction
σ_p	[rad]	Spread of the mean direction
P_w	[W/m]	Wave power

Wind data is also available for the location. It is measured using an anemometer located at the Port of Hanstholm. The data is continuously transmitted to www.hyde.dk. The data presented in table 4 from the website is added to the DanWEC database.

Table 4. Weather data included in the DanWEC database

Parameter	Unit
End sample time (UTC) -	
Water level	[m]
Mean wind speed	[m/s]
Wind direction	[°]
Wind gust	[m/s]
Pressure	[bar]
Temperatur	[°C]

3 Forecast model for the DanWEC test site

The forecast model for the DanWEC test site has been based on the hindcast model for the area. The later will be introduced together with its validation against data from wave measuring buoys. The forecast model will afterwards be presented.

3.1 Hindcast model

The forecast model is based on a hindcast model of the DanWEC test site. For the hindcast model the numerical model used is the MIKE 21 Spectral Wave (SW) model version 2016 [7]. MIKE 21 SW includes the following physical phenomena:

- Wave growth by action of wind
- Non-linear wave-wave interaction (quadruplet and triad-wave interactions)
- Dissipation due to white-capping
- Dissipation due to bottom-friction
- Dissipation due to depth-induced wave breaking
- Refraction and shoaling (due to depth variations and currents)
- Wave-current interaction
- Effect of time-varying water depth and currents

Wave diffraction and wave reflection are not included in this study as no island, headland or other obstruction are present in the area under study. The effect of ice coverage on the wave field is also not relevant for the area under study. The frequency discretization was 25 bins with a minimum frequency of 0.033 Hz

and a logarithmic frequency increment factor of 1.15 resulting in resolved wave periods in the interval [1.2, 33.3] s ([0.033, 0.945] Hz).

The output wave data covers the period from January 15, 1981 to December 31, 2015, a total of 35 years. The number of azimuthal directions in the numerical model is 24. A maximum (adaptive) computational time step of 300 s was applied and the output time step is 1 hour. The bathymetry and grid resolution used in the numerical model are presented in Fig. 2.

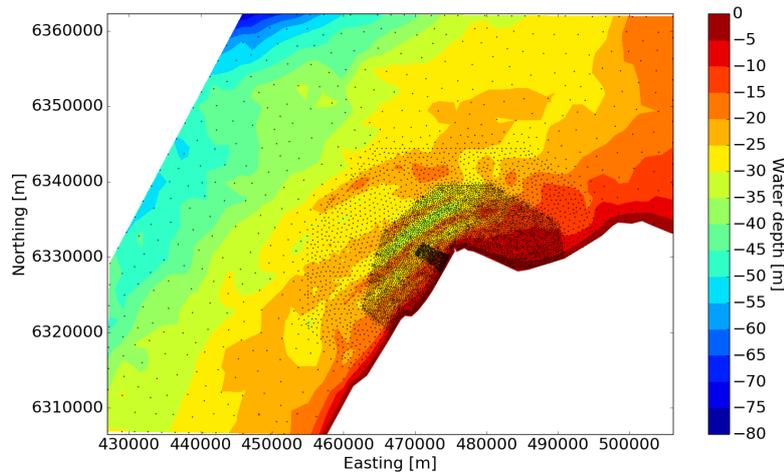


Fig. 2. Bathymetry and mesh resolution used in MIKE 21 Spectral Wave model to obtain the 35 years hindcast data at the DanWEC test site.

Wind forcing was applied with an uncoupled air-sea interaction process. The wind energy momentum transfer to the water was calibrated through the Charnock parameter, which directly influences the amount of energy transferred from the wind to the build-up waves. A Charnock parameter of 0.0185 was applied, which is commonly used for coastal areas. A cap was introduced for the ratio of friction velocity to wind speed (U/U_{10}) [8]. The cap was set at 0.055 [8]. This cap limits the momentum transfer and is based on the documented concept of a saturation of the drag coefficient at extreme storm wind speeds.

Depth-induced wave breaking is a process by which waves dissipate energy when the waves are too high to be stable at the local water depth, i.e. exceeding a limiting wave height to depth ratio. The breaking parameter, γ , varies significantly depending on the wave conditions and the bathymetry. The γ parameter controlling the limiting water depth and the α parameter controlling the rate of dissipation were applied with the default (average) values of $\gamma = 0.8$ and $\alpha = 1$ (see [9, 10]).

Bottom friction was described through the Nikuradse roughness, meaning that the bottom friction varies with the orbital characteristics of the wave close to the bottom. The applied roughness was $k_N = 0.04$ m.

For the boundary conditions, spectral wave data from the Spectral wave model Northern Europe (SWNE) is applied along the open boundaries. The SW_{NE} was validated with data from two different stations: Fjaltring NE and Hirtshals west. This model takes into account coastal reflections. More information is available in [11].

White-capping, a process by which waves dissipate energy, is primarily controlled by the steepness of the waves. The C_{dis} coefficient is a proportional factor on the white-capping dissipation source function and thus controls the overall dissipation rate. The $\text{DELTA}_{\text{dis}}$ coefficient controls the weight of dissipation in the energy/action. These parameters were found from calibration, and are within the range of typically adopted parameters in coastal applications.

3.2 Hindcast model validation

The model was validated against wave measurements from Buoy I, Buoy II and Buoy III. Scatter plots of modelled and observed data is presented in [11]. As an example, a very good agreement between measured and modelled data is shown in Fig. 3 where a snap shot in time between the 12th of September 2015 and the 9th of October 2015 is taken. The results of the validation are summarized in 5, 6 and 7. The quality indices are defined in terms of the observed data (X), the modelled data (Y), and the number of synchronized data used for the validation (N).

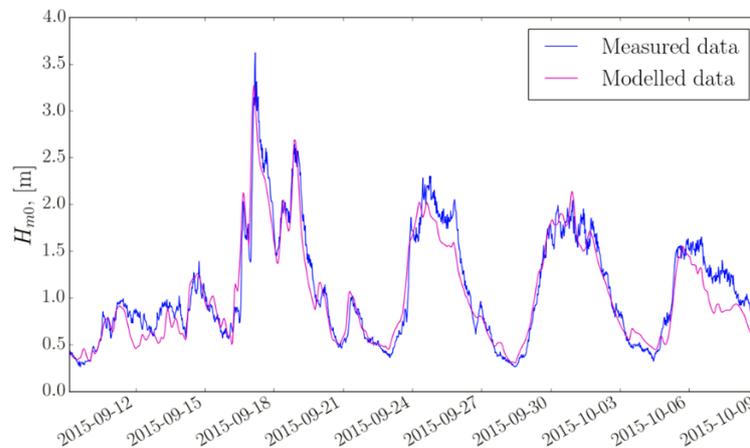


Fig. 3. Measured and modelled data between the 12th of September 2015 and the 9th of October 2015 at the reference point.

\bar{Y} stands for the average of the modelled data Y :

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i \quad (1)$$

\bar{X} stands for the average of the observed data X :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (2)$$

BIAS is the mean difference:

$$\text{BIAS} = \frac{1}{N} \sum_{i=1}^N (Y - X)_i \quad (3)$$

AME is the absolute mean difference:

$$\text{AME} = \frac{1}{N} \sum_{i=1}^N (|Y - X|)_i \quad (4)$$

RMSE is the root mean square difference:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y - X)_i^2} \quad (5)$$

SI is the scatter index unbiased:

$$\text{SI} = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (Y - X - \text{BIAS})_i^2}}{\frac{1}{N} \sum_{i=1}^N |X_i|} \quad (6)$$

EV is the explained variance:

$$\text{EV} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2 - \sum_{i=1}^N [(X_i - \bar{X}) - (Y_i - \bar{Y})]^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (7)$$

CC is the correlation coefficient:

$$\text{CC} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2}} \quad (8)$$

PR is the peak ratio of N_p highest events:

$$\text{PR} = \frac{\sum_{i=1}^{N_p} Y_i}{\sum_{i=1}^{N_p} X_i} \quad (9)$$

Table 5. Summary of the quality indices for the validation of the model against data from Buoy I

	H_{m0}	$T_{z(H_{m0}>0.50 \text{ m})}$
N	169630 (9.7 years)	140991 (8.0 years)
\bar{Y}	1.17 m (94.4%)	4.35 s (95.9%)
BIAS	-0.07 m (-5.6%)	-0.18 s (-4.1%)
AME	0.18 m (14.7%)	0.39 s (8.6%)
RMSE	0.26 m (21.2%)	0.53 s (11.7%)
SI	0.20 (Unbiased)	0.11 (Unbiased)
EV	0.89	0.61
CC	0.94	0.84
PR	1.02 ($N_p = 19$)	0.91 ($N_p = 16$)

Table 6. Summary of the quality indices for the validation of the model against data from Buoy II

	H_{m0}	$T_{z(H_{m0}>0.50 \text{ m})}$
N	13964 (290.9 days)	12325 (256.8 days)
Mean	1.37 m (103.4%)	4.30 s (91.3%)
BIAS	0.04 m (3.4%)	-0.41 s (-8.7%)
AME	0.16 m (11.8%)	0.50 s (10.6%)
RMSE	0.22 m (16.3%)	0.63 s (13.4%)
SI	0.16 (Unbiased)	0.10 (Unbiased)
EV	0.93	0.80
CC	0.97	0.90
PR	0.98 ($N_p = 2$)	0.91 ($N_p = 1$)

Table 7. Summary of the quality indices for the validation of the model against data from Buoy III

	H_{m0}	$T_{z(H_{m0}>0.50 \text{ m})}$
N	13443 (280.1 days)	12112 (252.3 days)
Mean	1.46 m (99.8%)	4.47 s (91.2%)
BIAS	-0.00 m (-0.2%)	-0.43 s (-8.8%)
AME	0.16 m (10.7%)	0.53 s (10.9%)
RMSE	0.22 m (14.8%)	0.67 s (13.7%)
SI	0.15 (Unbiased)	0.10 (Unbiased)
EV	0.94	0.79
CC	0.97	0.89
PR	0.93 ($N_p = 2$)	0.90 ($N_p = 1$)

3.3 Forecast model

The forecast model for the DanWEC test site is based on the hindcast model where the grid output resolution is reduced as shown in Fig. 4. The forecast model updates a 5 day-horizon twice every 24 hours [4]. The model forcing comprises input from regional DHI models and forecast wind fields. The list of output parameters for the forecast model are given in Fig. 8.

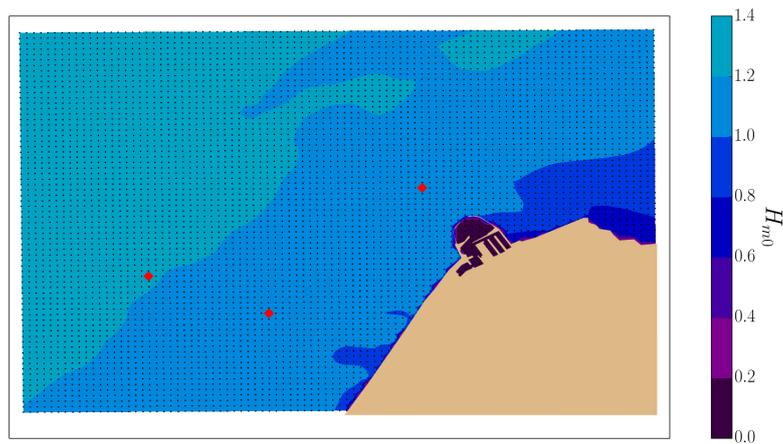


Fig. 4. Example of forecast H_{m0} for the total area provided for the forecast model at DanWEC. The three markers mark the position of the wave measuring buoys and the dots correspond to the mesh of the model, i.e. values for each point can be extracted.

The absolute mean difference (see previous section) for H_{m0} for the forecast model for the month of January 2018 is presented in Fig.5. For the calculation of the error, a point in the forecast modeling area close to the location of Buoy III was chosen together with measurement from Buoy III. Note that the first 48 points correspond to hindcast data. More work is needed to be able to correctly quantify the accuracy of the forecast model.

The ultimate goal is to couple the forecast wave, current and wind conditions to an operation and maintenance (O&M) tool for the test center. By improving the forecast, better planing of the operation can be achieved, leading to a decrease in the levelized cost of energy of offshore renewable energy.

4 Conclusion

In this paper, the forecast model for the DanWEC test center was introduced. The model is strongly based on a hindcast model for the area, which has been

Table 8. List of output parameters for the forecast model updated twice every 24 hours throughout the modelling area (Fig.4)

Parameter	Unit
H_{m0}	[m]
H_{max}	[m]
T_p	[s]
T_{01}	[s]
T_{02}	[s]
Wave direction	[°]
Wind speed	[m/s]
Current speed	[m/s]

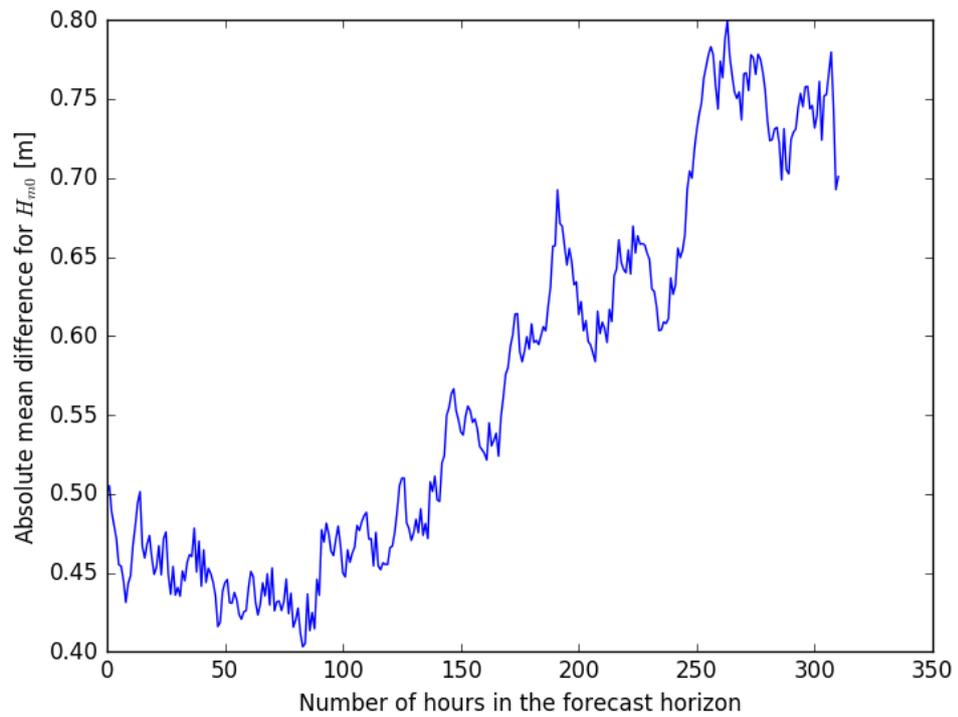


Fig. 5. Absolute mean difference for H_{m0} for the forecast model for the month of January 2018.

validated using measured data from wave measuring buoys placed at the test center. The forecast model provides a 5 day-horizon for a range of parameters enabling a better planing of O&M activities at the test center. Future work includes data assimilation from the wave measuring buoys into the forecast model in order to increase the accuracy of the prediction and thereby a better planing of the O&M at the test center.

Acknowledgments. The authors gratefully acknowledge the financial support from the Danish Energy Agency under The Energy Technology Development and Demonstration Program (EUDP) (*Resource Assessment, Forecasts and WECs O&M strategies at DanWEC and beyond*) which render this work possible.

References

1. Brodersen, H. J., Nielsen, K., Kofoed, J.P.: Development of the Danish test site DanWEC. In: 10th European Wave and Tidal Energy Conference (EWTEC), Aalborg, Denmark (2013)
2. DHI: MIKE 21, Spectral Wave Module, Scientific Documentation. DHI Hørsholm, Denmark (2015)
3. Tetu, A., Kofoed, J.P.: Long-term wave climate at DanWEC. DCE Contract Report No. 188, Department of Civil Engineering, Aalborg Univeristy (2017)
4. Rugbjerg, M., Sørensen, O. R., Jacobsen, V.: Wave Forecasting For Offshore Wind Farms. In: 9th International Workshop on Wave Hindcasting and Forecasting, Victoria, B.C., Canada (2006)
5. Nielsen, K., Remmer, M., Beatie, W.C.: Elements of large wave power plants. In: 1st European Wave and Tidal Energy Conference (EWTEC). NEL-Renewable Energy, Edinburgh, UK (1993)
6. Datawell BV 2014: Datawell Waves4 Manual, User Manual. Datawell BV oceanographic instruments (2014)
7. MIKE 21 SW, <https://www.mikepoweredbydhi.com/products/mike-21/waves>
8. Jensen, R. E., Cardone, V. J., Cox, A.T.: Performance of third generation wave models in extreme hurricanes. In: 9th International Wind and Wave Workshop, Victoria, B.C., Canada (2006)
9. Battjes, J., Stive, M.: Calibration and verification of a dispersion model for random breaking waves. *Geophys. Research.* 112, 307–319 (1985)
10. Kaminsky, G.: Evaluation of depth-limited wave breaking criteria. In: 2nd International Symposium on Ocean Wave Measurement and Analysis, New Orleans, USA (1993)
11. Jensen, P. M.: DanWEC EUDP, Establishment of Wave Hindcast. Technical report, DHI (2016)

ABSTRACT

DENSITY FORECAST COMPARISON FOR DISAGREGGATED MACROECONOMIC RANDOM VARIABLES USING BAYESIAN VAR MODELS, BAYESIAN GLOBAL VAR MODELS AND LARGE BAYESIAN VAR MODELS WITH STOCHASTIC VOLATILITY

Roberto Morales Arsenal
UCM

Miguel Ángel Gómez Villegas
CUNEF

Multivariate macroeconomic models require a set of variables to increase the available information set in order to improve the predictive capacity of the model. In many cases, this set of random variables shows an interdependence or *feedback* that must be taken into account. Macroeconomic models of simultaneous equations, the so-called Vector Autoregressive (VAR) (Sims, 1980), are able to capture this property of interdependence. The problem with these models is the high number of parameters to estimate, a number that increases when we consider time-varying parameters (TVP-VARS) and also allows that the error covariance matrix to change over time. This gives rise to a problem of *overparameterization*. This has led to the use of Bayesian procedures that, using prior information, impose restrictions (*shrinkage*) on the parameters to estimate in the model in a logical and consistent way. In this work we compare a point forecast and the predictive density function obtained from three alternative approaches:

1. **Bayesian vector autoregressive models (BVAR):** Hemos especificado modelos Bayesian VARs with different priors, time varying parameters (TVP) Bayesian VARs (Canova, 1993), and TVP-FAVARs (Kose, Otrok and Whiteman, 2003).
2. **Bayesian Global vector autoregressive models (BGVAR):** The Global VAR (GVAR) approach, originally proposed in Pesaran et al. (2004), provides a relatively simple way of modelling complex high-dimensional systems such as the global economy. The GBVAR can be summarized in two steps: 1) the specification of the individual models (VARX* models) and 2) VARX* models are stacked and solved simultaneously. We impose Bayesian shrinkages (Huber, 2016).
3. **Bayesian large scale vector autoregressive models (BLSVAR):** As alternative to factor models we estimate these models able to handle with a large set of variables incorporating Bayesian shrinkages (Bánbura et al, 2008).

In the previous models, we assume that the volatility of the series remains constant or that it does not change substantially. However, several studies show an increase in the volatility of the macroeconomic series in industrialized countries, especially since 2000. This leads to incorrect density predictions (too broad or too narrow). Therefore, it may be necessary to consider processes with heteroscedasticity by a more flexible specification of the covariance variances matrix. In this sense we introduce models with stochastic volatility Clark (2011),

Clark and Ravazzolo (2015) and (Huber, 2016). Additionally we compare these models using a disaggregated vs aggregated framework. The debate about aggregation versus disaggregation in economic modelling goes back to Theil (1954) and Grunfeld and Griliches (1960). One strand of the literature has focused on the effect of contemporaneous aggregation of forecast accuracy. There are two main arguments for aggregation forecast of disaggregated variables instead of forecasting the aggregated variable of interest directly. One rationale is that disaggregated variables can be better modelled by taking their different dynamic properties into account, and therefore, can be predicted more accurately than the aggregated variable. Modelling disaggregated variables may involve using a larger and more heterogeneous information set, and the specifications may vary across the disaggregated variables (Barker and Pesaran, 1990). A second argument in favor of disaggregation is that forecast errors of disaggregated components might cancel partly, leading to more accurate predictions of the aggregate. Following to Hendry (2003) it is consider forecasting a contemporaneously aggregated variable defined as a variable consisting of the sum or weighted sum of a number of different disaggregated subcomponents at time t .

The disaggregation can be interesting for many countries and economic areas because very often it turns out that components show, among other things, different behaviors (trends, cycles, ect.) and this increase of the information set could be beneficial in the forecast.

The aim of disaggregation is increase the information set. This information set can be enlarged in different non-exclusive directions Espasa et al. (1987):

1. frequency enlargement, integrating more frequent data,
2. enlargement by means of functional and geographical disaggregation of macro variable,
3. enlargement with other related variables.

We apply these models to the Harmonized Consumer Price Index (HCPI) in the Euro Area. The obtained results show the superiority of the Global Bayesian VAR procedures using disaggregated information.

References

George, Edward I., Dongchu Sun, and Shawn Ni (2008) "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, Vol. 142, No. 1, pp. 553 – 580.

Geweke, John (1996) "Bayesian reduced rank regression in econometrics," *Journal of Econometrics*, Vol. 75, No. 1, pp. 121–146, November.

Geweke, John and Charles Whiteman (2006a) "Bayesian Forecasting," in G. Elliott, C. Granger, and A. Timmermann eds. *Handbook of Economic Forecasting*, Vol. 1: Elsevier, 1st edition, Chap. 01, pp. 3–80.

Huber, Florian and Martin Feldkircher (2017) "Adaptive Shrinkage in Bayesian Vector Autoregressive Models," *Journal of Business & Economic Statistics*, Vol. 0, No. 0, pp. 1–13.

Simple estimators for higher-order stochastic volatility models and forecasting *

Md. Nazmul Ahsan[†]
McGill University

Jean-Marie Dufour[‡]
McGill University

September 6, 2018

* The authors thank Manabu Asai, Aman Ullah, Russell Davidson, René Garcia, Eric Renault, Lynda Khalaf, John Galbraith, Victoria Zinde-Walsh, Pascale Valéry, Firmin Doko Tchatoka, Masaya Takano, Purevdorj Tuvaandorj, Byunguk Kang, Jinjing Liu for useful comments and constructive discussions. Earlier versions of this paper were presented at Annual Conference of IAAE (Montreal, 2018), 52nd Annual Conference of the CEA (Montreal, 2018), CIREQ Econometrics Conference (Recent Advances in the Method of Moments, Montreal, 2018), University of Bergamo (Bergamo, 2018), University of Southern California (Los Angeles, 2018), Camp Econometrics (New York, 2018), York University (Toronto, 2018), CIREQ-McGill Lunch Seminar (Montreal, 2016, 2017), 11th World Congress of the Econometric Society (Montreal, 2015), and 11th CIREQ PhD Students' Conference (Montreal, 2015). This paper was previously circulating under the title "Simple estimators and inference for higher-order stochastic volatility models" and "Closed-Form Estimator for Higher-order Gaussian Stochastic Volatility Models".

This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universidad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche sur la société et la culture (Québec).

[†] Ph.D. Candidate, Department of Economics, McGill University and Centre interuniversitaire de recherche en analyse des organisations (CIRANO). Mailing address: Department of Economics, McGill University, Leacock Building, 855 Sherbrooke Street West, Montreal, Quebec H3A 2T7, Canada; e-mail: md.ahsan@mail.mcgill.ca.

[‡] William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 414, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 6071; FAX: (1) 514 398 4800; e-mail: jean-marie.dufour@mcgill.ca. Web page: <http://www.jeanmariedufour.com>

ABSTRACT

We focus on higher-order stochastic volatility models [SV(p)] and propose several estimators: two simple estimators [moment-based and ARMA-based] and GMM estimators. The estimation and inference are challenging in SV models due to the inherent problem of evaluating the likelihood function – a general feature of most non-linear latent variable models. Most of the existing estimation methods of SV models are confined to estimate an SV(1) model and are computationally expensive, inflexible [not easy to generalize for SV(p) models], difficult to implement and additionally inefficient. The estimation of SV(p) models is even more challenging than SV(1) model. So these models are rarely considered in the literature. Compared to the existing methods for SV(p) models, our simple estimators are computationally simple and very easy to implement. These estimators do not require choosing a sampling algorithm, initial parameter values, and an auxiliary model. Also, we suggest winsorized versions of the simple ARMA-based estimator to ensure the stationarity condition, especially in a small sample or in the presence of outliers. We develop recursive estimation procedures which allow for recursive-in-order calculation of the parameters of higher-order SV processes. We derive asymptotic theories for simple estimators and show the usefulness of these estimators in the context of simulation-based inference technique, i.e., Monte Carlo (MC) tests. By simulation, we compare our proposed estimators to the Bayesian MCMC estimator. The results show that the simple winsorized ARMA-based estimator is uniformly superior to other estimators in terms of bias and root mean square error. We present empirical applications related to SV(p) models and the ARMA-based estimator. First, using the daily return on the S&P 500 index from 1928 to 2016, we find that higher-order SV models may be preferable for in-sample model fitting and this result confirms by both asymptotic and finite-sample tests. Second, using different volatility proxies [the squared return of S&P 500 index and the realized volatility of S&P 500, FTSE100, NASDAQ100, N225, SSMI20 indices], we conduct two out-of-sample forecast experiments: (1) we forecast a moderately volatile period after the late-2000s Financial Crisis; (2) we forecast a highly volatile period, i.e., the core Financial Crisis. We compare the accuracy of volatility forecasts among SV(p) models, GARCH models, and Heterogenous Autoregressive model of Realized Volatility (HAR-RV) models. The results suggest that SV(p) models perform better than other models in most cases. This finding holds even if a high volatility period (such as Financial Crisis) is included in the estimation sample or the forecasted sample. Formal prediction tests, i.e., model confidence set procedure, also support these inferences. Our findings highlight the usefulness of higher-order SV models for volatility forecasting.

Key words: generalized method of moments, Markov Chain Monte Carlo, Monte Carlo tests, stochastic volatility, asymptotic distribution, stock returns, realized variance, volatility forecasting, high frequency data.

Journal of Economic Literature classification: C15, C22, C53, C58.

Entropy-based Channel Selection using Supervised Temporal Patterns in MI Tasks

No Author Given

No Institute Given

1 **Abstract.** Brain-Computer Interfaces decode brain activity to transmit
2 commands to external devices. One of the most known applications uses
3 Motor Imagery paradigm, the imagination of a motor action, based on
4 the similarities at the neural level between imagination and execution.
5 However, there are many issues hampering straightforward translation
6 of experimental applications into real-world tools as the limited signal
7 quality and task-irrelevant and redundant channels. Moreover, motor
8 imagery patterns depend on complex brain activations including non-
9 time-locked or induced responses. This work proposes a channel selection
10 based on Supervised Temporal Patterns and Sample Entropy estimation
11 to find the channels that maximize the difference between MI tasks. We
12 evaluated the proposed methodology on a well-known BCI dataset. The
13 attained performance showed that the method is able to select discrimi-
14 nant channels to both highlighting the MI-induced and reducing evoked
15 activity.

16 **Keywords:** Channel Selection · Supervised Temporal Patterns · EEG · Entropy
17 · Motor Imagery

18 1 Introduction

19 Brain-Computer Interfaces (BCI) interpret the brain activity to transmit com-
20 mands to external devices. Recently, BCI systems have been widely used in the
21 development of tools to improve the live quality of physically-disabled people,
22 enhancing the experience of games, and improve the attentional and concentra-
23 tion brain networks [19]. Although brain activity is both observed and registered
24 by several noninvasive neuroimage techniques, the high temporal resolution and
25 low cost of the electroencephalogram (EEG) signals, that are records the neural
26 activation at electrodes placed over the scalp, make it widely used to study brain
27 dynamics in several applications as evoked potentials, investigating epilepsy and
28 locating seizure origin [16,9]. Regarding brain motor activity, Motor imagery
29 (MI) paradigm uses the imagination to provide a communication link between
30 the brain and the devices based on the similarities at the neural level between
31 imagination and execution action [18,7].

32 However, there are many remaining issues hampering straightforward transla-
33 tion of experimental BCIs into real-world applications as the limited quality (low

34 signal-to-noise) and task-irrelevant and redundant channels [11,3,2,12]. In [17],
35 they proposed channel selection employing the Event-Related Desynchronization
36 (ERD) and Synchronization (ERS). ERD/S show the channel-wise temporal dy-
37 namics within an EEG channel as the average energy relative to a reference time
38 segment. Despite highlighting changes in the excitation at the cortical motor re-
39 gions and frequency bands, ERD/S demands the application of a large bank
40 of narrowband filters to find dynamic changes [20,4]. Moreover, the MI neural
41 patterns depend on complex brain activations including visual stimuli, selection,
42 and generation of the appropriate MI [1]. The above activations generate non-
43 time-locked or induced responses, which indicates that the MI activity varies
44 over trials [15,6]. To solve such an issue, the temporal design of filters from data
45 highlights the EEG time dynamics in [5].

46 However, there are many issues hampering straightforward translation of
47 experimental MI tools into real-world applications as the limited quality (low
48 signal-to-noise) and task-irrelevant and redundant channels [11,3,2,12]. In [17],
49 they proposed channel selection employing the Event-Related Desynchronization
50 (ERD) and Synchronization (ERS). ERD/S show the channel-wise temporal
51 dynamics within an EEG channel as the average energy relative to a reference
52 time segment. Despite highlighting changes in the excitation at the cortical motor
53 regions and frequency bands, ERD/S demands the application of a large bank of
54 narrowband filters to find dynamic changes [20,4]. However, MI patterns depend
55 on complex brain activations including visual, attentional, and sensory-motor
56 tasks generating non-time-locked or induced responses [1,15,6]. To solve such an
57 issue, the temporal design of filters from data highlights the EEG time dynamics
58 in [5]. Aiming to overcome the redundant channels, this work proposes a channel
59 selection based on the estimation of Supervised Temporal Patterns (STP) that
60 decode brain dynamics in MI paradigm. The initial time embedding of EEG
61 channels tackles the temporal variability. Then, the STP result from the solution
62 of a generalized eigenvalues problem as the impulse response of a linear filter.
63 Later, Sample Entropy (SampEn) estimates the amount of information and a
64 statistical test finds the channels that maximize the difference between MI tasks
65 at the significance level.

66 We evaluated the proposed methodology on a well-known BCI dataset. The
67 obtained results indicated the STP is allowed to pick discriminative channels by
68 identifying the temporal dynamics from EEG data. The attained performance
69 showed that the method is able to select discriminant channels, taking the 39% of
70 the channel dataset without significant accuracy differences to both highlighting
71 the MI-induced and reducing evoked activity. Additionally, the selected EEG
72 channels are consistent with the neurophysiological findings. The remainder of
73 this work is organized as follows: Section 2.1 contains the theoretical background
74 of channel selection based on Supervised Temporal Patterns estimation. Section 3
75 describes the performed experiments. Section 4 discusses the attained results.
76 Finally, Section 5 presents the conclusions and future work.

77 2 Methods

78 2.1 Channel Selection based on Supervised Temporal Patterns

79 Let $\{\mathbf{X}_i \in \mathbb{R}^{C \times T} : i \in [1, N]\}$ be a set of N EEG trials with C channels, lasting
 80 T time instants, and $l_i \in \{+, -\}$ be a bi-class set of MI tasks. For each i -th
 81 trial, we compute the channel-wise time-embedded matrices for reference and
 82 MI activity conditions, $\mathbf{R}_i^c \in \mathbb{R}^{W \times M}$ and $\mathbf{A}_i^c \in \mathbb{R}^{W \times M}$, where W is the embedding
 83 dimension and M denotes the number of time-lagged windows that fit within
 84 the segmented time window. Then, provided a class $c \in C$, the average temporal
 85 covariances are given by $\Sigma_R^c = \mathbb{E}_i \{\mathbf{R}_i^c \mathbf{R}_i^{c\top}\}$ and $\Sigma_A^c = \mathbb{E}_i \{\mathbf{A}_i^c \mathbf{A}_i^{c\top}\}$ for both brain
 86 activity conditions: the reference state (that is, there is no elicitation) and MI
 87 state (during the elicitation), respectively. Notation $\mathbb{E}_\xi \{\cdot\}$ stands for expectation
 88 operator across variable ξ .

89 Relying on a couple of introduced matrices, $\Sigma_R^c \in \mathbb{R}^{W \times W}$ and $\Sigma_A^c \in \mathbb{R}^{W \times W}$,
 90 the temporal pattern $\mathbf{w}_c \in \mathbb{R}^W$ of class c is calculated through the Rayleigh quo-
 91 tient as below:

$$J(\mathbf{w}_c) = \frac{\mathbf{w}_c^\top \Sigma_A \mathbf{w}_c}{\mathbf{w}_c^\top (\Sigma_A + \Sigma_R) \mathbf{w}_c}, J(\mathbf{w}_c) \in \mathbb{R}^+ \quad (1)$$

92 Optimization of Eq. (1) is performed by solving a generalized eigenvalue prob-
 93 lem, for which the first eigenvector \mathbf{w}_c is computed as a result of maximization
 94 of the variability of MI activity conditions.

95 Aiming to compute further the optimal temporal basis function of the super-
 96 vised temporal patterns (STP), a time-delay-embedding matrix is built, in which
 97 its additional rows are created from time-delayed versions of existing rows [5].
 98 Therefore, the computed STP set holds a weighted combination of successive time
 99 points, which are optimized to distinguish between both brain neural states at
 100 the channel level. Given the linear nature of STP, the convolution with \mathbf{w}_c filters
 101 the corresponding channel of each trial $\mathbf{x}_i^c \in \mathbb{R}^T$.

102 Lastly, due to the high variability among trials, we perform the channel selec-
 103 tion by means of a two-sample t -test for equal means to determine whether the
 104 average amount of information is the same for the two MI tasks. Since account-
 105 ing for the complexity of a time series, the Sample Entropy (SampEn) estimates
 106 the amount of information as the difference between the log-probability of two
 107 different-sizing time embedding to hold a unique sample within the same time
 108 window, according to a predefined distance function and a similarity threshold.
 109 Then, the smaller the SampEn - the more self-similar the time segment or less
 110 informative. We compute the channel-wise SampEn from the envelope of each
 111 filtered trial. Therefore, the statistical test finds the channels that maximizes
 112 the difference between either MI task (+ or -), at the significance level $\alpha \in \mathbb{R}^+$.

113 3 Experimental Set-Up

114 3.1 EEG dataset

115 The BCI Competition (IV Dataset 1) includes EEG signals from four real sub-
 116 jects (labeled as S1, S2, S6, and S7) using 59 electrode positions, most densely
 117 distributed over the sensorimotor area and downsampled at $F_s=100$ Hz. The
 118 dataset consists of six runs, performed without feedback, divided into two runs
 119 for calibration and four runs for evaluation. The proposed analysis only consid-
 120 ers calibration data due to its acquisition conditions. Fig. 1 explains the trial
 121 timing for calibration data. First, the screen shows a fixation cross for 2 s. Next,
 122 a 4-second arrow indicates to the subject to perform the MI task by pointing
 123 left, right, or down (left hand, right hand, or foot, respectively). Lastly, a blank
 124 screen designates the end of the trial during 2 s. For each subject, we selected
 125 left hand and right hand from the three MI task, each of them with 100 trials.

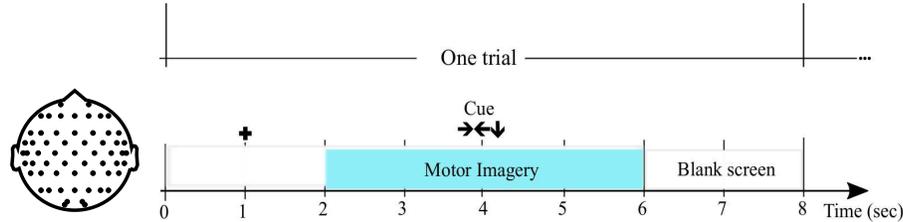


Fig. 1: Trial timing of the tested MI task (Cue MI: left hand, right hand).

126 3.2 Parameter set-up

127 Fig. 2 shows the framework to validate the performance of the proposed CSSTP
 128 including three stages: *i*) channel selection analysis based on the introduced tem-
 129 poral patterns; *ii*) the feature representation, which comprises Common Spatial
 130 Patterns (CSP); and *iii*) the classification stage by means of a Linear Discrimi-
 131 nant Analysis (LDA).

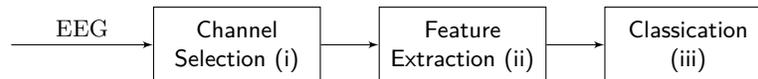


Fig. 2: Tested framework to validate the proposed CSSTP.

132 To estimate the temporal patterns, we use as reference the period during the
 133 subject is watching a fixation cross ($[0, 2]$ s), and as activity windows the inter-
 134 val that includes the neural activity of the subject performing the selected MI

135 tasks ($[2, 4]$ s). The time embedding size is empirically-fixed to $M=180$, yielding
136 $N=200$ time windows. Besides, we contrast the CSSTP against the state-of-the-
137 art FSDE [17], which relies on the *Event-Related Desynchronization* (ERD) and
138 *Synchronization* (ERS) on narrow-bands. Spectral analysis is performed by fil-
139 tering each EEG channel with a linearly-distributed filter-bank with 2 Hz band-
140 width and 1 Hz overlap within $[2, 40]$ Hz; and the temporal analysis is carried
141 out by computing the SampEn on sliding windows within $[2.5, 3.5]$ s, as proposed
142 by the authors. For both compared channel selection approaches, we generate a
143 performance curve by adding, one at a time, the selected channels to the feature
144 representation and classification stages. In this regard, CSP extracts six features
145 from the MI window ($[2, 5 - 4, 5]$ s) of each trial [3], while LDA is tested using a
146 five-fold cross-validation scheme.

147 4 Results and Discussion

148 Left-side column of Fig. 3 displays the spectral representation of all CSTP ob-
149 tained for a varying number of channels: 6 more relevant (6Ch), all relevant
150 channels that have been selected (NSCh), and the whole number of channels
151 (59Ch). For the representation convenience, all spectra are plotted in log scale.
152 Although there is a certain similarity between patterns, the influence of adding
153 more channels affects differently on each patient. As seen in Table 1, the dif-
154 ference between the 6Ch and NSCh patterns is very close to the one between
155 NSCh and 59Ch, meaning that the behavior remains alike across all channels.
156 This situation holds for patients $S1$, $S6$, and $S7$. By contrast, the 6Ch and NSCh
157 patterns vary the most for patient $S2$, while NSCh and 59Ch spectra are similar,
158 pointing out on the diverse behavior of brain activity across space.

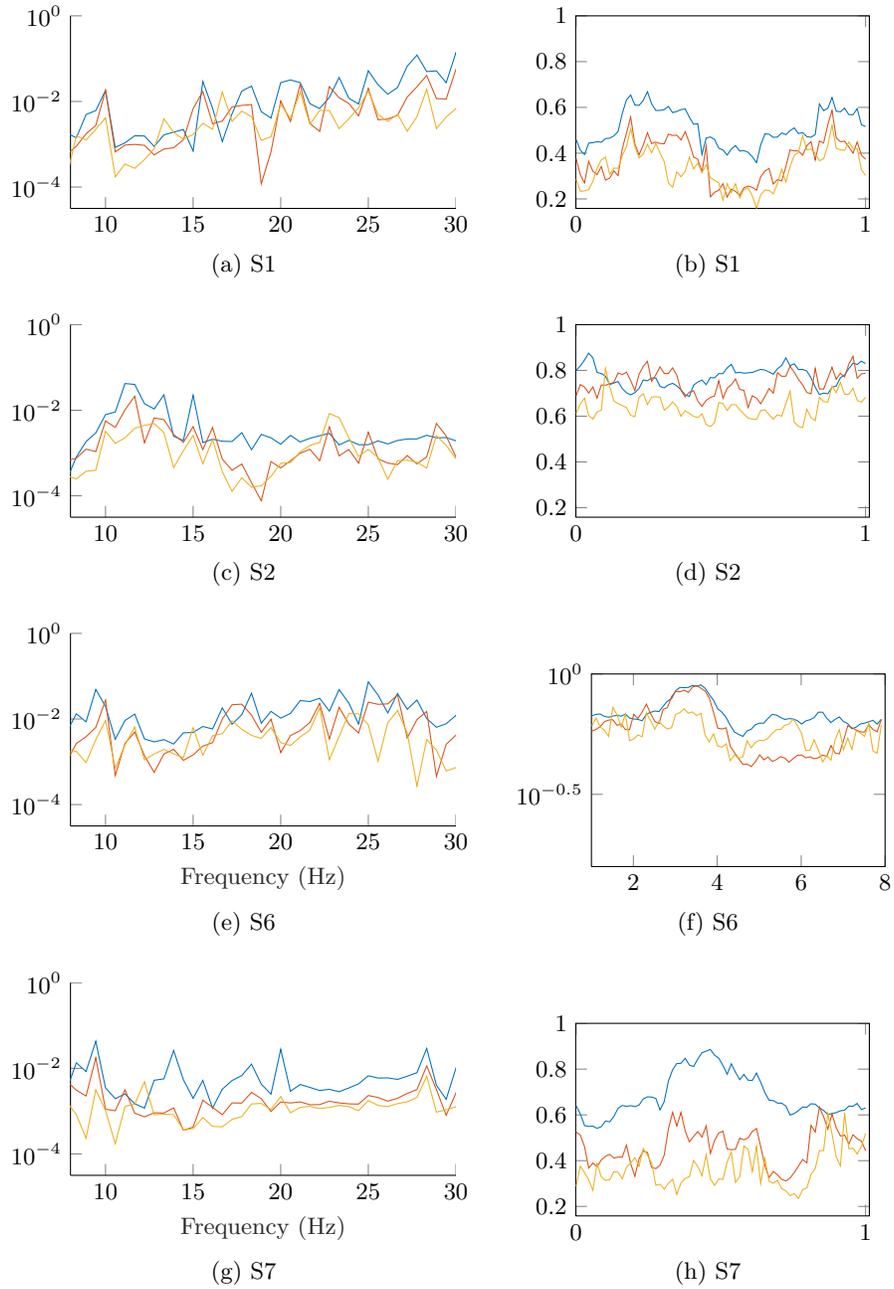


Fig. 3: Estimated CSTP spectra and Rayleigh quotients along the time domain, employing different number of channels

159 Likewise, in the right-side column, the time-varying Rayleigh quotients are
 160 shown, making explicit the timing event synchronization as the number of chan-
 161 nels increases. So, although their brain activity behaves differently during the
 162 cue elicitation, patients *S1*, *S6*, and *S7* adequately respond to the first cue flank
 163 (close to 2-3 s from the star). However, the Rayleigh quotient of patient *S2* evi-
 164 dences a very weak influence of ERP dynamics regardless of the involved chan-
 165 nels. Moreover, this patient evidences a high variability between the quotient
 166 shapes, highlighting the diversity of neural activity through the space domain
 167 as seen in Table 1.

Table 1: Correlation index values, measuring the relationship between shapes of CSTP spectra and Rayleigh quotient

Channels	CSTP spectra				Rayleigh quotient			
	<i>S1</i>	<i>S2</i>	<i>S6</i>	<i>S7</i>	<i>S1</i>	<i>S2</i>	<i>S6</i>	<i>S7</i>
6Ch - NSCh	0.71	0.58	0.92	0.75	0.87	0.02	0.83	0.40
6Ch - 59Ch	0.57	0.72	0.82	0.73	0.72	0.58	0.58	0.07
NSCh - 59Ch	0.75	0.91	0.91	0.99	0.75	0.50	0.58	0.39

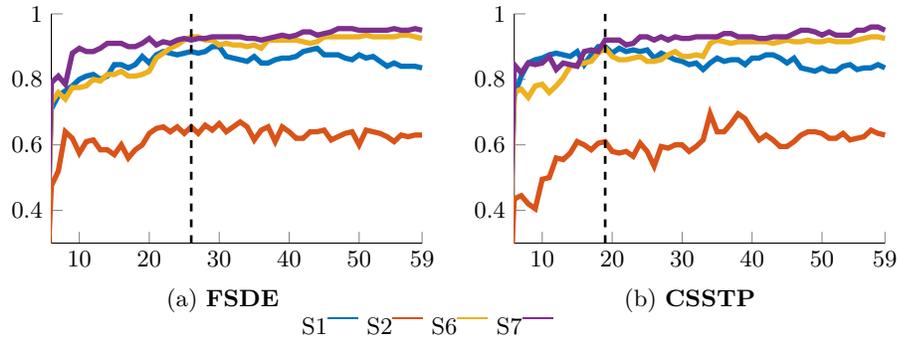


Fig. 4: Performance plots performed by FSDE and CSSTP methods.

168 Fig. 4 shows the accuracy results according to the channel selection stage for
 169 FSDE and CSTP methods. From the attained results, it is possible to notice that
 170 the number of selected channels for the maximum system performance varies
 171 between methods and subjects. The results of FSDE in Fig. (a) shows that
 172 using the first 26 selected channels, it achieves the maximum accuracy without
 173 significant differences by adding more channels. For comparison, the proposed
 174 CSSTP method in Fig. (b), uses in general, the first 19 selected channels to
 175 obtain the maximum accuracy without significant differences by adding more

176 channels. Although, the subject *S2* attains the maximum performance using the
 177 35 selected channels exhibiting the lower performance from all analyzed subjects
 for both methods.

Table 2: Classification results (average accuracy \pm standard deviation) and selected number of channels for FSDE and CSSTP methods.

Subjects	FSDE	N. Channels	CSSTP	N. Channels
S1	88.5% \pm 6.02%	22	90.0% \pm 8.48%	18
S2	65.5% \pm 8.37%	22	69.5% \pm 6.47%	35
S6	93.0% \pm 3.26%	26	89.5% \pm 3.26%	19
S7	92.5% \pm 3.54%	28	92.0% \pm 2.09%	19
Mean	84.9% \pm 5.33%	25	85.3%\pm 5.08%	23

178
 179 The attained performance results in Table 2 displays the selected number of
 180 channels using FSDE and CSSTP for all subjects. In general, CSSTP selects a
 181 less amount of channels and improves the accuracy results than FSDE method.
 182 The temporal patterns filter the EEG signals highlighting brain dynamics that
 183 are related to both induced and evoked activity present in MI task. However, all
 184 subjects demand different among channels, it is explained by the fact that each
 185 subject presents different cognitive behavior and the EEG data is frequently
 186 contaminated by artifacts [8,13].

187 From Fig. 5, it is possible to see the selected channels by FSDE and CSSTP
 188 methods. Mainly, FSDE selects the 48% of channels of the analyzed dataset. For
 189 all subjects (Figs. (a),(b),(c),(c)), FSDE picks channels over the occipital region.
 190 Occipital lobe exhibits brain activity associated with image perception and not
 191 with MI tasks. This activity is related to the time-locked activity (the activity
 192 that manifests the same brain response at roughly the same time) displayed in
 193 the ERD/S [12]. Also, FSDE selects channels over the sensorimotor cortex and
 194 spread activity over frontal, parietal, and temporal brain regions. Although these
 195 areas are related to motor and cognitive activity associated with MI, it is impor-
 196 tant to mark the fact a few channels are selected over the primary motor cortex
 197 (M1) which is principally linked to MI. For its part, CSSTP chooses 32% of the
 198 dataset channels. Particularly, it does not select the occipital related activity
 199 for subjects with the highest accuracy results, *S1* and *S7* and highlights the
 200 sensorimotor activity for subjects *S1* (Fig. (e)), *S6* (Fig. (g)) and *S7* (Fig. (h)).
 201 However, subject *S2* (Fig. (f)) does not select channels over the motor cortex.
 202 *S2* performs the lower accuracy for both methods that could be related to both
 203 psychological and physiological states generate a low and the MI-illiteracy hav-
 204 ing a less-developed brain network that is incapable of MI tasks [1]. For all the
 205 subjects CSSPT finds discriminative channels over the M1 and others secondary
 206 motor cortices as the posterior parietal cortex (PP) and supplementary motor
 207 area (SMA), which is consonant with MI neurophysiology. Movement prepara-

208 tion connects the M1 (sensory processing), the PP (translate visual information
 209 into motor commands and generating mental movement representations), and
 210 SMA (planning and coordinating tasks) brain regions [14,10].

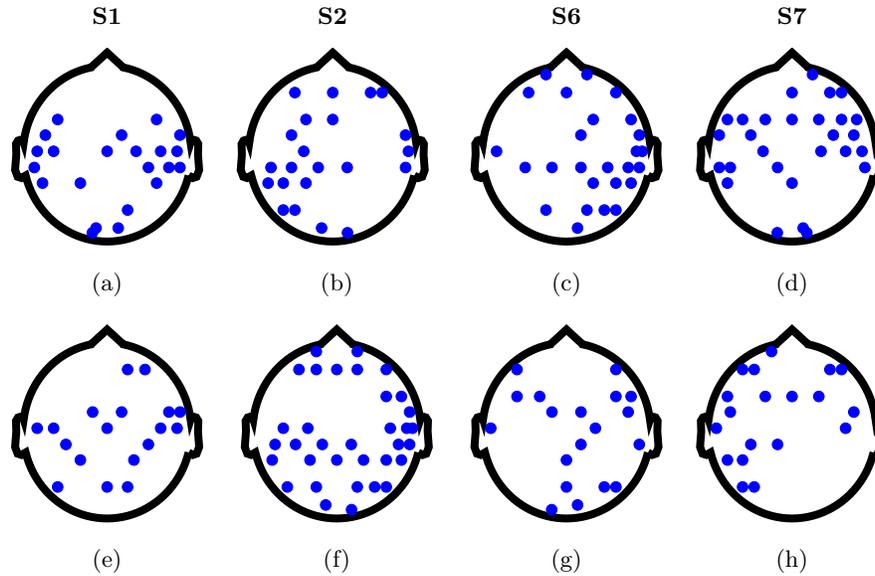


Fig. 5: Selected channels. Top row: FSDE. Bottom row: CSSTP.

211 5 Conclusions and Future Work

212 The present work proposes a channel selection analysis through the supervised
 213 temporal patterns estimation allowing to pick discriminative channels by iden-
 214 tifying the temporal dynamics from EEG data. We tested the proposed CSSTP
 215 method using a well-known BCI data set based on MI paradigm and compared
 216 the results with a state-of-the-art method. The attained performance showed
 217 that CSSTP is able to select discriminant channels, taking the 39% (23) of the
 218 channel dataset without significant accuracy differences to both highlighting the
 219 MI-induced and reducing evoked activity. The chosen brain regions are consistent
 220 with the neurophysiological findings. As future work, we propose two research
 221 directions. Firstly, we purpose adding a preprocessing stage for artifact reduc-
 222 tion. Secondly, we plan to test additional information measures to SamEn to
 223 find the channels that maximize the differences between the two MI tasks.

224 **References**

- 225 1. Ahn, M., Jun, S.C.: Performance variation in motor imagery brain-computer inter-
226 face: a brief review. *Journal of neuroscience methods* **243**, 103–110 (2015)
- 227 2. Al-Ani, A., Koprinska, I., Naik, G.: Dynamically identifying relevant eeg channels
228 by utilizing channels classification behaviour. *Expert Systems with Applications*
229 **83**, 273–282 (2017)
- 230 3. Alimardani, F., Boostani, R., Blankertz, B.: Weighted spatial based geometric
231 scheme as an efficient algorithm for analyzing single-trial eegs to improve cue-
232 based bci classification. *Neural Networks* **92**, 69–76 (2017)
- 233 4. Bian, Y., Qi, H., Zhao, L., Ming, D., Guo, T., Fu, X.: Improve-
234 ments in event-related desynchronization and classification performance
235 of motor imagery using instructive dynamic guidance and com-
236 plex tasks. *Computers in Biology and Medicine* **96**, 266 – 273 (2018).
237 <https://doi.org/https://doi.org/10.1016/j.combiomed.2018.03.018>
- 238 5. Cohen, M.X.: Using spatiotemporal source separation to identify prominent fea-
239 tures in multichannel data without sinusoidal filters. *European Journal of Neuro-*
240 *science* (2017)
- 241 6. Cohen, M.X.: *Analyzing neural time series data: theory and practice*. MIT press
242 (2014)
- 243 7. Emami, Z., Chau, T.: Investigating the effects of visual distractors on the per-
244 formance of a motor imagery brain-computer interface. *Clinical Neurophysiology*
245 **129**(6), 1268–1275 (2018)
- 246 8. Giakoumis, D., Tzovaras, D., Hassapis, G.: Subject-dependent biosignal features
247 for increased accuracy in psychological stress detection. *International Journal of*
248 *Human-Computer Studies* **71**(4), 425–439 (2013)
- 249 9. Guerrero-Mosquera, C., Navia-Vázquez, A.: Automatic removal of ocular artefacts
250 using adaptive filtering and independent component analysis for electroencephalo-
251 gram data. *IET signal processing* **6**(2), 99–106 (2012)
- 252 10. Hanakawa, T., Immisch, I., Toma, K., Dimyan, M.A., Van Gelderen, P., Hallett, M.:
253 Functional properties of brain areas associated with motor execution and imagery.
254 *Journal of neurophysiology* **89**(2), 989–1002 (2003)
- 255 11. Kee, C.Y., Ponnambalam, S., Loo, C.K.: Multi-objective genetic algorithm as chan-
256 nel selection method for p300 and motor imagery data set. *Neurocomputing* **161**,
257 120–131 (2015)
- 258 12. Lo, C.C., Chien, T.Y., Chen, Y.C., Tsai, S.H., Fang, W.C., Lin, B.S.: A wearable
259 channel selection-based brain-computer interface for motor imagery detection. *Sen-*
260 *sors* **16**(2), 213 (2016)
- 261 13. OâĀĹŽRegan, S., Faul, S., Marnane, W.: Automatic detection of eeg artefacts arising
262 from head movements using eeg and gyroscope signals. *Medical engineering &*
263 *physics* **35**(7), 867–874 (2013)
- 264 14. Saiote, C., Tacchino, A., Bricchetto, G., Roccatagliata, L., Bommarito, G., Cordano,
265 C., Battaglia, M., Mancardi, G.L., Inglese, M.: Resting-state functional connectiv-
266 ity and motor imagery brain activation. *Human brain mapping* **37**(11), 3847–3857
267 (2016)
- 268 15. Samek, W., Nakajima, S., Kawanabe, M., Müller, K.R.: On robust parameter esti-
269 mation in brain-computer interfacing. *Journal of neural engineering* **14**(6), 061001
270 (2017)
- 271 16. Sanei, S., Chambers, J.A.: *EEG signal processing*. John Wiley & Sons (2013)

- 272 17. Wang, D., Miao, D., Blohm, G.: Multi-class motor imagery eeg decoding for brain-
273 computer interfaces. *Frontiers in neuroscience* **6**, 151 (2012)
- 274 18. Wang, J., Feng, Z., Lu, N., Sun, L., Luo, J.: An information fusion scheme based
275 common spatial pattern method for classification of motor imagery tasks. *Biomed-*
276 *ical Signal Processing and Control* **46**, 10–17 (2018)
- 277 19. Yang, Y., Chevallier, S., Wiart, J., Bloch, I.: Subject-specific time-frequency
278 selection for multi-class motor imagery-based bcis using few laplacian eeg
279 channels. *Biomedical Signal Processing and Control* **38**, 302 – 311 (2017).
280 <https://doi.org/https://doi.org/10.1016/j.bspc.2017.06.016>
- 281 20. Yuan, H., He, B.: Brain–computer interfaces using sensorimotor rhythms: current
282 state and future perspectives. *IEEE Transactions on Biomedical Engineering* **61**(5),
283 1425–1435 (2014)

Maximum Entropy Methodologies in Large-Scale Data

Maria da Conceição Costa and Pedro Macedo

Department of Mathematics and CIDMA – Center for Research and Development in Mathematics and Applications, University of Aveiro, 3810-193, Aveiro, Portugal
{lopescosta, pmacedo}@ua.pt
<http://www.ua.pt>

Abstract. *It was already in the fifties of the last century that the relationship between information theory, statistics, and maximum entropy was established, following the works of Kullback, Leibler, Lindley and Jaynes. However, the applications were restricted to very specific domains and it was not until recently that the convergence between information processing, data analysis and inference demanded the foundation of a new scientific area, commonly referred to as Info-Metrics [1], [2]. As huge amount of information and large-scale data have become available, the term "big data" has been used to refer to the many kinds of challenges presented in its analysis: many observations, many variables (or both), limited computational resources, different time regimes or multiple sources. In this work, we consider one particular aspect of big data analysis which is the presence of inhomogeneities, compromising the use of the classical framework in regression modelling. A new approach is proposed, based on the introduction of the concepts of info-metrics to the analysis of inhomogeneous large-scale data. The framework of information-theoretic estimation methods is presented, along with some information measures. In particular, the normalized entropy is tested in aggregation procedures and some simulation results are presented.*

Keywords: Big Data, Info-Metrics, Maximum Entropy

1 Introduction

Inference and processing of limited information is still one of the most fascinating universal problems. As stated by Amos Golan in [2], a very recent publication, "[...] the available information is most often insufficient to provide a unique answer or solution for most interesting decisions or inferences we wish to make. In fact, insufficient information - including limited, incomplete, complex, noisy and uncertain information - is the norm for most problems across all disciplines." Also, regardless of the system or question studied, any researcher observes only a certain amount of information or evidence and optimal inference must take into account the relationship between the observable and the unobservable, [3].

Info-Metrics is a constrained optimization framework for information processing, modelling and inference with finite, noisy or incomplete information. It

is at the intersection of information theory, statistical methods of inference, applied mathematics, computer science, econometrics, complexity theory, decision analysis, modelling and the philosophy of science, [2].

As Info-Metrics generalizes the Maximum Entropy (ME) principle by Jaynes, [4], [5], which in turn relies on the maximization of Shannon's entropy, the notions of information, uncertainty and entropy are fundamental to the understanding of the methodologies involved. Each scientist and discipline have their own interpretation and definition of information within the context of their research and understanding but, in the context of Info-Metrics, it refers to the meaningful content of data, its context and interpretation and how to transfer data from one entity to another. As for uncertainty, it arises from a proposition or a set of possible outcomes where none of the choices or outcomes is known with certainty (a proposition is uncertain if it is consistent with knowledge but not implied by knowledge). Therefore, these outcomes are represented by a certain probability distribution. The more uniform the distribution, the higher the uncertainty that is associated with this set of propositions or outcomes. Finally, the concept of entropy reflects what, on average, we expect to learn from observations and it depends on how we measure information. Technically, entropy is a measure of uncertainty of a single random variable. As such, entropy can be viewed as a measure of uniformity.

For a brief discussion of entropy, let us consider the set $\mathbf{A} = \{a_1, a_2, \dots, a_K\}$ to be a finite set and \mathbf{p} a proper probability mass function on \mathbf{A} . The amount of information needed to fully characterize all of the elements of this set consisting of K discrete elements is defined by the Hartley's formula, $I(\mathbf{A}_K) = \log_2 K$. Shannon's information content of an outcome a_k is $h(a_k) = h(p_k) \equiv \log_2 \frac{1}{p_k}$. Shannon's entropy reflects the expected information content of an outcome and is defined as

$$H(\mathbf{p}) \equiv \sum_{k=1}^K p_k \log_2 \frac{1}{p_k} = - \sum_{k=1}^K p_k \log_2 p_k = E \left[\log_2 \left(\frac{1}{p(X)} \right) \right], \quad (1)$$

for the random variable X . This information criterion, expressed in bits, measures the uncertainty of X that is implied by \mathbf{p} . The entropy measure $H(\mathbf{p})$ reaches a maximum when $p_1 = p_2 = \dots = p_K = \frac{1}{K}$ and a minimum with a point mass function. The entropy $H(\mathbf{p})$ is a function of the probability distribution \mathbf{p} and not a function of the actual values taken by the random variable.

The remainder of the paper is laid out as follows: in Section 2, maximum entropy and generalized maximum entropy estimation are briefly discussed. Section 3 illustrates some traditional aggregation procedures and a new proposal based on normalized entropy. Section 4 presents simulation results. Some conclusions and topics for future research are given in Section 5.

2 Generalized Maximum Entropy Estimator

The ME principle was discussed by Golan, Judge and Miller, [6], in order to develop analytical and empirical methods for recovering the unobservable pa-

rameters of a pure linear inverse problem. Considering then

$$\mathbf{y} = \mathbf{X}\mathbf{p}, \tag{2}$$

where \mathbf{y} is the vector ($N \times 1$) of observations, \mathbf{X} is a non-invertible matrix ($N \times K$) with $N < K$, and \mathbf{p} is the vector ($K \times 1$) of unknown probabilities, the ME principle consists in choosing \mathbf{p} that maximizes Shannon's entropy

$$H(\mathbf{p}) = - \sum_{k=1}^K p_k \ln p_k = -\mathbf{p}' \ln \mathbf{p}, \tag{3}$$

subject to the data consistency restriction, $\mathbf{y} = \mathbf{X}\mathbf{p}$, and the additivity restriction, $\mathbf{p}'\mathbf{1} = 1$. Formally, the ME estimator is given by

$$\operatorname{argmax}_{\mathbf{p}} \{-\mathbf{p}' \ln \mathbf{p}\}, \tag{4}$$

subject to the model consistency and additivity constraints,

$$\begin{cases} \mathbf{y} = \mathbf{X}\mathbf{p} \\ \mathbf{1}'\mathbf{p} = 1 \end{cases}. \tag{5}$$

There is no closed-form analytical solution, but a numerical approximation can be obtained using the Lagrange multipliers. It can be said that the Jaynes maximum entropy formalism has enabled us to solve the pure inverse problem with this optimization (maximization) procedure, regarding it as an inference problem. The ME principle is the basis for transforming the information in the data into a probabilistic distribution that reflects our uncertainty about individual outcomes.

To extend the ME estimator to the linear regression model represented by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{6}$$

where, as usually, \mathbf{y} denotes a ($N \times 1$) vector of noisy observations, $\boldsymbol{\beta}$ is a ($K \times 1$) vector of unknown parameters, \mathbf{X} is a known ($N \times K$) matrix of explanatory variables, and \mathbf{e} is the ($N \times 1$) vector of random disturbances (errors), Golan, Judge and Miller, [6], considered each β_k as a discrete random variable with a compact support and $M \geq 2$ possible outcomes and each e_n as a finite and discrete random variable with $J \geq 2$ possible outcomes. The error vector is considered here as another vector of unknown parameters to be estimated simultaneously with the vector $\boldsymbol{\beta}$. In this context, the linear regression model is represented as

$$\mathbf{y} = \mathbf{X}\mathbf{Z}\mathbf{p} + \mathbf{V}\mathbf{w}, \tag{7}$$

where

$$\boldsymbol{\beta} = \mathbf{Z}\mathbf{p} = \begin{bmatrix} z'_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & z'_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & z'_K \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_K \end{bmatrix}, \tag{8}$$

and

$$\mathbf{e} = \mathbf{V}\mathbf{w} = \begin{bmatrix} \mathbf{v}'_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{v}'_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{v}'_N \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_N \end{bmatrix}. \quad (9)$$

Matrices \mathbf{Z} ($K \times KM$) and \mathbf{V} ($N \times NJ$) are the matrices of support values and vectors \mathbf{p} ($KM \times 1$) and \mathbf{w} ($NJ \times 1$) are the vectors of unknown probabilities to be estimated. Thus, the generalized maximum entropy (GME) estimator is given by

$$\underset{\mathbf{p}, \mathbf{w}}{\operatorname{argmax}} \{ -\mathbf{p}' \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w} \}, \quad (10)$$

subject to the consistency (with the model) and additivity (for \mathbf{p} and \mathbf{w}) constraints,

$$\begin{cases} \mathbf{y} = \mathbf{XZ}\mathbf{p} + \mathbf{V}\mathbf{w}, \\ \mathbf{1}_K = (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}, \\ \mathbf{1}_N = (\mathbf{I}_N \otimes \mathbf{1}'_J)\mathbf{w}, \end{cases} \quad (11)$$

where \otimes represents the Kronecker product. The optimal probability vectors, $\hat{\mathbf{p}}$ and $\hat{\mathbf{w}}$, are used to obtain point estimates of the unknown parameters and the unknown errors with $\hat{\boldsymbol{\beta}} = \mathbf{Z}\hat{\mathbf{p}}$ and $\hat{\mathbf{e}} = \mathbf{V}\hat{\mathbf{w}}$.

3 Large-Scale Data and Aggregation

Large-scale data or big data usually refers to datasets that are large in different ways: many observations, many variables (or both); observations are recorded in different time regimes or are taken from multiple sources. Some difficult issues arise in dealing with this kind of data like, for instance, retaining optimal (or, at least, reasonably good) statistical properties with a computationally efficient analysis; or dealing with inhomogeneous data that does not fit in the classical framework: data is neither i.i.d. (exhibiting outliers or not belonging to same distribution) nor stationary (time-varying effects may be present).

Standard statistical models (linear or generalized linear models for regression or classification) fail to capture inhomogeneity structure in data, compromising estimation and interpretation of model parameters, and, of course, prediction. On the other hand, statistical approaches for dealing with inhomogeneous data (such as varying-coefficient models, mixed effects models, mixture models or clusterwise regression models) are typically very computationally cumbersome.

Ignoring heterogeneity in data, computational burden can be addressed with the following procedure, [7]: firstly, construct g groups from the large-scale data (groups may be overlapping and may not cover all observations in the sample); then, for each group compute an estimator, $\hat{\boldsymbol{\beta}}_g$, through standard techniques (e.g., OLS, ridge or LASSO); finally, considering the ensemble of estimators, aggregate them into a single estimator, $\hat{\boldsymbol{\beta}}$.

3.1 Traditional Aggregation Procedures

Several aggregation procedures have been already proposed in literature. Three of them are presented next.

1. Bagging: this procedure results in less computational complexity and even allows for parallel computing. It simply averages the ensemble estimators with equal weight to obtain the aggregated estimator, [8]:

$$\hat{\beta} := \sum_{g=1}^G w_g \hat{\beta}_g, \tag{12}$$

where $w_g = \frac{1}{G}$ for all $g = 1, 2, \dots, G$. The estimates $\hat{\beta}_g$ are obtained from bootstrap samples, where the groups are sampled with replacement from the whole data. It is a simple procedure and the weights do not depend on the response \mathbf{y} , but it is not suitable for inhomogeneous data.

2. Stacking: instead of assigning a uniform weight to each estimator, [9] and [10] proposed the aggregated estimator

$$\hat{\beta} := \sum_{g=1}^G w_g \hat{\beta}_g, \tag{13}$$

where

$$\mathbf{w} := \operatorname{argmin}_{\mathbf{w} \in \mathbf{W}} \left\| \mathbf{y} - \sum_{g=1}^G w_g \hat{\mathbf{y}}_g \right\|_2, \tag{14}$$

and, using a ridge constraint, $\mathbf{W} = \{\mathbf{w} : \|\mathbf{w}\| \leq s\}$, for some $s > 0$, or using a sign constraint, $\mathbf{W} = \{\mathbf{w} : \min_g w_g \geq 0\}$, or using a convex constraint, $\mathbf{W} = \{\mathbf{w} : \min_g w_g \geq 0 \text{ and } \sum_{g=1}^G w_g = 1\}$. The idea is to find the optimal linear or convex combination of all ensemble estimators, but it is also not suitable for inhomogeneous data.

3. Magging: corresponds to maximizing the minimally “explained variance” among all data groups, [7], such that

$$\hat{\beta} := \sum_{g=1}^G w_g \hat{\beta}_g, \tag{15}$$

where

$$\mathbf{w} := \operatorname{argmin}_{\mathbf{w} \in \mathbf{W}} \left\| \sum_{g=1}^G w_g \hat{\mathbf{y}}_g \right\|_2, \tag{16}$$

and $\mathbf{W} = \{\mathbf{w} : \min_g w_g \geq 0 \text{ and } \sum_{g=1}^G w_g = 1\}$. The idea is to choose the weights as a convex combination to minimize the $\|\cdot\|_2$ of the fitted values, $\hat{\mathbf{y}}$. If the solution is not unique, it is considered the solution with lowest $\|\cdot\|_2$ of

the weight vector among all solutions. This procedure was the first that we are aware of that was proposed for heterogeneous data. The main idea is that if an effect is common across all groups, then it cannot be “averaged away” by searching for a specific combination of the weights. The common effects will be present in all groups and will be retained even after the minimization of the aggregation scheme.

We believe the question as to whether the effects are really common across all groups may not be answered straightforwardly. If the groups carry information about the whole dataset and there are inhomogeneities, why should we consider that, with random sub-sampling, all groups are equally informative?

These considerations led us to the idea of choosing the groups according to their “information content”.

3.2 Proposed Aggregation Procedure

To measure the information content in a system and to measure the importance of the contribution of each piece of data or constraint in reducing uncertainty, Golan, Judge and Miller, [6], stated that, in the ME formulation, the maximum level of entropy-uncertainty results when the information-moment constraints are not enforced and the distribution of probabilities over the K states is uniform. As each piece of effective data is added, there is a departure from the uniform distribution, which implies a reduction of uncertainty. The proportion of the remaining total uncertainty is measured by the normalized entropy (NE),

$$S(\hat{\mathbf{p}}) = -\frac{\sum_k \hat{p}_k \ln \hat{p}_k}{\ln(K)}, \quad (17)$$

where $S(\hat{\mathbf{p}}) \in [0, 1]$ and $\ln(K)$ represents maximum uncertainty (the entropy level of the uniform distribution with K outcomes). A value $S(\hat{\mathbf{p}}) = 0$ implies no uncertainty and a value $S(\hat{\mathbf{p}}) = 1$ implies perfect uncertainty. Related to the normalized entropy, the information index (II) is defined as $1 - S(\hat{\mathbf{p}})$ and measures the reduction in uncertainty.

In this work, we propose a new aggregation scheme that is based on identifying the information content of a given group through the calculation of the normalized entropy. The proposed NE aggregated estimator is then

$$\hat{\boldsymbol{\beta}} := \sum_{g=1}^G w_g \hat{\boldsymbol{\beta}}_g, \quad (18)$$

where w_g is defined by normalized entropy using GME,

$$S(\hat{\mathbf{p}})_g = \frac{-\hat{\mathbf{p}}' \ln \hat{\mathbf{p}}}{K \ln M}, \quad (19)$$

for the signal, $\mathbf{X}\boldsymbol{\beta}$, such that $\sum_{g=1}^G w_g = 1$. This aggregation procedure is a weighted average of the collection of regression coefficient estimates as in Bagging, Stacking and Madding. The idea is almost as simple as Bagging and it is

expected to provide similar results if the data is homogeneous. However, since the weights in (18) will depend on the information content of each group according to (19), or some function of it, the weights will be, in general, non-uniform (as in Stacking and Madding) if the data is inhomogeneous.

Following section reports some simulated situations for which the NE aggregated estimator was calculated and compared to the aggregated estimator based on Bagging.

4 Simulation Study

A linear regression model was considered, where \mathbf{X} is the simulated matrix of explanatory variables, drawn randomly from normal distributions; $\boldsymbol{\beta}$ is a vector of parameters, \mathbf{e} is the vector of random disturbances, drawn randomly from normal distributions and \mathbf{y} is the constructed vector of noisy observations. For this simulation, $\boldsymbol{\beta}$ was considered as

$$\boldsymbol{\beta} = [1.8, 1.2, -1.4, 1.6, -1.8, 2.0, -2.0, 0.2, -0.4, 0.6, 0.8] \quad (20)$$

and two cases for the error distributions were considered, both with mean value zero and different standard deviations. Necessary reparameterizations were done considering $M = 5$ and $J = 3$ and different matrices \mathbf{Z} containing the supports for the parameters. The support matrix \mathbf{V} containing the supports for the errors, was set considering symmetric and zero-centred supports using the three-sigma rule with the empirical standard deviation of the noisy observations.

4.1 Simulation – Part I

Initial simulations were done with small size data, considering \mathbf{X} a (100×11) matrix; $\boldsymbol{\beta}$ a (11×1) vector; \mathbf{e} a (100×1) vector and \mathbf{y} a (100×1) vector. Two error distributions were considered, both with mean value zero, and standard deviations of 1 and 5. Random sub-sampling with replacement was done considering 5 groups and different number of observations per group. The Euclidean norm of the difference between the aggregated estimator $\hat{\boldsymbol{\beta}}$ and the true parameter $\boldsymbol{\beta}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$, is calculated for each simulated case and the results are given in Tables 1 – 4. For each case, three different solutions are presented, namely,

1. NE1: the chosen $\hat{\boldsymbol{\beta}}$ corresponds to the GME estimate for the group with lower normalized entropy, (NE). This solution does not correspond, in fact, to an aggregated estimator; it corresponds to a chosen estimate amongst all groups;
2. NE2: the chosen $\hat{\boldsymbol{\beta}}$ corresponds to the weighted average of the GME estimates of all groups, weighted by the information index (1-NE);
3. Bgg: the $\hat{\boldsymbol{\beta}}$ chosen corresponds to Bagging (mean average of the OLS estimates of all groups).

Table 1. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-10, 10]$.

n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
20	NE1	1.6217	4.2507
	NE2	1.7188	3.9252
	Bgg	3.5027	4.1104
40	NE1	1.6016	4.3768
	NE2	1.5754	4.2818
	Bgg	3.4736	4.1026
60	NE1	1.4911	4.7092
	NE2	1.5190	4.6012
	Bgg	3.8703	4.0965
80	NE1	1.5290	4.7649
	NE2	1.4300	4.7525
	Bgg	3.6974	4.1030
100	NE1	1.6484	4.8736
	NE2	1.5150	4.8415
	Bgg	3.7622	4.0932

Table 2. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-8, 8]$.

n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
20	NE1	1.8125	3.0865
	NE2	2.0742	3.4649
	Bgg	3.5072	4.1191
40	NE1	1.6205	3.8020
	NE2	1.6986	4.0805
	Bgg	3.8165	4.0992
60	NE1	1.5918	4.2832
	NE2	1.5365	4.2697
	Bgg	3.6198	4.1058
80	NE1	1.8352	4.5161
	NE2	1.4995	4.5134
	Bgg	3.6514	4.0947
100	NE1	1.5487	4.6573
	NE2	1.4779	4.5995
	Bgg	3.6136	4.0915

It can be concluded that, as the amplitude of the support vectors diminishes, $\|\hat{\beta} - \beta\|_2$ tends to be higher for all the cases where the aggregated estimator is given through Bagging. For the higher amplitude support vectors, there are some situations for which the “best” aggregated estimator (in terms of having lower Euclidean norm) is the one provided by Bagging, but, as the amplitude of the mentioned intervals gets smaller, these situations tend to disappear and the best solution comes from one of the normalized entropy methodologies.

Table 3. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-6, 6]$.

n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
20	NE1	2.1479	3.3230
	NE2	2.3099	3.5611
	Bgg	3.7682	4.0640
40	NE1	1.6081	4.0060
	NE2	1.7206	3.6234
	Bgg	3.4006	4.0931
60	NE1	1.4676	3.9680
	NE2	1.6202	3.8246
	Bgg	3.6237	4.1073
80	NE1	1.4535	4.1885
	NE2	1.5945	4.0395
	Bgg	3.4946	4.0988
100	NE1	1.2722	4.1910
	NE2	1.5613	4.0918
	Bgg	3.8570	4.0964

Table 4. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-4, 4]$.

n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
20	NE1	2.6369	3.3976
	NE2	2.8533	3.7063
	Bgg	3.4096	4.1027
40	NE1	2.1150	3.4465
	NE2	2.2036	3.6406
	Bgg	3.9626	4.0986
60	NE1	1.9771	3.2674
	NE2	1.9576	3.4074
	Bgg	3.5776	4.1045
80	NE1	1.7218	3.1931
	NE2	1.7847	3.4290
	Bgg	3.7653	4.1020
100	NE1	1.7628	3.3152
	NE2	1.6954	3.4890
	Bgg	3.7286	4.1016

4.2 Simulation – Part II

Second part of the simulation study was done considering a bigger size data, where \mathbf{X} is a $(50\,000 \times 11)$ matrix; β is again a (11×1) vector; \mathbf{e} , a $(50\,000 \times 1)$ vector and \mathbf{y} , a $(50\,000 \times 1)$ vector. Two error distributions were considered, both with mean value zero, and standard deviations of 1 and 5. Random subsampling with replacement was done considering 5 and 10 groups and different number of observations per group. The Euclidean norm of the difference between

the aggregated estimator $\hat{\beta}$ and the true parameter β is calculated for each simulated case and the results are given in Tables 5 – 8, for the same three different aggregating schemes, NE1, NE2 and Bgg.

Table 5. Euclidean norm of the difference $\hat{\beta} - \beta$, with $z_k = [-10, 10]$.

n.g.	n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
5	50	NE1	1.6639	4.4905
		NE2	1.5669	4.4620
		Bgg	3.6644	4.1011
5	100	NE1	1.1021	4.9127
		NE2	1.4665	4.9026
		Bgg	3.6434	4.1043
5	200	NE1	1.1879	5.2478
		NE2	0.9633	5.2311
		Bgg	3.6170	4.1064
10	50	NE1	1.7925	4.6382
		NE2	1.6789	4.4868
		Bgg	4.2933	4.3083
10	100	NE1	1.0860	4.9707
		NE2	1.3152	4.9284
		Bgg	4.2305	4.3087
10	200	NE1	1.5453	5.2464
		NE2	1.4044	5.2143
		Bgg	4.2121	4.3092

From the second part of the simulation study, the same conclusion may be drawn: for lower amplitude support vectors, normalized entropy based aggregation schemes tend to provide better solutions, in what relates to $\|\hat{\beta} - \beta\|_2$.

5 Concluding Remarks

The idea of an aggregation procedure based on normalized entropy is promising as it is clear from the simulation study that there are situations where this approach provides very satisfactory solutions. Although not mentioned here due to space limitations, preliminary results with real data examples indicate that NE aggregation procedures can provide substantially lower mean squared error than Bagging. This observation suggests that further simulation analysis with different error structures or severe inhomogeneities may reveal substantial differences between normalized entropy aggregation schemes and Bagging, eventually penalizing the second. These analysis will be conducted in future work, along with investigation of other scenarios, such as the detection of zero coefficients, non-normal regressors, presence of high collinearity and other violations of the classical framework. Also, the comparison with Magging is a very important analysis that remains to be explored.

Table 6. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-8, 8]$.

n.g.	n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
5	50	NE1	1.8915	4.0802
		NE2	1.7639	4.1039
		Bgg	3.7762	4.1128
5	100	NE1	1.5698	4.7071
		NE2	1.6890	4.5564
		Bgg	3.9517	4.1040
5	200	NE1	1.2229	5.0010
		NE2	1.3650	4.9828
		Bgg	3.7637	4.0995
10	50	NE1	1.7169	4.2028
		NE2	1.7333	4.1304
		Bgg	4.2710	4.3080
10	100	NE1	1.5071	4.7164
		NE2	1.5930	4.6391
		Bgg	4.0948	4.3090
10	200	NE1	1.0548	4.9964
		NE2	1.3527	4.9798
		Bgg	4.1865	4.3077

Table 7. Euclidean norm of the difference $\hat{\beta} - \beta$, with $\mathbf{z}_k = [-6, 6]$.

n.g.	n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
5	50	NE1	1.7821	3.5938
		NE2	1.8027	3.5212
		Bgg	3.6412	4.0959
5	100	NE1	1.5745	4.0764
		NE2	1.6760	4.0109
		Bgg	3.6909	4.1040
5	200	NE1	1.3594	4.4428
		NE2	1.4045	4.4564
		Bgg	3.7164	4.1025
10	50	NE1	1.6363	3.3771
		NE2	1.7805	3.5501
		Bgg	4.3290	4.3102
10	100	NE1	1.4194	4.1365
		NE2	1.6203	4.0650
		Bgg	4.1384	4.3072
10	200	NE1	1.4564	4.5868
		NE2	1.5775	4.4612
		Bgg	4.2337	4.3101

Acknowledgments. This work was supported in part by the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnolo-

Table 8. Euclidean norm of the difference $\widehat{\beta} - \beta$, with $\mathbf{z}_k = [-4, 4]$.

n.g.	n.obs.g.	Solution	$e_i \sim N(0, 1)$	$e_i \sim N(0, 25)$
5	50	NE1	2.0097	3.1750
		NE2	2.0887	3.0208
		Bgg	3.8063	4.1012
5	100	NE1	1.7791	3.0203
		NE2	1.8030	3.0822
		Bgg	3.8743	4.1033
5	200	NE1	1.6830	3.3193
		NE2	1.7278	3.2865
		Bgg	3.6825	4.1020
10	50	NE1	2.0330	2.9980
		NE2	2.1206	3.1846
		Bgg	4.2913	4.3088
10	100	NE1	1.8520	3.1174
		NE2	1.8528	3.1217
		Bgg	4.2690	4.3079
10	200	NE1	1.6465	3.2994
		NE2	1.6905	3.3019
		Bgg	4.2192	4.3082

gia) through CIDMA – Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

References

- Golan, A.: On the State of Art of Info-Metrics. In: Uncertainty Analysis in Econometrics with Applications. Huynh, V.N., Kreinovich, V., Sriboonchitta, S., Suriya, K. (eds.), pp. 3–15. Springer-Verlag, Berlin (2013)
- Golan, A.: Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information. Oxford University Press, New York (2018)
- Golan, A.: On the Foundations and Philosophy of Info-Metrics. In: Cooper, S.B., Dawar, A., Lowe, B.L. (eds.) CiE2012. LNCS, vol. 7318, pp. 238–245. Springer-Verlag, Heidelberg (2012).
- Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. 106, 620–630 (1957)
- Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. 108, 171–190 (1957)
- Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics - Robust Estimation with Limited Data. John Wiley & Sons, Chichester (1996)
- Bühlmann, P., Meinshausen, N.: Magging: Maximin Aggregation for Inhomogeneous Large-Scale Data. In: Proceedings of the IEEE 104 (1): Big Data: Theoretical Aspects, pp. 126–135, IEEE Press, New York (2016)
- Breiman, L.: Bagging Predictors. Mach. Learn. 24, 123–140 (1996)
- Wolpert, D.: Stacked Generalization. Neural Netw. 5, 241–259 (1992)
- Breiman, L.: Stacked Regressions. Mach. Learn. 24, 49–64 (1996b)

Forecasting time series using topological data analysis

N. Gabdrakhmanova

People's Friendship University of Russia, Moscow, Russia

`gabd-nelli@yandex.ru`

Abstract. The dynamics of data traffic intensity is examined using traffic measurements at the interface switch input. The wish to prevent failures of trunk line equipment and take the full advantage of network resources makes it necessary to be able to predict the network usage. The research tackles the problem of building a predicting neural-net model of the time sequence of network traffic. Topological data analysis methods are used for data preprocessing. Nonlinear dynamics algorithms are used to choose the neural net architecture.

Topological data analysis methods allow the computation of time sequence invariants. The probability function for random field maxima cannot be described analytically. However, computational topology algorithms make it possible to approximate this function using the expected value of Euler's characteristic defined over a set of peaks. The expected values of Euler's characteristic are found by constructing persistence diagrams and computing barcode lengths. A solution of the problem with the help of R-based libraries is given. The computation of Euler's characteristics allows us to divide the whole data set into several uniform subsets. Predicting neural-net models are built for each of such subsets. Whitney and Takens theorems are used for determining the architecture of the sought-for neural net model. According to these theorems, the associative properties of a mathematical model depend on how accurate the dimensionality of the dynamic system is defined. The sub-problem is solved using nonlinear dynamics algorithms and calculating the correlation integral. The goal of the research is to provide ways to secure the effective transmission of data packets.

Keywords: computational topology, persistence, stability, neural network

1 Introduction

The topicality of the study is determined by the following reasons. The continuing development of telecommunication and Internet services sets new requirements for the bandwidth of telecommunication channels. The presence of a great deal of various services in a single physical transmission medium at peak hours can bring about the overloading of switching and routing devices in trunk lines and, as result, a reduction of many services. The wish to prevent failures of trunk line equipment and take the full advantage of network resources makes the problem of effective use of the telecommunications channel bandwidth very important (the direct widening of the bandwidth inevitably leads to an increase of service costs). It is necessary to have effective traffic control methods that could use statistical data to predict the traffic intensity. A

lot of modern publications deal with mathematical models of different types of network traffic [1, 2, 3]. The complexity and relevance of this problem urge further research in the field.

2 The topological data analysis

The topological data analysis is a new theoretical trend in the field of data analysis. The approach allows the determination of topological data structures. Recent advancements in the field of computational topology make it possible to find topological invariants in data collections [4, 5, 6].

The point of the analysis is that stable properties are to be immune to noise, distortions, errors, lack of data. The practice of using the analysis in different fields shows that the supposition is true and stable topological properties can provide a lot of information about data collections. Persistence diagrams are one of basic tools of computational topology. They make it possible to get useful information about excursion sets of a function. Below are the basic definitions (according to [4]).

Let X be a topological space being triangulated, f be a continuous tame function defined over space X . Let us introduce the notation $U_a = f^{-1}(-\infty, a]$ for $a \in \mathbb{R}$. When moving upwards, components U_a can merge or produce new components. It is possible to trace how the sub-level topology changes with a by examining homologies of these sets with, say, persistence homologies. Parameter $a \in \mathbb{R}$ is called the homological critical value if for certain k the homomorphism induced by nesting $f_* : H_k(U_{a-\varepsilon}) \rightarrow H_k(U_{a+\varepsilon})$ is not an isomorphism for any sufficiently small $\varepsilon > 0$

(homology groups are considered with coefficients in \mathbb{Z}_2). Continuous function f is called tame function if it has a finite number of homological critical values. When $b \leq a$, then $U_b \subseteq U_a$. Let us denote a set of connectivity components as $C(a) = C(U_a)$. It is possible to define a functional – Euler characteristic – over a set of sub-levels of U_a .

Let $X \subset \mathbb{R}^2$. Then, in the terms of algebraic topology, Euler's number is $\chi(U_a) = \beta_0 - \beta_1$, where β_0, β_1 are the ranks of the first two homology groups. Functional $\chi(U_a)$ measures the field topological complexity on the sub-level set. Note that for function f it is possible to deal with a set of super-levels $U_a = f^{-1}[a, \infty)$ instead of sub-levels.

Let us define the persistence diagram according to [5]. Let $f: X \rightarrow \mathbb{R}$ be a tame function. Let $a_1 < a_2 < \dots < a_n$ be critical homological values. Let us consider interjacent values $b_0, b_1, \dots, b_n : b_{i-1} < a_i < b_i$. Let us supplement the chosen points in the following way: $b_{-1} = a_0 = -\infty; b_{n+1} = a_{n+1} = +\infty$.

Let us define the multiplicity of point (a_i, a_j) for each couple of indices $0 \leq i < j < n+1$ by setting $\mu_i^j = \beta_{b_{i-1}}^{b_j} - \beta_{b_i}^{b_j} + \beta_{b_i}^{b_{j-1}} - \beta_{b_{i-1}}^{b_{j-1}}$, where $\beta_x^y = \dim(\text{Im}(f_x^y))$, $f_x^y : H_k(U_x) \rightarrow H_k(U_y)$. Persistence diagram $D(f) \subset \mathbb{R}^2$ of function f stands for a set of points (a_i, a_j) ($i, j = 0, \dots, n$).

1, ..., n+1) adjusted for multiplicity μ_i^j in combination with a set of diagonal points $\Delta = \{(x, x) | x \in R\}$ adjusted for infinite multiplicity.

The immunity of a persistence diagram to perturbations of function f is its remarkable feature. Persistence diagrams can be used to calculate the lengths of the barcodes of connectivity components. Here the term barcode stands for the component lifetime. Let us denote the summarized lengths of barcodes of two homology groups H_0 and H_1 as L_0 and L_1 correspondingly. Then the mean of the Euler characteristic can be determined [6] as

$$\chi = L_0 - L_1. \quad (1)$$

3 Setting the problem

A second-level interfacial switch of a backbone line provider is taken as a test object in the paper. The traffic coming to each port of the switch is integrated traffic from user groups belonging to a particular region. The explanatory drawing is given in Figure 1. The Cacti software (SNMP interface protocol) was used to gather statistic data. The information about the degree of network usage is more useful in practice. The knowledge of the number of packets in unit time can be misleading. For this reason the aggregate quantity $x(t)$ – traffic intensity (in bits) at moment t – is taken as an observable variable. The extension of data is 10080 points or 7 days. The plot of traffic intensity measured at port GE 0 is shown in Figure 2. Each point in this plot represents a number of bits going through the trunk in one minute's time.

So the goal is to construct a mathematical model for the m -step prediction of traffic intensity using observations $\{x(t), t = 1, 2, \dots, N\}$, where N is the number of points. The estimates of Euler's characteristics are used here as indication of network usage. The following algorithm is proposed. The whole data collection is to be divided in clusters with different Euler's characteristics. A neural-net prediction model is to be built for each cluster using nonlinear dynamics methods. Below is the result of the experimentation.

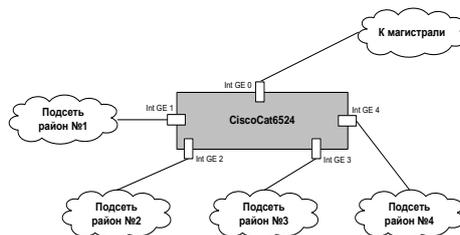


Fig. 1. The measurement arrangement.

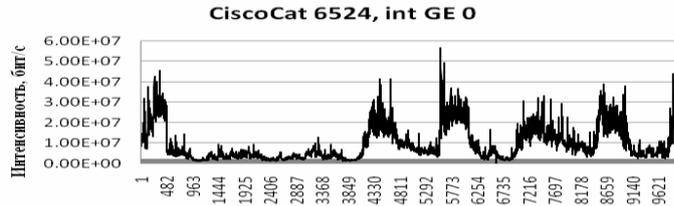


Fig. 2. The traffic intensity plot at port GE 0.

4 Topological invariants calculated for a traffic intensity sequence

Packet TDA from a public repository of R packets was used as a library for finding stable homologies. The packet has a broad toolkit for topological data analysis by topological methods.

Before finding topological characteristics, the whole data collection was divided in some portions. Each portion held data acquired in two hours' time. For each portion persistence diagrams, barcodes were determined and Euler's characteristic estimates were calculated by formula (1).

The following algorithm was used to find estimates of Euler's characteristic in the TDA packet. A triangulation grid was first built using function `Grid()`. Then function `gridDiag` was used to produce matrix `Diag`. Function `gridDiag` evaluates the actual value of the function by the triangulation grid, generates simplex filtration using these values, and calculates constant homologies from the filtration. Figure 3 shows the persistence diagrams for one portion of data. The birth time of a component is plotted as abscissas; the death time is plotted as ordinates. The dots correspond to zero-dimensional simplexes, the triangles mark single-dimensional simplexes. Figure 4 presents the barcode chart of zero-dimensional simplexes. Table 1 gives the estimates computed for different ($n = 15$) portions of the object. The following notation is used in the table: n is the number of a portion (interval), L_0 and L_1 are the summarized barcode lengths of zero- and single-dimensional simplexes, χ is the estimate of Euler's characteristic (1). The plot in Figure 5 shows Euler's characteristic as function of n . The horizontal axis represents the number of an interval and Euler's characteristic is measured on the vertical axis.

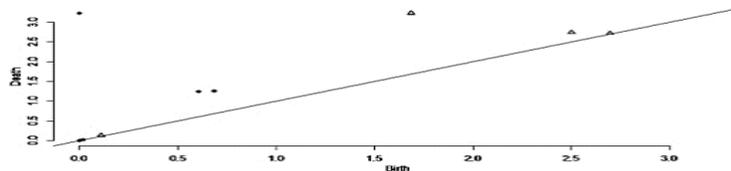


Fig. 3. The plot on the right shows the persistence diagram of the superlevel sets of the KDE.

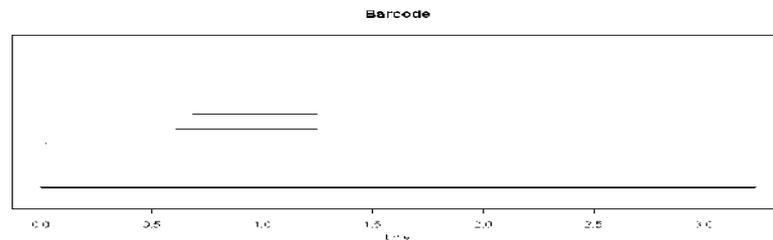


Fig. 4. Barcode

Table 1. The estimates of Euler's characteristic

n	1	2	3	4	5	6	7	8	9	10	11	12
L_0	3,7	4,1	3,6	3,7	2,2	1,7	2,0	2,0	1,8	2,4	3,4	3,6
L_1	2,3	1,6	1,7	1,8	2,5	1,7	2,0	2,0	1,5	3,3	1,8	2,3
χ	1,5	2,6	1,9	1,8	-0,3	-0,1	0,03	0,04	0,3	-0,8	1,5	1,4

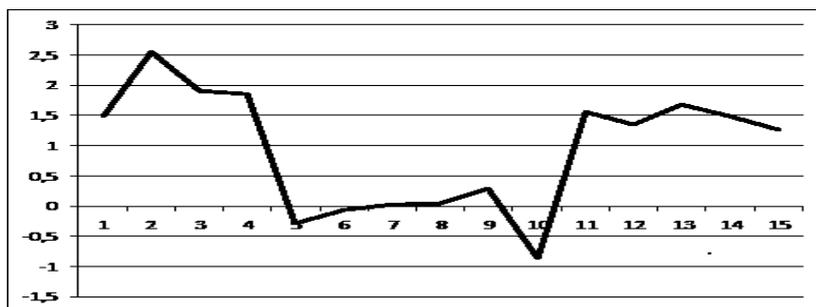


Fig. 5. Euler's characteristic as function of n .

The results prove that Euler's numbers are a stable characteristic of traffic intensity. At the next stage the portions with the same $[\chi]$ (where $[\cdot]$ is the integer of a number) are united in a single cluster. A neural-net prediction model is built for each cluster.

5 Building the neural-net model of the data

Methods of nonlinear dynamics are used to construct a neural-net model for a selected cluster. The subproblem is set as follows. Let $\{x(t)\}_{t=1}^N$ be measurements of a particular observable scalar component of a d_1 -dimensional dynamic system \bar{y} . On the whole, the dimensionality and behavior of the dynamic system are not known. For a given time sequence it is necessary to build a model that would incorporate the dy-

namics responsible for the generation of observations $x(t)$. According to Takens' theorem, the geometrical structure of the dynamics of a multivariable system can be restored using observations $\{x(t)\}_{t=1}^N$ in a D -dimensional space built around new vector $\bar{z}(t) = \{x(t), x(t-1), \dots, x(t-(D-1))\}^T$ (where $D \geq 2d_1 + 1$). The evolution of points $\bar{z}(t) \rightarrow \bar{z}(t+1)$ in the restored space corresponds to the evolution of points $\bar{y}(t) \rightarrow \bar{y}(t+1)$ in the initial space. The procedure of searching for a suitable D is called nesting. The least value of D at which the dynamic restoration is achieved is called the dimension of the nesting. The algorithm offered by P. Grassberger and I. Procaccia in 1983 makes it possible to evaluate D using a time sequence.

After D is estimated, the problem at hand can be formulated in the following way. There is time series $\{x(t)\}_{t=1}^N$ and restoration parameters ($D = 11$ in our case) are set. For N_1 vectors $\bar{z}(t) = \{x(t), x(t-1), \dots, x(t-(D-1))\}^T$ the values of the sought-for function $F(t) = F(\bar{z}(t))$ are known (because the terms of the time series following $\bar{z}(t)$ are known). It is necessary to find the value of the sought-for function at new point $\bar{z}(t)$, $x = F(\bar{z})$.

Neural nets of the multiple-layer perceptron type [7] are used to tackle the problem. Only the key results are given below. Figure 6 shows the graph of traffic intensity on a set of test points. The horizontal axis represents time, the vertical axis shows the normalized traffic intensity; the solid line corresponds to experimental data x , the dashed line represents theoretical results \hat{x} .

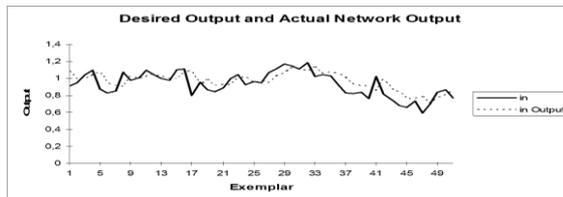


Fig. 6. Traffic intensity on a set of test points

6 Conclusions

The goal of the paper was to test the hypothesis that the use of the topological data analysis would make it possible to build traffic intensity prediction models due to finding additional characteristics that cannot be discovered by conventional data analysis. The data of network traffic intensity in a week's time were examined. The computations showed that the traffic intensity dynamics can be described by Betti numbers and Euler's characteristics. The algorithm using Euler's characteristics was used in the paper to build a model makes it possible to increase the prediction accuracy by an order of magnitude (as compared with methods not using Betti numbers). The

paper gives the results of first steps towards the application of topological data analysis for predicting the network traffic intensity. The results proved the prospectiveness of further research in the field.

References

1. Heyman D.P., Tabatabai A., Lakshman T.V. Statistical analysis and simulation study of video teleconference traffic in ATM networks // IEEE Transactions on Circuits and Systems for Video Technology. 1992. V.2.P.49-59.
2. M.F. Zhani, H. Elbiaze. Analysis and Prediction of Real Network Traffic. // Journal of Networks, Vol.4.No 9. November 2009.
3. Potapov A.B. Time-series analysis: when dynamical algorithms can be used // Proc. 5 Int. Specialist Workshop Nonlinear Dynamics of Electronic Systems. Moscow. June 26-27, 1997. P.388-393.
4. H. Edelsbunner, D.Letsscher, A.Zomorodian, Topological persistence and simplification//Discrete Comput.Geom.,28(2002),511-533.MR 1949898
5. G. Carlsson, A. Zomorodian, Computing persistent homology//In “proc.20th Ann.Sympos.Comput.Geom.,2004”, 347-356.
6. M.F. Zhani, H. Elbiaze. Analysis and Prediction of Real Network Traffic. // Journal of Networks, Vol.4.No 9. November 2009.
7. Simon Haykin. Neural Networks: A Comprehensive Foundation Second Edition, 2006. – 1104p.

Forecasting Subtidal Water Levels and Currents in Estuaries. Assessment of Management Scenarios

MÁ Reyes Merlo, R Siles-Ajamil* and M Díez-Minguito

Andalusian Institute for Earth System Research (IISTA), University of Granada
Avda. del Mediterraneo, s/n, E- 18006, Granada, Spain.

<http://www.springer.com/lncs>

Abstract. Floods are one of the most harmful extreme events occurring in estuaries, which are influenced by tides and freshwater discharges. The Guadalquivir estuary (SW Spain) has experienced flooding multiple times in recent years. In this estuary, high-resolution time-series from a long-term monitoring campaign are analyzed to assess the impacts of two different management strategies: the 15% reduction of discharges, and the 23% deepening of the navigation channel. Auto-regression models were applied to characterize subtidal water levels and currents. The best fit for levels corresponds to a linear superposition of tides and discharge, while the best fit for currents involves a non-linear interaction of both. The obtained relationships were used to assess, on a medium-term basis, the effect of the scenarios. The comparison between scenarios results and present conditions of the estuary revealed that subtidal levels decrease when the freshwater input reduces, and increase as a consequence to deepening the channel.

Keywords: Guadalquivir Estuary, flood risk, tides, river discharge, management

1 Introduction

Estuaries are identified as the most productive coastal systems; but are also one of the most vulnerable in environmental, social and economical terms. Faced with an increasing use and demands and under the scenario of global change and flood risk, estuaries are complex framework.

Nevertheless, since estuaries and human communities are in permanent feedback, the ability to model and predict both the natural and the stressed response to forcing agents, and their spatio-temporal shift, is an issue under permanent study. Two major forcings are considered for the simulations in the Guadalquivir River estuary (GRE): tidal motion and freshwater discharge.

The GRE is located in the SW Iberian Peninsula, and its waters mixes with those of the Gulf of Cádiz (Atlantic Ocean). The main tidal channel extends

* Corresponding author: reyess@ciccp.es

110 km converging from its only mouth at Snlucar de Barrameda to Alcal del Ro head weir, upstream from Seville. The GRE is navigable up to the Port of Seville, 85 km from the mouth. A minimum depth of 6.5 meters is maintained for navigational purposes. The channel is convergent. Tides are semidiurnal, with M2 (12.42 h) tides as the main tidal constituent.

During 2008-2011 a comprehensive research study was carried out with the purpose of regulating the different economic and environmental uses of the estuary [1]. This involved the installation of a remote, real-time monitoring network (RTMN) [2] which accounted for tidal gauges and current meters, amongst others.

The subtidal levels and currents are the variables analyzed. In the latter case, Section 2 applies a non-stationary statistical approach to representing its stochastic nature. Predictions of subtidal water levels and currents are performed with common regression models using the method of Monte Carlo, as presented in Section 3. Short-term analysis for subtidal elevations is also done in that Section.

2 River discharge: Non-stationary modeling

The objective of this characterization is to obtain the marginal distribution that represents the best the forcings behavior, used to simulate a long time series of such variable. In our case study, the variable is the river discharge in the Guadalquivir estuary, with noticeable seasonality. After the stationary analysis (Section 2.1), the non-stationary approach developed in [3] is applied (Section 2.2).

2.1 Stationary modeling

Usual and mixture models for the marginal distribution of the river discharge are implemented. Usual distributions are the exponential (EX), Lognormal (LN), Weibull (WB) and Gamma (GM). The mixture models are intended to incorporate the central and extreme populations into a single model. In these, the central regime is a truncated distribution, where the upper and lower tails are represented by means of Generalized Pareto distribution (GPD). The analyzed resulting distributions are the LN-GPD and the WB-GPD [4]. The parameters of the mixture models are estimated by maximum likelihood. Kolmogorov-Smirnov test, with 5% significance level, assesses the goodness of fit. To avoid numerical inconsistencies with null discharge (4% of the data), and considering the precision of the measurements (10^{-2}), the time series is uniformly completed with values between $5 \cdot 10^{-3}$ and $1 \cdot 10^{-2}$ m³/s.

Fig. 1 (left panel) shows the empirical, as well as the modeled cumulative distribution function (cdf) of the river discharge. According to the test, the best fit corresponds to LN distribution. Fig. 1 (right panel) shows the cdf of LN-GPD and WB-GPD. The best distribution is the LN-GPD. Comparing the best usual and mixture distributions, the LN and LN-GPD, the later one improves the fit,

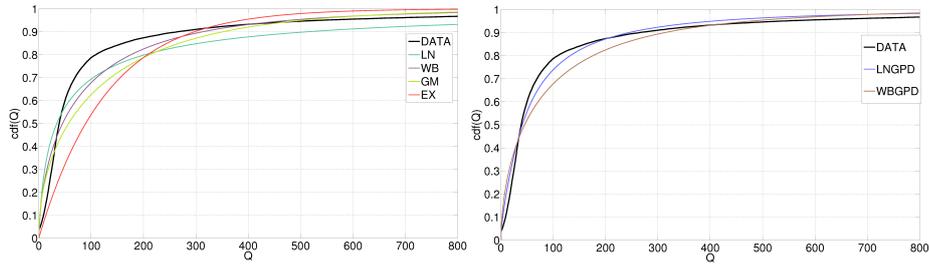


Fig. 1. Cdf for the river discharge with the usual models (left panel) and the mixture models (righth panel) described in Section 2.1.

especially in the tails. Nevertheless, none of them can capture the trend with low freshwater discharges. The poor fit with usual models and the strong seasonality observed in river discharge, justify the use of non-stationary distributions.

2.2 Non-stationary modeling

The non-stationary distribution that best fit the data is the LN-GPD-NE, that includes the seasonality in the parameters of the distribution using a Fourier time Series. For the discharges in the Guadalquivir, the model with minimum Bayesian Information Criterion has an order of approximation, for the Fourier series in the parameters $(\mu_{LN}, \sigma_{LN}, \xi_2)$, of $(4,2,2)$. The pdf and cdf considering the seasonality in the parameters of the distribution are modeled too. Although figure is not included, the improvement in the fit is noticeable.

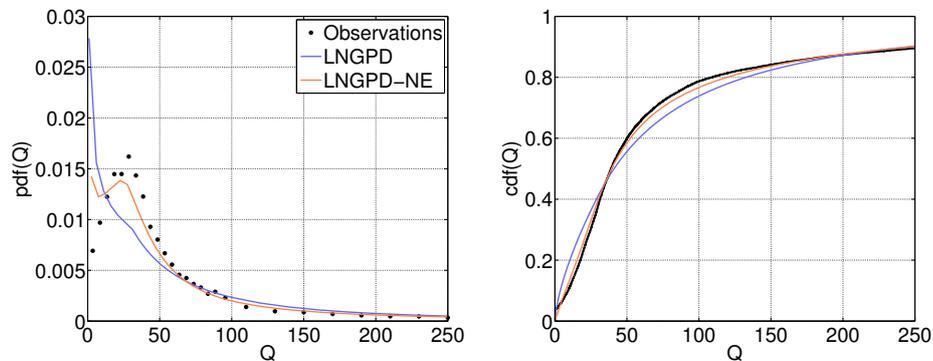


Fig. 2. Pdf and cdf with the LN-GPD and LN-GPD-NE distributions.

The results show the quantiles corresponding to the empirical accumulated probability values and those obtained when the LGN-GPD-NE. A moving window of one month was used to obtain the empirical quantiles. The lower-most

part of the tail is not well reproduced, for quantiles ≤ 0.1 ($\sim 10 \text{ m}^3/\text{s}$). The reason is that the system is highly regulated, and further considerations should be performed to overcome this situation. Nevertheless, the rest of the quantiles are well reproduced. Then, this is the selected model when performing the simulations in the rest of the sections.

3 Regression Models & Management Scenarios Simulations

3.1 Regression Models

Three regression models are applied and compared. The regression models corresponds to those detailed in: [5] (linear superposition of river discharge Q_d and tidal range H), [6] (accounts for non-linear interaction between Q_d and H) and Mixed (partial combination of both).

A regression analysis is carried out in the Guadalquivir estuary. Values are subtidally averaged for a time step of 25 hours. Only the currents projected along the channel are fitted. [5] used a linear combination of tidal contribution (tidal range) and water level variation due to river discharge.

$$\begin{aligned}\bar{\eta}_G &= s_{\bar{\eta},1}^G H + s_{\bar{\eta},2}^G Q_d + s_{\bar{\eta},3}^G, \\ \bar{u}_G &= s_{\bar{u},1}^G H + s_{\bar{u},2}^G Q_d + s_{\bar{u},3}^G.\end{aligned}\quad (1)$$

This regression is applied to all the tidal gauges and current meters installed in the Guadalquivir using the local tidal range and the Q_d realised from the Alcalá del Río dam. Water level and current observations were recorded between 2008 and 2011 by eight tidal gauges, β_i , and six current meters, α_i , located at different stations along the main channel, [2]. [6] used the following regression (non-linear scaling):

$$\begin{aligned}\bar{\eta}_K &= s_{\bar{\eta},1}^K H_0^2 Q_d^{-4/3} + s_{\bar{\eta},2}^K Q_d^{2/3} + s_{\bar{\eta},3}^K, \\ \bar{u}_K &= s_{\bar{u},1}^K H_0^2 Q_d^{-4/3} + s_{\bar{u},2}^K Q_d^{2/3} + s_{\bar{u},3}^K,\end{aligned}\quad (2)$$

where $s_{\bar{\eta},k}^G$ and $s_{\bar{\eta},k}^K$ in Eq. 1 and 2 are fitted coefficients, and H_0 is the tidal range (station β_0). The variables are referred to the same time step.

In these expressions, the exponents correspond to the cited authors. The following expression, designed as mixed, is proposed to combine both models without indiscriminately increasing the number of parameters:

$$\begin{aligned}\bar{\eta}_M &= s_{\bar{\eta},1}^M H + s_{\bar{\eta},2}^M Q_d + s_{\bar{\eta},3}^M + s_{\bar{\eta},4}^M H^{s_{\bar{\eta},5}^M} Q_d^{s_{\bar{\eta},6}^M}, \\ \bar{u}_M &= s_{\bar{u},1}^M H + s_{\bar{u},2}^M Q_d + s_{\bar{u},3}^M + s_{\bar{u},4}^M H^{s_{\bar{u},5}^M} Q_d^{s_{\bar{u},6}^M}.\end{aligned}\quad (3)$$

This mixture model intends to find consistent exponents, in the nonlinear term, that better represent the GRE. The fitted parameter are model performance S_k and correlation coefficient R . Fig. 3 shows the observed and the predicted subtidal elevation in instruments β_S , β_5 , β_1 and β_0 . The figure also shows the observed and the predicted subtidal currents in instruments α_5 and α_0 .

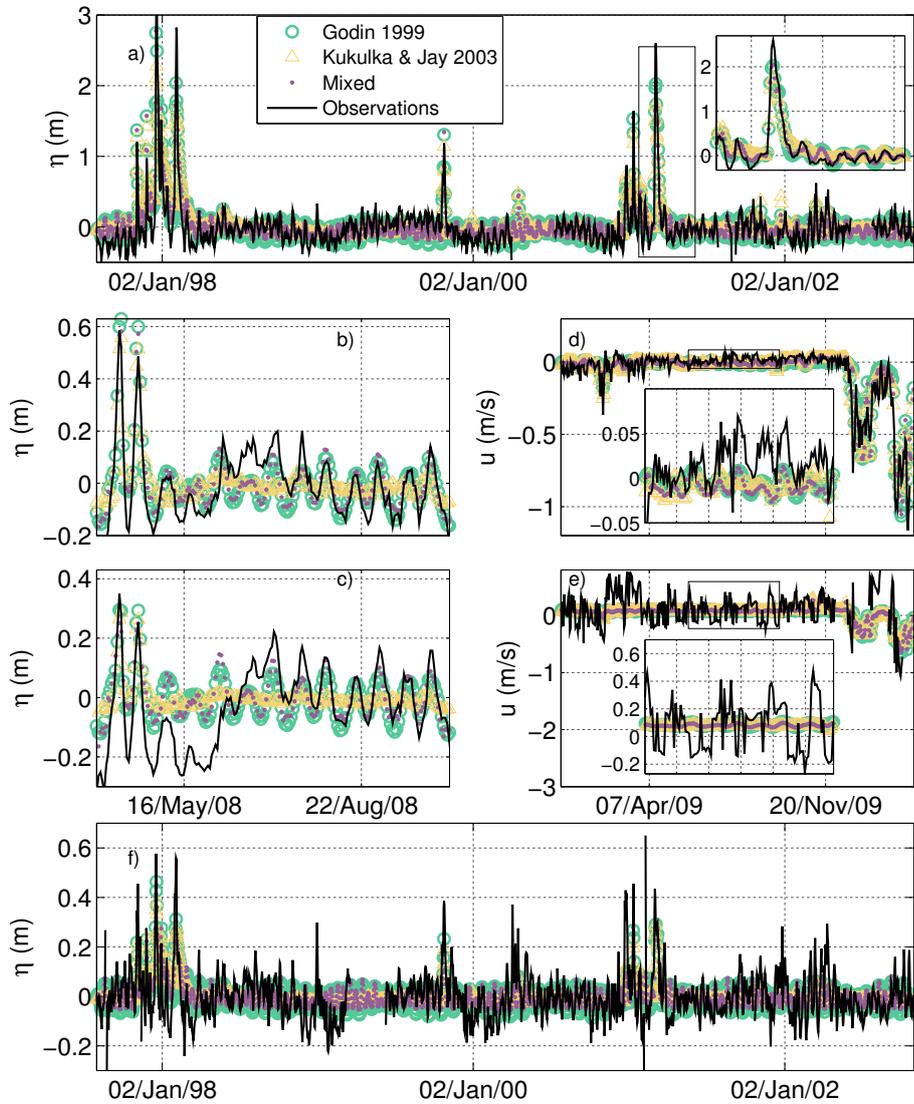


Fig. 3. Panels (a) and (f): Calibration and validation at β_5 and β_0 , respectively, of [5] (green circles), [6] (orange triangles) and Mixed model (purple dots) regression formulae against subtidal water level obtained from observations (black curve). Panels (b) and (c): Fit of these models for water levels at β_5 and β_1 , respectively. Panels (d) and (e): Fit of these models for subtidal currents at α_5 and α_0 , respectively. Insets zoom into rectangled areas. Symbols are greater than the 95% error confidence interval.

Overall, the models reproduce better the signal at the landward-most stations. The fit coefficients with Godin and Kukulka formulation are same order of magnitude between instruments, e.g.: $s_{\eta,1}^G$ for is $\sim 10^{-1}$. This consistency is not observed in the Mixed model. Results suggest that, despite the improvement in the model performance with the Mixed model, the understanding of the dynamics, from a global point of view, is lost. For the sake of simplicity (parsimony principle), the interpretations are done with the first two models.

The best fit for the elevations is achieved with [5] and Mixed models. Focusing on former and according to the coefficients, with low discharges the dynamic behavior of the subtidal elevation in the GRE is mostly controlled by the astronomical tidal range. With discharges around $400 \text{ m}^3/\text{s}$ and higher, the situation changes. In this case, the results are in agreement with [7], where the subtidal elevations can be linearly related to the fresh water discharge. The [6] fit seems to be not adequate with small discharges. Since there is no linear term with the tidal range, the subtidal behavior for the elevations cannot be reproduced: important spring-neap variations are dumped by the term $H_0^2/Q^{3/4}$. In addition, it should be noticed that the model considered the tidal range at the mouth, not local, which also affects the goodness of fit.

The pdf and cdf for instruments β_6 and β_0 are depicted in Fig. 4. When plotting the probability density and cumulative distribution functions with the observations and the models in Fig. 4, we obtained that the upper tails are better reproduced than the lower. This fact must be considered when evaluating the results of our simulations, for example, if we intend to assess the navigability of the channel.

Markov Chain technique was considered in the regression analysis with [5], [6] and Mixed formulation, in an attempt to improve the fit in low regime. The values of the new coefficients, which are not included, present the same features explained before, where the most important is the hysteretic or past-time term. The calibration improves, but the validation is almost the same (values and figure not included).

The overall performance in the validation is similar, though slightly lower than without Markov Chain ($\sim 10^{-2}$). The proposed regression models does not significantly improve the fit. The higher subtidal elevations and currents time step (25 h), that possibly attenuate or includes the past effect, can explain this fact. Also, changing the way to consider the inertia or the variables of the models, could improve the results, such as including the wind and other lateral effects as the secondary circulation. Either way, from this point, Markov Chain is not considered with subtidal elevations and currents.

3.2 Management Scenario Simulations

The proposed regression models were used to assess the effects of different scenarios into the subtidal water levels, from short- to mid- temporal scales. The mid-term simulations consider four different scenarios, related to the reduction of the freshwater discharge and the increase of the water depth after a dredging intervention. Short term simulation assesses the variation of the subtidal level

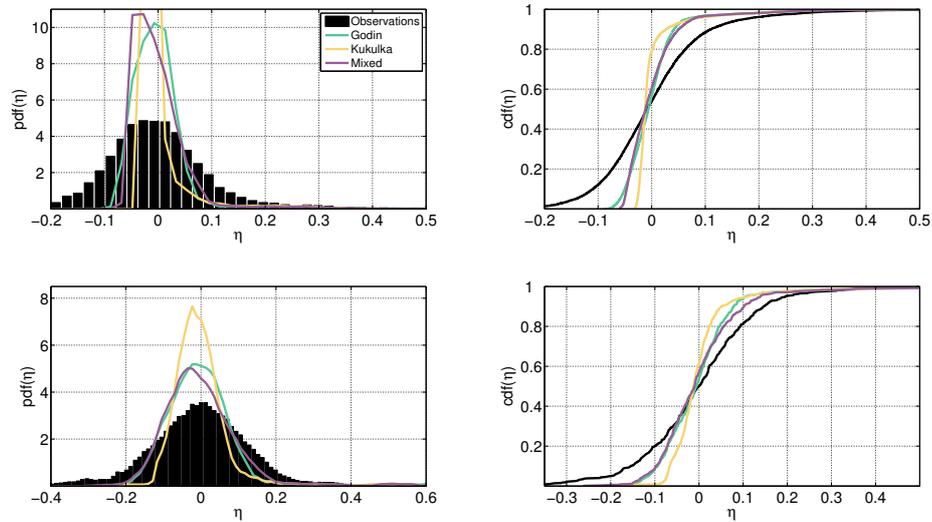


Fig. 4. Pdf and cdf for subtidal elevations in β_6 (first row) and β_0 (second row) for the observed data (black color), [5] (green line), [6] (orange line) and Mixed model (purple line).

after a peak discharge, supposing both the present state of the system and the response, to the same forcing, after the dredging intervention. The coefficients of the regressions models are kept fixed during the simulations.

Mid-term Simulations Subtidal water levels depend on the freshwater discharge and the astronomical tidal range. For the different scenarios, 50 simulations with 25 years of duration each, were run. River discharges were stochastically simulated with the method of Monte Carlo using the non-stationary distribution. A simple one-dimensional tidal model ([8]) assessed the changes induced by the dredging intervention (rise in the water depth). In this model, friction and the rest of parameters are supposed to remain as in the present situation. Thus, the shift in the tidal motion was applied by changing the amplitude and phase of the semidiurnal component M_2 .

The considered scenarios are as follow:

- S1: Similar discharge (Q_d) distribution with the current water depth (h), ($\sim Q_d, \sim h$).
- S2: Decrease in the freshwater discharge by a 15%, according to [9]. The water depth does not change ($\downarrow Q_d, \sim h$).
- S3: Increase in the water depth, from 7 to 8.5 m, according to the Port Authority of Seville. The fresh water regime does not change ($\sim Q_d, \uparrow h$).
- S4: Combination of S2 and S3 ($\downarrow Q_d, \uparrow h$).

Fig. 5 depicts the pdf and cdf in β_S with the different models.

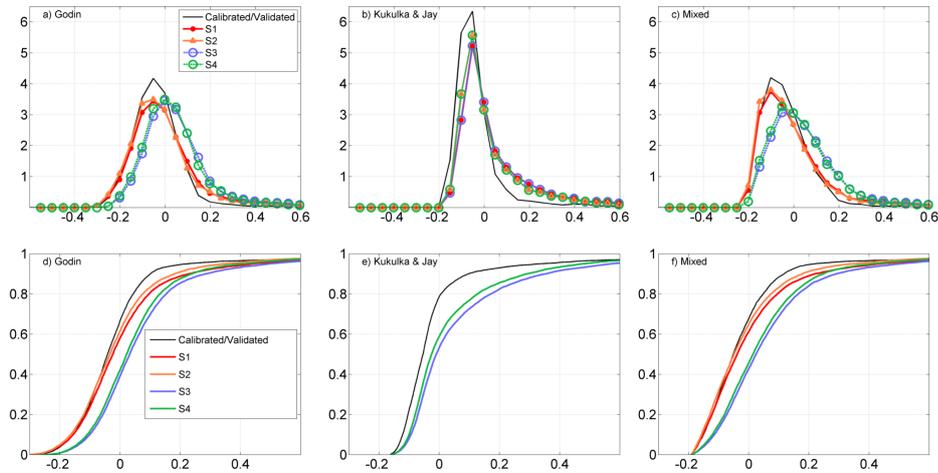


Fig. 5. Pdf (first row) and cdf (second row) for subtidal elevations in β_S for the calibration and validation period (black line) and scenarios S1 (red), S2 (orange), S3 (blue) and S4 (green) for the regression models. In [6] (second column), S1 overlaps S3, and S2 does the same with S4.

[5] and Mixed model, despite different values, present similar behavior (Fig. 5 panels a-c, d-f). As expected with these models, S1 (red line) reproduces better the observed values (black line) in this instrument than with [6]. With this formulation, S1 overlaps S3, as well as S2 and S4. The reason is that the authors built their model with the astronomical tidal range at the mouth station. Since the one-dimensional tidal model fixes the harmonics in the mouth after the dredging, [6] can only capture a variation in the freshwater regime. In all scenarios with all models but with the mixed one, subtidal water level distributions do not have marked changes in the dispersion and tails behavior in comparison to the simulation of present state S1.

Fig. 6 displays the mean values of the subtidal levels for each scenario with the regression models. Comparing the scenarios with the current simulated situation S1, we observe that the subtidal levels decrease when the freshwater input reduces (S2), and increase as a response to deepening the navigation channel (S3). The mixed scenario (S4) is in between S2 and S3. According to the proposed values for S4 and in comparison to S1, it is possible to locate where the subtidal elevations decrease (lower stretch) or increase (upper stretch).

Short-Term Simulations This section assesses the effects of a flooding in the subtidal water levels. The purpose is two-fold: firstly, to show that the presented regression models can be used as an early warning system since managers can control the river discharge released by the dam; and secondly, evaluate the response of the estuary against a flooding with or without the dredging intervention. [5] is the used model in this section.

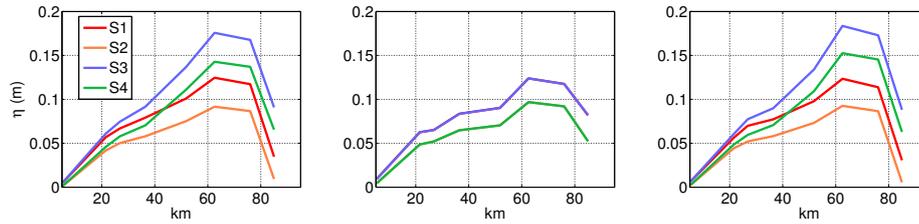


Fig. 6. Along channel representation of the mean subtidal levels with the differens scenarios (S1,S2,S3 and S4) with the formulation of [5] (left panel), [6] (central panel) and Mixed (right panel).

The first step is to test the model against a high discharge in many instruments as possible. This corresponds to the discharges of April 2008 with peak values of $550 \text{ m}^3/\text{s}$. The global coefficient of determination R^2 , using all the stations, is 0.67. The second step is to explore the along-estuary response with a higher discharge. The selected period corresponds to the winter of 2009/2010. Only β_0 and β_S recorded the elevations. In this case, the fit improves, with a global R^2 of 0.81. Finally, we focus on the second period with high discharge, from 15-February 2010 and 13-March 2010. The recorded subtidal elevations in β_S were higher than 0.4 m. The performance of the observations and the predictions is $R^2=0.87$. Thus, the regression model can fairly predict the subtidal elevations in flooding. From a management point of view and delimiting, beforehand, the value of the discharge, the model can be used to draw a plan of action against flooding in the riverine communities with 25 hours in advance.

We chose this later discharge to fulfill the short-term analysis. This time, the objective was the assessment of what would have happened if the depth of the navigational channel had been deeper. In this case, the shift is introduced by modifying the harmonics. Plotting the differences in the subtidal elevations between the current situation and the increased depth scenario, the situation worsen in terms of safety upstream, from pk 35 to the weir. It must be noticed that this is a subtidal assessment of 25 hours. Probably, the results after accounting for higher frequency motions (discharge regarded as a ‘bore’) would be more adverse.

4 Results and Conclusions

The seasonality of the freshwater discharge in the Guadalquivir was characterized through a non-stationary mixture distribution model, with a Lognormal for the central regime and Generalized Pareto distribution for both upper and lower tails. Though the lower-most part of the discharge distribution is not captured due to the regulation of the dam, quintiles greater than 0.1 are well reproduced.

Linear and non-linear regression models were applied to characterize subtidal water levels and currents. These accounted for astronomical tidal range

and freshwater discharge. The best fit for elevations corresponds to a linear superposition of tidal range and discharge, while the best fit for currents includes non-linear interaction. In the Guadalquivir, these models capture the elevations better than the currents. The elevation signal is better reproduced at the landward-most stations. An analysis of the subtidal levels distribution revealed that the upper tails are better modeled than the lowers, which is a relevant fact when assessing the navigability in the channel. Markov Chain processes were analyzed, but not considered in the simulations: despite the performance of the calibration notably improves with the Markov process, the validation remains the same.

The obtained relationships were used to predict on a medium-term basis the effect of different management strategies. Monte Carlo techniques, with regression models, were used for subtidal water levels simulations. During the subtidal water levels and currents simulations, the coefficients that relate each of the processes within the models are fixed. The changes associated to the forcings shift is introduced in their distribution. Two scenarios were analyzed: a freshwater discharge decrease by 15% for the next years, and the deepening of the navigation channel by increasing the water depth 23%. A new non-stationary distribution computed for the freshwater discharge reduction. The shifts associated to the deepening were introduced by changing the amplitude and phase of the semidiurnal component M2 through a one-dimensional tidal model. Considering the subtidal elevations simulations, the comparison between the present state and the scenarios revealed that subtidal levels decrease when the freshwater input reduces, and increase as a consequence to deepening the channel. A short term scenario was studied to assess the response of the system in case of flooding. Results point that the increase in the subtidal elevations is more significant from km 35 upstream to the dam. Overall, these simulations contribute to a better knowledge of the subtidal water level distributions with tidal motion and freshwater discharge.

Acknowledgements

The authors acknowledge support from the Programa Estatal de Investigación, Desarrollo e Innovación orientada a los RETOS de la sociedad (Ref. CTM2017-89531-R);

References

1. Ruiz, J. and Losada, M. A.: Propuesta Metodológica para Diagnosticar y Pronosticar las Consecuencias de las Actuaciones Humanas en el Estuario del Guadalquivir. Tech. rep. Spanish National Research Council (CSIC), University of Granada (UGR) and University of Córdoba (UCO).(2010).
2. Navarro, G., Gutierrez, F. J., Díez-Minguito, M., Losada, M. A., and Ruiz, J.: Temporal and Spatial Variability in the Guadalquivir Estuary: A challenge for Real-Time Telemetry. *Ocean Dynamics* 61.6, pp. 753765, (2011).
3. Solari, S. and Losada, M. A. : Non-stationary Wave Height Climate Modeling and Simulation. In: *Journal of Geophysical Research: Oceans* 116, (2011).

4. Solari, S.: Metodologías de Simulación de Agentes Naturales y Desarrollo de Sistemas. Modelo de Verificación y Gestión de Terminales Portuarias. Aplicación al Puerto de la Bahía de Cádiz. PhD thesis. University of Granada, Spain, (2011).
5. Godin, G.: The Propagation of Tides up Rivers with Special Considerations on the Upper Saint Lawrence River. *Estuarine, Coastal and Shelf Science* 48.3, pp. 307324, (1999).
6. Kukulka, T. and Jay, D. A.: Impacts of Columbia River Discharge on Salmonid habitat: 2. Changes in Shallow-Water Habitat. *Journal of Geophysical Research: Oceans* 108.3294 (C9),(2003).
7. Díez-Minguito, M., Baquerizo, A., Ortega-Sánchez, M., Navarro, G., and Losada, M. A.: Tide Transformation in the Guadalquivir Estuary (SW Spain) and Process-Based Zonation. *Journal of Geophysical Research: Oceans* 117, (2012).
8. Prandle, D. and Rahman, M.: Tidal Response in Estuaries. *Journal of Physical Oceanography* 10.10, pp. 15521573,(1980).
9. CEDEX: Evaluación del Impacto del Cambio Climático en los Recursos Hídricos en Régimen Natural. Tech. rep. Centro de Estudios y Experimentación de Obras Públicas, p.22, (2011).

Nonstationary time series forecasting of wind and waves, combining hindcast, measured and satellite data

Christos Stefanakos

SINTEF Ocean, Postboks 4762 Torgard, 7465 Trondheim, Norway

Abstract. In a series of previous papers, the well-known FIS/ANFIS systems have been successfully combined with a nonstationary time series modelling for improved predictions of wind and wave parameters. The initial time series is first decomposed into a seasonal mean value and a residual stationary part multiplied by a seasonal standard deviation. Then, the FIS/ANFIS models are applied to the stationary part. Then, they are combined with the already estimated seasonal patterns (mean value and standard deviation) to obtain forecasts of the full time series. In the present paper, different sources of data are combined for the estimation of different parts of the time series (hindcast or buoy for the stationary part and buoy or satellite for the seasonal patterns). In this way, one data from different sources and from different time periods can be combined with very good results. The performance of forecasting procedures is assessed by means of well-known error measures.

Keywords: FIS/ANFIS, wind and wave data, forecasting, satellite, buoy, hindcast

1 Introduction

The study of wind and wave climate is very important for a number of applications, including among others design of coastal and offshore structures, coastal erosion and sediment transport, wave energy resource evaluation etc. There is a number of sources of wind and wave data, among which *in situ* buoy measurements consist the most reliable data source. However, measurement campaigns are considerably costly, and buoy networks do not have a good spatial coverage of the seas, providing us with a relatively small number of long-term records of wind and wave measurements.

A very useful alternative is the long-term hindcast wind and wave databases based on third generation spectral models [24, 23, 2]. They provide us with data of good spatial and time resolution without gaps, and, thus, can be used for forecasting purposes (either off-line or in near real-time). In addition, some of them are freely available; see, e.g., [13, 5, 6]. Wave hindcasts are generated in a global grid, and they are based on improved wind input and assimilated observations concerning atmospheric, oceanic, land, and ice information [21]. These datasets are continuously assessed for their accuracy; see, e.g., [3, 22].

A third source of wind and wave data used in studying ocean wave climate is satellite measurements, mainly from SAR and altimetry [11, 9]. SAR images provide wave spectra, but the time and space resolution is not very good failing to capture properly some wave characteristics. On the other hand, altimeter wind and wave data is a very useful source with measurements presently available almost continuously over a 26-year time period from several altimeter missions [12, 8].

Each source of data presents advantages and disadvantages concerning the time and spatial resolution offered and/or the number of parameters provided. For example, buoy and model data can provide with the full directional wave spectrum, while satellite altimeters only with significant wave height and wind speed. Model and satellite data have a long-term (in time) and global (in space) coverage, while buoy measurements are usually available only for few locations around the world and not for so many years [14].

The main motivation for this paper is to combine various sources of wind and wave data to produce predictions of future values by exploiting the advantages of each source in the estimation of different characteristics in different time scales. For this, a nonstationary modelling will be applied, which has been developed by the author in a series of works for the analysis, modelling and simulation of time series of wind and wave [1, 15], and maritime parameters [16].

According to this modelling, the initial nonstationary series is decomposed into a seasonal mean value, and a residual time series multiplied by a seasonal standard deviation. The seasonal components are estimated using mean monthly values, and the residual time series is modelled as stationary series. Then, FIS/ANFIS models are applied only to the stationary part to obtain predictions of future values. Nonstationary modelling is finally used for the synthesis of the full simulated time series. The results of this combination of the nonstationary modelling with the FIS/ANFIS models has been published in a series of works with predictions of wind and wave parameters in short- [25], long- [18, 19] and very long-term [20, 17] horizon.

In the previous works only one source of data has been used at a time (either hindcast or buoy data). In the present work, two different sources are combined: satellite or buoy data are used for the estimation of the seasonal patterns (mean value and standard deviation), while hindcast data are used for the estimation of the parameters of FIS/ANFIS model.

The forecasting procedure is applied to four different locations in four different oceans (Pacific, N.Atlantic, S.Atlantic and Indian), and results are compared with values kept for for this purpose. In addition, a location in the Norwegian Sea (Haltenbanken) has been chosen, because all three sources (buoy, hindcast and satellite) are available there.

The assessment of the predictions by means of well-known error measures shows a very promising performance of the proposed methodology. Further results in a global scale are under development and are going to be presented in the near future.

2 Methodology

The nonstationary stochastic model under discussion in the present work has been presented in its univariate form in [1], and been extended to its multivariate version in [15]; see also [16]. It can be described as follows; see also Fig.1.

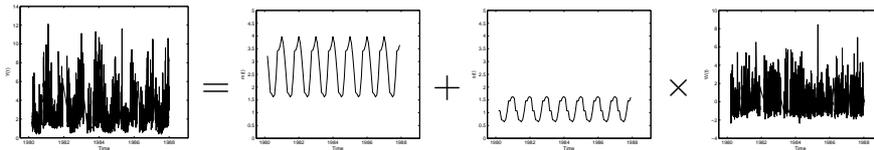


Fig. 1: Nonstationary time series modelling

A many-year long time series of wind and wave data can be treated as a nonstationary stochastic process with yearly periodic mean value and standard deviation. That is, it admits of the following decomposition:

$$Y(t) = m(t) + s(t) W(t), \tag{1}$$

where $m(t)$ and $s(t)$ are deterministic periodic functions with a period of one year, and $W(t)$ is a zero-mean, stationary, stochastic process. The functions $m(t)$ and $s(t)$ are seasonal mean value and seasonal standard deviation, respectively, and describe the exhibited seasonal patterns.

The seasonal patterns (mean value and standard deviation) are easily obtained by:

$$\tilde{\mu}_3(m) = \frac{1}{J} \sum_{j=1}^J \frac{1}{K_m} \sum_{k=1}^{K_m} Y(j, m, \tau_k), \tag{2}$$

$$\tilde{\sigma}_3(m) = \frac{1}{J} \sum_{j=1}^J \sqrt{\frac{1}{K_m} \sum_{k=1}^{K_m} \left[Y(j, m, \tau_k) - \frac{1}{K_m} \sum_{k=1}^{K_m} Y(j, m, \tau_k) \right]^2}, \tag{3}$$

with $m=1,2,\dots,12$. It has been shown that, periodic extensions of quantities $\tilde{\mu}_3(m)$ and $\tilde{\sigma}_3(m)$ are good estimates of periodic functions $m(t)$ and $s(t)$ [14].

In this way, the information contained in the time series $Y(t)$ is decomposed into two parts:

- one deterministic $[m(t), s(t)]$, containing info about features such as seasonal variability, interannual variability, climatic trends, and evolving more slowly in time, and
- one stochastic $[W(t)]$, containing info about the dependency (correlation) structure of the successive values of the series, and evolving more rapidly in time.

Then, the FIS/ANFIS forecasting methodology described in [18], is applied to the stationary part $W(t)$. The membership functions to form the fuzzy input sets are simple linear functions and the FIS systems are established assuming the following IF-THEN rules:

(a) wind speed W_S :

$$W_S(t+1) = f_1(W_S(t)) \quad (4)$$

(b) significant wave height H_S :

$$H_S(t+1) = f_2(W_S(t), H_S(t)) \quad (5)$$

Finally, using Eq. (1), the forecasted time series $\widehat{W}(t)$ is combined with seasonal components $m(t)$ and $s(t)$ estimated by another source of data to give a forecasted values of the initial (nonstationary) series $\widehat{Y}(t)$.

3 Data used

Data of significant wave height H_S and wind speed W_S have been used for this work, coming from three different sources as follows:

(i) *Hindcast data* from WAVEWATCH III model and GFS analysis winds. For more details, see [4]. Four selected points have been chosen to represent the four major areas of the oceans; see Fig. 2. The coordinates of the points are:

P1: (180E, 25N), Pacific Ocean,
 P2: (40W, 40N), N.Atlantic Ocean,
 P3: (20W, 20S), S.Atlantic Ocean,
 P4: (80E, 40S), Indian Ocean.

At each datapoint, three-hourly time series of significant wave height H_S and wind speed W_S are available, spanning a thirteen-year long period (2005–2017).

In addition, point (7.5E,65N) has been used in combination with measured data (see below).

(ii) *Measured data* from Haltenbanken buoy in the Norwegian Sea (7.6 E, 65.1 N); see Fig. 3. The dataset consists of 3-hourly measurements of H_S and W_S , covering a period of 8 years (1980–1987).

(iii) *Satellite data* from nine altimeter missions from the archive of IFREMER have been used. In Table 1, information about the various satellite missions is given following [12]. For the purpose of the present study, a 25-year long period (1992–2016) has been chosen. In Fig. 4, an example of satellite orbits around a hindcast data point is shown.

Comparisons with buoy data [7, 10] show that the H_S -estimate from the altimeter is in general in agreement with the in-situ data, with standard deviations of differences of the order of 0.30 m, but tends to slightly overestimate low values and to underestimate high values of H_S . Then, corrections to H_S have been established. For a complete list of the corrections per satellite, one should look into [12].

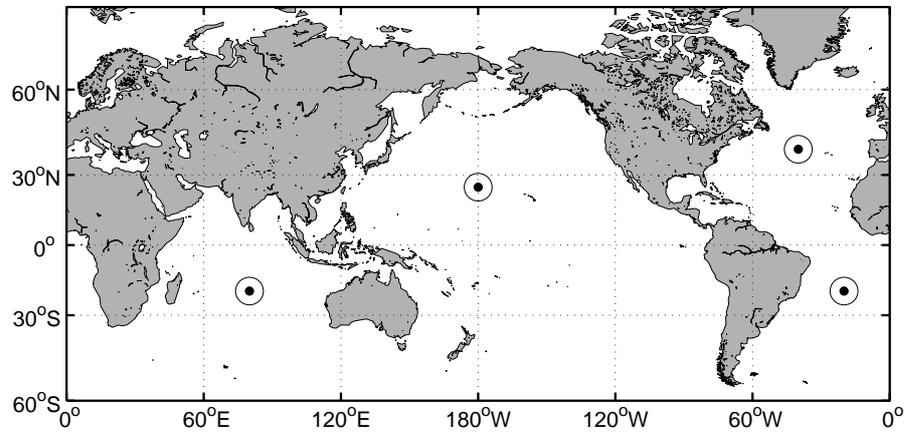


Fig. 2: Data points used. (i) Hindcast

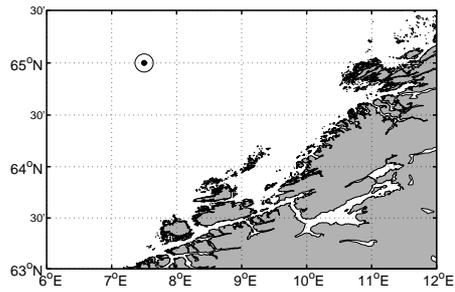


Fig. 3: Data points used. (ii) Measured (Haltenbanken)

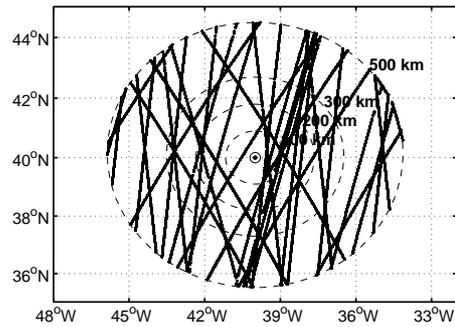


Fig. 4: Data points used. (iii) Satellite altimeter

Table 1: Metadata information on satellite altimeter data

Satellite	Product	Time Period
ERS1	OPR	01-08-1991 to 30-03-1992
		14-04-1992 to 20-12-1993
		24-12-1993 to 10-04-1994
		10-04-1994 to 21-03-1995
		24-03-1995 to 02-06-1996
ERS2	OPR	15-05-1995 to 04-07-2011
ENVISAT	GDR	14-05-2002 to 08-04-2012
TOPEX/ Poseidon	M-GDR	25-09-1992 to 08-10-2005
Jason-1	GDR	15-01-2002 to 21-06-2013
Jason-2	GDR	04-07-2008 to 17-01-2017
GEOSAT FO	GDR	07-01-2000 to 07-09-2008
Cryosat-2	IGDR	14-07-2010 to 17-01-2017
SARAL	GDR	14-03-2013 to 17-01-2017

4 Numerical results and discussion

4.1 Seasonal analysis

In a previous publication [14], it has been proven that the seasonal mean value and standard, deviation defined in Eqs. (2)–(3), can be equally well estimated by means of satellite data by assuming a spatially homogeneous area S_R , surrounding a specific site of interest. Then, all satellite observations within the area S_R can be associated to this site. The extent of this area has to be defined. In the present work, four different areas S_R have been tested of radius $R=100, 200, 300, 500$ km; see Fig. 5.

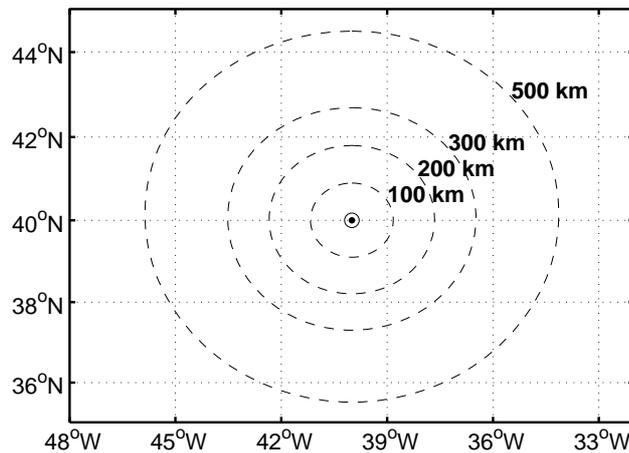


Fig. 5: Definition of a homogeneous area of satellite measurements

Since monthly values are to be estimated, the first thing that one should look into is the amount of available data per month. This will assure that a sufficient number of measurements exist per month. In addition, each yearly segment should contain all (or at least enough) months in order to depict the seasonality within the year.

In Fig. 6, where data from the Indian Ocean are shown, one can observe that only in the case of $R=500$ km one can find yearly segments containing twelve measurements (i.e., all months). On the other hand, when $R=100$ km, the number of months per year is not greater than six, which means that this case can hardly describe the seasonality within the year. Accordingly, if one looks into the number of years per month, there are cases where only 10 years are available instead of the nominal number of 25. This means that the dataset should be first filtered and keep only those which can reproduce the seasonality of the area.

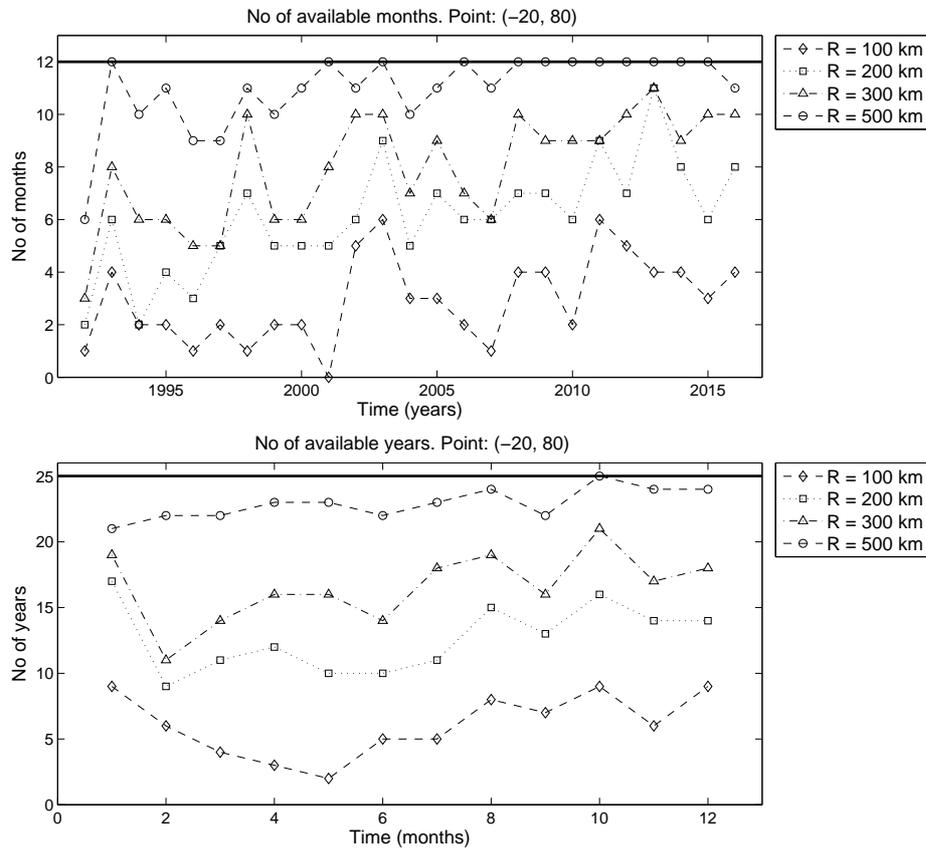


Fig. 6: Number of existing months and years (Indian Ocean)

Then, estimates of the seasonal mean value and standard deviation are obtained for both H_S and W_S in all four points. Two examples for points P2 and P4 are shown in Figs. 7–8. The results are equally good for the other two points, but due to lack of space the corresponding figures are omitted.

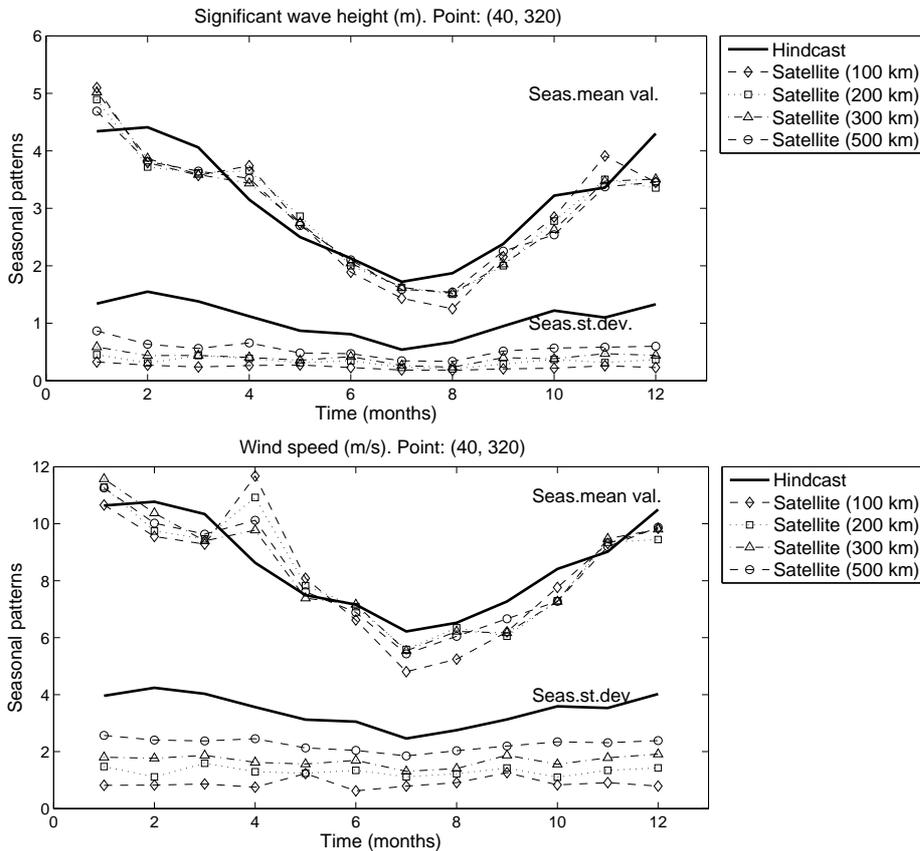


Fig. 7: Seasonal mean value and standard deviation (N.Atlantic Ocean)

Generally, the results for mean value of both parameters show an improvement as one goes from the S_{100} area ($R=100$ km) to the S_{500} . The estimated values based on satellite data differ from the ones based on hindcast in a percentage ranging from 1–10%, though there are some extreme cases where the difference is of the order of magnitude of 30%. The largest discrepancies for H_S are: P1 23% (Oct), P2 33% (Aug), P3 22% (Oct), P4 24% Jul). For W_S the corresponding values are: P1 20% (Oct), P2 35% (Apr), P3 19% (Oct), P4 24% (Apr).

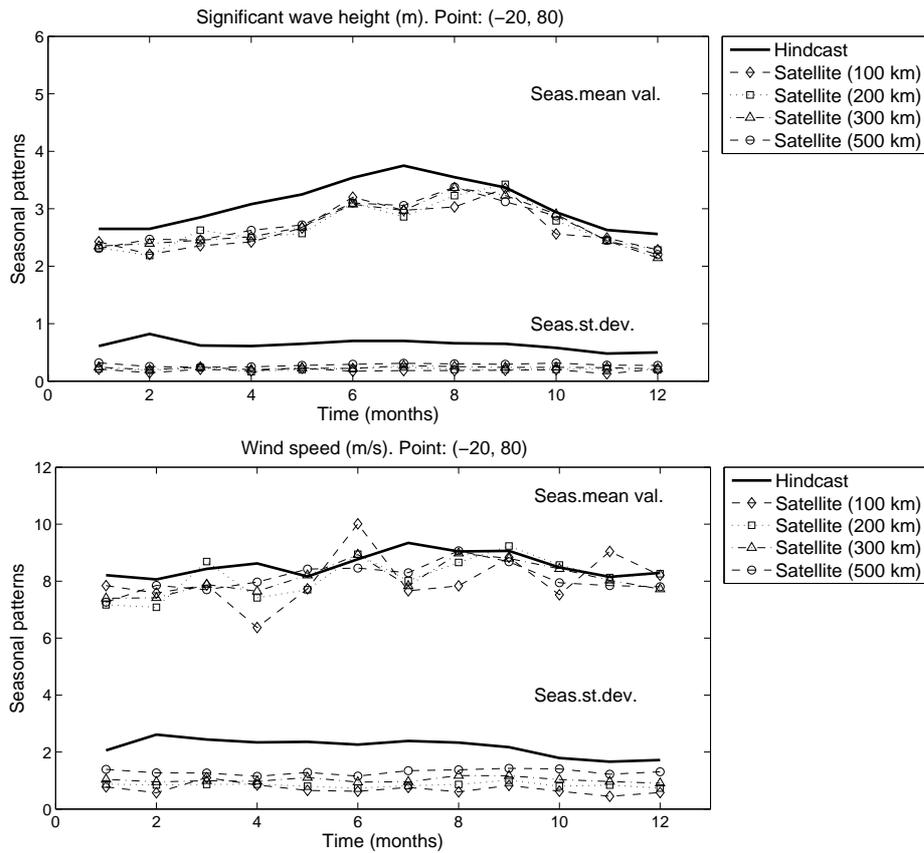


Fig. 8: Seasonal mean value and standard deviation (Indian Ocean)

Table 2: Difference between hindcast and satellite seasonal mean value and standard deviation as a percentage (%) of the hindcast values in P2: N.Atlantic Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	Mean value								Standard deviation							
	H_S				W_S				H_S				W_S			
	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}
Jan	18	13	16	8	0	6	9	6	76	66	56	35	79	63	55	35
Feb	14	16	12	13	11	10	4	7	83	80	72	59	81	74	58	43
Mar	12	11	12	10	10	8	9	7	83	68	68	59	79	61	54	41
Apr	19	16	9	12	35	27	13	17	76	65	64	41	79	64	54	31
May	9	15	10	8	8	4	1	2	68	64	59	45	61	60	50	32
Jun	11	6	4	1	8	1	0	4	72	57	48	42	80	56	45	33
Jul	17	6	6	8	23	10	11	13	66	57	51	36	68	55	47	25
Aug	33	20	19	18	20	3	5	7	73	69	65	50	67	56	49	26
Sep	9	16	15	5	15	17	15	8	78	70	58	46	60	55	40	30
Oct	12	14	18	21	8	13	14	14	82	70	69	53	77	69	57	35
Nov	16	4	3	0	3	4	5	2	76	71	57	47	74	62	49	34
Dec	20	22	18	20	6	10	7	6	83	72	67	55	80	64	52	41

Table 3: Difference between hindcast and satellite seasonal mean value and standard deviation as a percentage (%) of the hindcast values in P4: Indian Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	Mean value								Standard deviation							
	H_S				W_S				H_S				W_S			
	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}	S_{100}	S_{200}	S_{300}	S_{500}
Jan	8	12	11	13	4	13	10	12	66	65	60	48	62	57	49	32
Feb	17	17	10	7	6	12	8	3	82	77	76	69	78	68	64	51
Mar	17	8	14	14	7	3	7	9	67	63	61	61	55	65	59	48
Apr	21	19	18	15	26	14	11	8	70	73	67	59	64	63	58	51
May	18	21	18	16	5	6	1	3	66	69	65	58	72	66	53	46
Jun	10	12	13	12	14	2	2	4	75	69	68	58	72	68	58	49
Jul	21	24	21	18	18	14	16	11	74	65	61	55	68	66	60	44
Aug	15	9	5	5	13	4	1	0	72	69	62	54	74	63	50	41
Sep	0	2	4	7	2	2	3	4	71	68	63	55	62	53	46	34
Oct	13	5	1	2	11	1	1	6	65	64	58	46	65	55	42	21
Nov	5	7	7	7	11	0	1	4	73	57	51	42	73	49	42	26
Dec	11	11	16	14	1	0	7	6	58	60	57	45	66	57	47	24

It seems that, there is a slightly better agreement for H_S , where satellite data have been corrected based on buoy measurements [12]. On the other hand, the uncorrected wind speed measurements have been used for the present analysis, because the population of the corrected ones are not yet ready to be analysed.

In contrast, the situation is not so good in the estimation of seasonal standard deviation. The differences between the satellite and hindcast estimates range from 50% up to 85%. Also, the seasonal variation in the satellite estimates is not so eminent as in the hindcast estimates.

In Tables 2–3, the difference between hindcast and satellite estimates are given as a percentage (%) of the hindcast values for both parameters in points P2 and P4. The corresponding results for P1 and P3 are omitted due to lack of space, but they are available upon request.

Haltenbanken

In the sequel, a location in the Norwegian Sea has been chosen, because there are buoy measurements for that. The location name is Haltenbanken and it is a known oil and gas field.

For Haltenbanken, all three sources of data are available and compared; see Fig. 9 and Table 4. Especially, for the satellite dataset, the S_{500} area was used as it has shown a better performance in the previous cases.

Table 4: Difference between hindcast, buoy and satellite seasonal mean value and standard deviation as a percentage (%) of the hindcast values in Haltenbanken (M: Measured, S: Satellite)

	Mean value				St. dev.			
	H_S		W_S		H_S		W_S	
	M	S	M	S	M	S	M	S
Jan	11	15	10	12	23	32	3	34
Feb	8	16	9	26	22	33	15	33
Mar	2	13	5	14	17	40	23	30
Apr	1	20	2	19	6	39	9	24
May	4	7	4	16	16	41	5	21
Jun	12	17	1	5	3	26	6	14
Jul	7	3	5	1	11	48	4	27
Aug	8	6	3	7	1	45	8	33
Sep	1	12	8	9	3	57	8	35
Oct	18	16	3	8	22	47	4	32
Nov	8	6	1	3	7	45	1	39
Dec	1	7	4	6	7	43	5	32

It seems that measured data are in a better agreement with hindcast both for mean value and standard deviation. The discrepancies range from 1–18% for the mean value and from 1–23% for the standard deviation. Satellite data succeed to

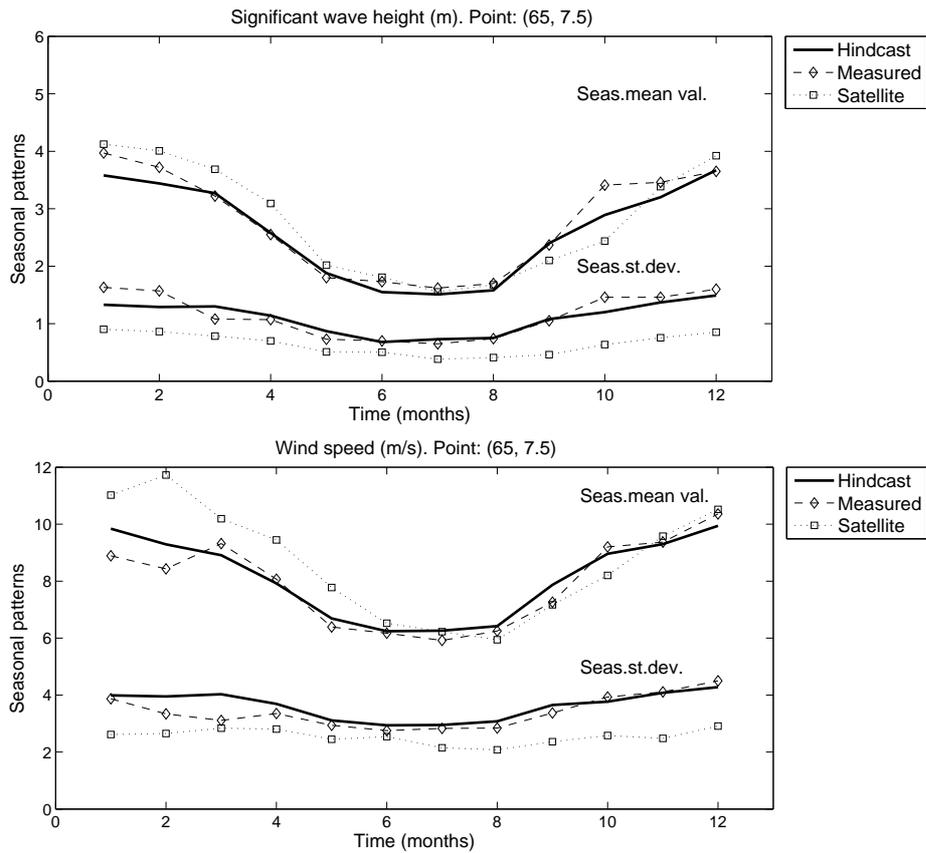


Fig. 9: Seasonal mean value and standard deviation (Haltenbanken)

estimate sufficiently well the mean value, but are not so good with the standard deviation. Especially, in the case of W_S , the discrepancies are up to 39%.

4.2 Forecasting

Having estimated the seasonal patterns, the initial time series is decomposed the FIS/ANFIS forecasting procedure is applied to the residual stationary part $W(t)$ for a forecasting horizon of one month. Then, different seasonal patterns are used in combination with the forecasts in order to obtain forecasts of the initial series. The final results are compared with the actual values of the series kept for this purpose; see Figs. 10–11.

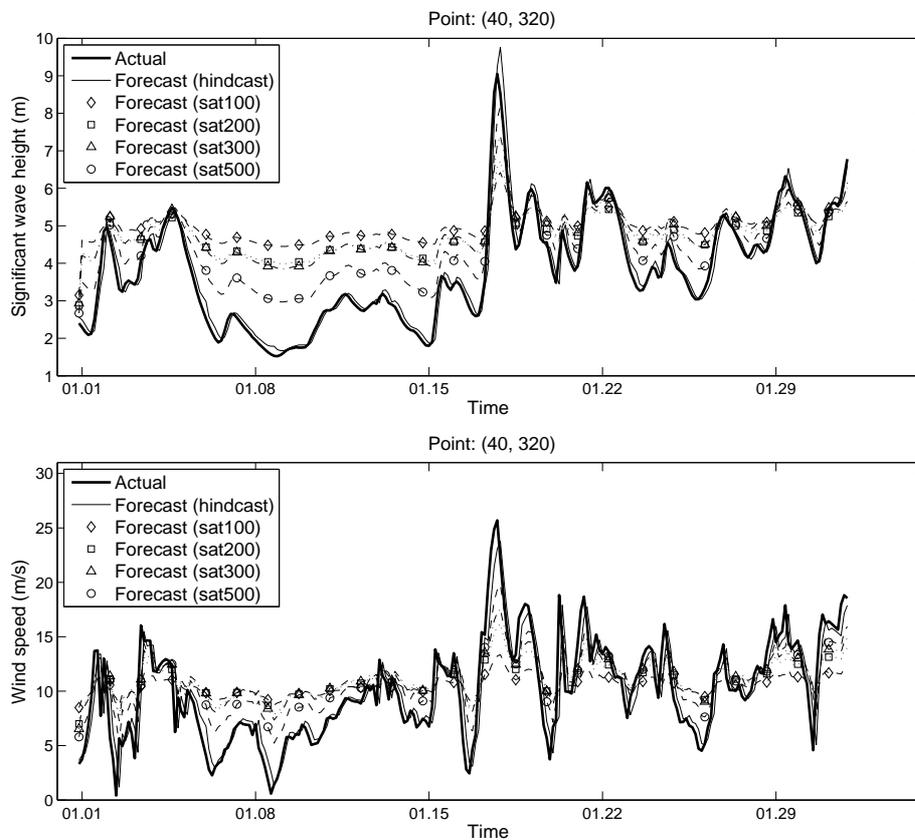


Fig. 10: Forecast of significant wave height and wind speed (N.Atlantic Ocean)

Various error measures, the definition of which is given in the Appendix, are calculated to quantify the performance of the forecasting procedure; see Tables 5–8.

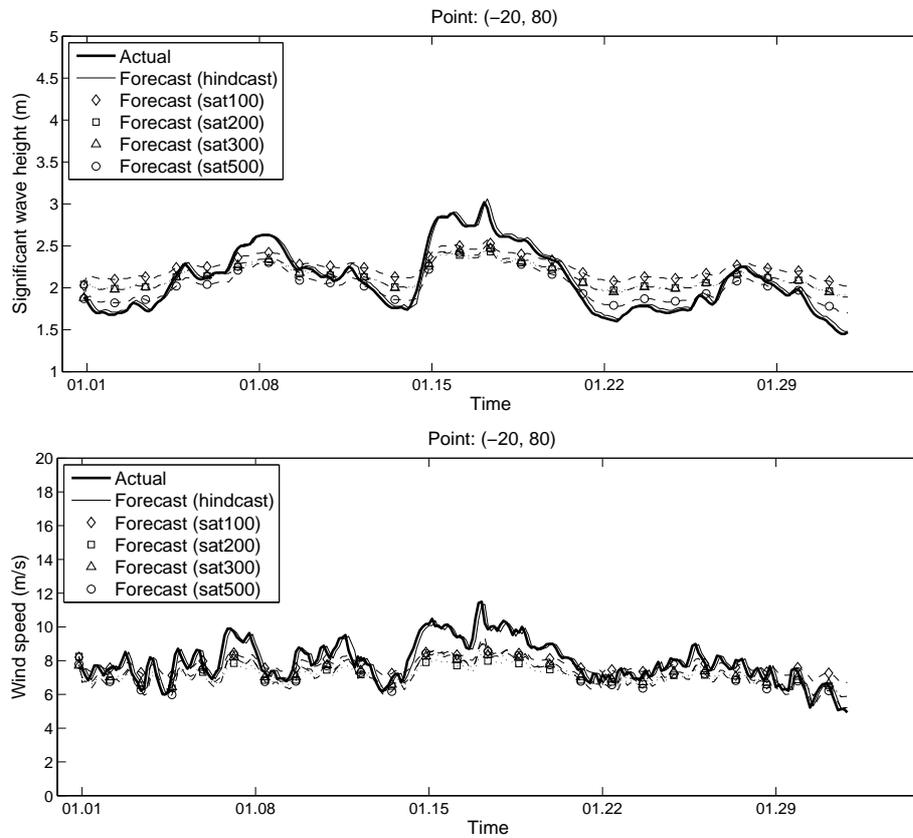


Fig. 11: Forecast of significant wave height and wind speed (Indian Ocean)

Table 5: Error measures for P1: Pacific Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	H_S					W_S				
	H	S_{100}	S_{200}	S_{300}	S_{500}	H	S_{100}	S_{200}	S_{300}	S_{500}
RMSE	0.14	0.71	0.69	0.60	0.56	1.23	3.24	2.79	2.30	1.73
MAPE	0.03	0.19	0.17	0.15	0.13	0.14	0.41	0.34	0.29	0.22
MASE	1.05	5.63	5.37	4.66	4.31	1.00	3.15	2.70	2.24	1.65
RMSSE	1.43	7.15	6.99	6.01	5.66	1.43	3.76	3.24	2.67	2.01
R^2	0.99	0.85	0.61	0.87	0.82	0.93	0.34	0.58	0.78	0.89
SI (%)	4.62	23.15	22.63	19.46	18.31	15.62	41.00	35.34	29.10	21.89
Bias	0.02	-0.09	-0.31	-0.15	-0.24	-0.08	-1.49	-1.44	-1.03	-0.68
Bias (max)	0.62	2.13	2.14	1.88	1.79	4.90	8.68	8.06	7.23	6.21

Table 6: Error measures for P2: N.Atlantic Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	H_S					W_S				
	H	S_{100}	S_{200}	S_{300}	S_{500}	H	S_{100}	S_{200}	S_{300}	S_{500}
RMSE	0.30	1.58	1.31	1.28	0.78	2.05	3.84	3.39	3.23	2.60
MAPE	0.06	0.50	0.41	0.40	0.24	0.21	0.60	0.56	0.54	0.41
MASE	0.88	5.51	4.55	4.49	2.73	0.98	2.16	1.92	1.83	1.45
RMSSE	1.26	6.55	5.43	5.31	3.25	1.41	2.64	2.33	2.21	1.78
R^2	0.98	0.29	0.47	0.53	0.84	0.89	0.81	0.76	0.76	0.86
SI (%)	7.93	41.21	34.17	33.36	20.43	20.22	37.86	33.45	31.82	25.61
Bias	0.04	1.15	0.89	0.98	0.55	-0.05	0.40	0.93	1.18	0.75
Bias (max)	1.39	2.95	2.52	2.38	1.69	11.19	12.47	9.78	8.62	9.52

Table 7: Error measures for P3: S.Atlantic Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	H_S					W_S				
	H	S_{100}	S_{200}	S_{300}	S_{500}	H	S_{100}	S_{200}	S_{300}	S_{500}
RMSE	0.04	0.13	0.13	0.10	0.10	0.59	1.02	1.00	0.85	0.73
MAPE	0.02	0.07	0.07	0.05	0.05	0.08	0.17	0.18	0.15	0.12
MASE	1.03	3.38	3.47	2.75	2.65	1.00	1.77	1.64	1.47	1.29
RMSSE	1.23	4.12	4.20	3.41	3.32	1.31	2.27	2.23	1.90	1.62
R^2	0.98	0.87	0.75	0.96	0.91	0.90	0.84	0.76	0.90	0.90
SI (%)	2.39	7.98	8.15	6.60	6.42	9.63	16.64	16.38	13.94	11.93
Bias	0.01	0.03	0.06	-0.02	-0.04	0.02	-0.17	0.33	-0.06	-0.10
Bias (max)	0.11	0.27	0.29	0.27	0.27	3.00	3.31	3.57	2.80	2.15

Table 8: Error measures for P4: Indian Ocean (H: Hindcast, S_{100} : Satellite $R=100$ km, S_{200} : Satellite $R=200$ km, S_{300} : Satellite $R=300$ km, S_{500} : Satellite $R=500$ km)

	H_S					W_S				
	H	S_{100}	S_{200}	S_{300}	S_{500}	H	S_{100}	S_{200}	S_{300}	S_{500}
RMSE	0.05	0.29	0.25	0.23	0.19	0.42	0.85	1.17	0.96	1.01
MAPE	0.02	0.13	0.10	0.10	0.07	0.04	0.08	0.11	0.09	0.10
MASE	1.28	7.62	6.29	5.94	4.73	1.00	2.00	2.85	2.29	2.60
RMSSE	1.54	8.79	7.48	7.07	5.84	1.26	2.55	3.52	2.90	3.04
R^2	0.99	0.63	0.94	0.94	0.95	0.94	0.87	0.47	0.63	0.64
SI (%)	2.44	13.94	11.86	11.21	9.26	5.30	10.69	14.75	12.18	12.76
Bias	0.03	0.15	0.04	0.05	-0.06	0.02	-0.18	-0.87	-0.66	-0.84
Bias (max)	0.16	0.59	0.58	0.54	0.54	1.30	2.74	3.31	2.90	2.66

One can observe first that, the performance of S_{500} is better than the others. However, due to lack of variation in the seasonal standard deviation, the final forecasts based on satellite data are not so good as the ones based on hindcast data. Especially, in point P2 (N.Atlantic), the forecasts are better for higher values, but not so good for the lower ones. So, further investigation should be done on this matter.

Haltenbanken

Here the performance of satellite data is better, especially in the case of W_S , and the range of error measures is the same as in the case of measured data; see Fig. 12 and Table 9. The measured data are in very good agreement with the hindcast ones.

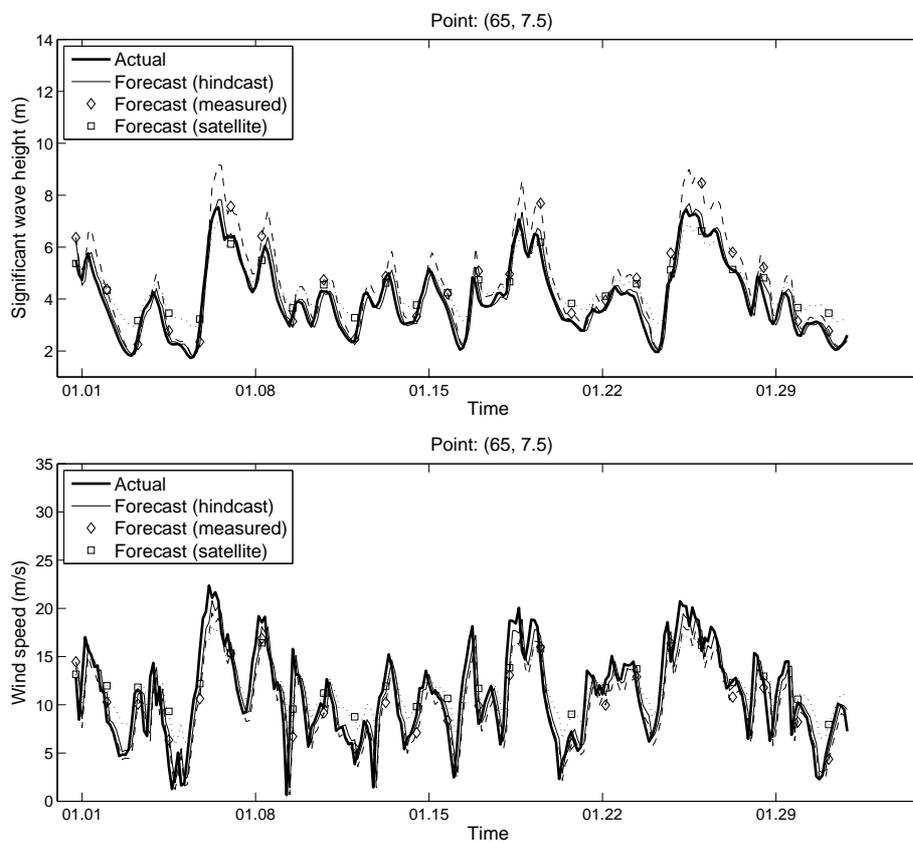


Fig. 12: Forecast of significant wave height and wind speed (Haltenbanken)

Table 9: Error measures for Haltenbanken (H: Hindcast, M: Measured, S_{500} : Satellite $R=500\text{km}$)

	H_S			W_S		
	H	M	S_{500}	H	M	S_{500}
RMSE	0.36	0.77	0.66	2.33	2.60	2.74
MAPE	0.07	0.15	0.18	0.27	0.26	0.38
MASE	0.89	1.95	1.78	0.97	1.10	1.20
RMSSE	1.13	2.42	2.08	1.27	1.41	1.49
R^2	0.97	0.92	0.89	0.87	0.83	0.84
SI (%)	8.83	18.89	16.29	21.01	23.39	24.68
Bias	0.05	0.56	0.42	-0.15	-1.12	0.65
Bias (max)	0.94	2.26	1.40	8.19	9.08	9.36

5 Concluding Remarks

In the present work, forecasts of significant wave height H_S and wind speed W_S have been obtained based on a newly introduced procedure by combining various sources of data (hindcast, measured and satellite). The methodology is applied to four points in the open ocean (Pacific, N.Atlantic, S.Atlantic and Indian Ocean), and one point in the Norwegian Sea (Haltenbanken).

Seasonal analysis showed that estimates based on satellite data from a sufficiently large area around the data point can be good replacements for seasonal mean value estimated based on hindcast data. However, estimates of seasonal standard deviation shows a lower variability and this matter should be further investigated.

The performance of the forecasting procedure based on the various datasets is assessed by means of various error measures, revealing that in general measured data can be used in combination with hindcast data while satellite ones need to be further tested. Further results in a global scale are under development and are going to be presented in the near future.

Acknowledgments This work has been partially funded by the NFR project “High-dimensional statistical modelling of changes in wave climate and implications for maritime infrastructure” (HDwave) under Contract No. 243814/E10, with partners Norsk Regnesentral (coordinator), DNV-GL and SINTEF Ocean.

References

1. Athanassoulis, G., Stefanakos, C.: A nonstationary stochastic model for long-term time series of significant wave height. *Journal of Geophysical Research, Section Oceans* 100(C8), 16149–16162 (1995)
2. Booij, N., Ris, R., Holthuijsen, L.: A third-generation wave model for coastal regions: 1. model description and validation. *Journal of Geophysical Research, Section Oceans* 104((C4)), 7649–7666 (1999)

3. Caires, S., Sterl, A., Bidlot, J.R., Graham, N., Swail, V.: Intercomparison of different wind wave reanalyses. *Journal of Climate* 17(10), 1893–1913 (2004)
4. Chawla, A., Spindler, D., Tolman, H.: Wavewatch iii hindcasts with re-analysis winds. initial report on model setup. Tech. rep., National Centers for Environmental Prediction (2011)
5. Chawla, A., Spindler, D.M., Tolman, H.L.: Validation of a thirty year wave hindcast using the Climate Forecast System Reanalysis winds. *Ocean Modelling* 70, 189–206 (2013)
6. Perez, J., Menendez, M., Losada, I.J.: GOW2: A global wave hindcast for coastal applications. *Coastal Engineering* 124, 1–11 (2017)
7. Queffeuilou, P.: Long term quality status of wave height and wind speed measurements from satellite altimeters. In: Proceedings of the ISOPE conference. Honolulu, Hawaii, USA (May 25-30 2003)
8. Queffeuilou, P.: Merged altimeter wave height data base. an update. In: ESA Living Planet Symposium. Edinburgh, UK (9-13 September 2013), <ftp://ftp.ifremer.fr>
9. Queffeuilou, P., Croizé-Fillon, D.: La mesure satellite de hauteur de vague par altimètre. État des lieux, application à la climatologie et à la modélisation des états de mer. In: AMA 2009, Les ateliers de modélisation de l’atmosphère. Toulouse (27-29 janvier 2009), <ftp://ftp.ifremer.fr>
10. Queffeuilou, P.: Long-term validation of wave height measurements from altimeters. *Marine Geodesy* 27(3-4), 495–510 (2004)
11. Queffeuilou, P., Bentamy, A.: Analysis of wave height variability using altimeter measurements: Application to the mediterranean sea. *Journal of Atmospheric and Oceanic Technology* 24(12), 2078–2092 (2007)
12. Queffeuilou, P., Croizé-Fillon, D.: Global altimeter swh data set. Tech. rep., Laboratoire d’Océanographie Physique et Spatiale, IFREMER, Plouzané, France (February 2017), <ftp://ftp.ifremer.fr>
13. Rasclé, N., Ardhuin, F.: A global wave parameter database for geophysical applications. Part 2: Model validation with improved source term parameterization. *Ocean Modelling* 70, 174–188 (2013)
14. Stefanakos, C., Athanassoulis, G., Barstow, S.: Time series modeling of significant wave height in multiple scales, combining various sources of data. *Journal of Geophysical Research, Section Oceans* 111(C10), 10001–10012 (2006)
15. Stefanakos, C., Belibassakis, K.: Nonstationary stochastic modelling of multivariate long-term wind and watedata. In: 24th International Conference on Offshore Mechanics and Arctic Engineering, OMAE’2005. Halkidiki, Greece (12-17 June 2005 2005)
16. Stefanakos, C., Schinas, O.: Forecasting bunker prices; a nonstationary, multivariate methodology. *Transportation Research Part C: Emerging Technologies* 38(1), 177 – 194 (2014)
17. Stefanakos, C., Vanem, E.: Nonstationary fuzzy forecasting of wind and wave climate in very long-term scales. *Journal of Ocean Engineering and Science* (2018)
18. Stefanakos, C.: Fuzzy time series forecasting of nonstationary wind and wave data. *Ocean Engineering* 121, 1–12 (2016)
19. Stefanakos, C.: Nonstationary Prediction of Wind and Waves in the Pacific Ocean using Fuzzy Inference Systems. In: 26th International Offshore and Polar Engineering Conference, ISOPE’2016. International Society of Offshore & Polar Engineers, Rhodes (Rodos), Greece (June 26-July 2, 2016 2016)
20. Stefanakos, C., Vanem, E.: Climatic forecasting of wind and waves using fuzzy inference systems. In: 37th International Conference on Ocean, Offshore and Arctic Engineering OMAE’2017. Trondheim, Norway (June 25-30, 2017 2017)

21. Stopa, J.E.: Wind forcing calibration and wave hindcast comparison using multiple reanalysis and merged satellite wind datasets. *Ocean Modelling* 127, 55 – 69 (2018)
22. Stopa, J.E., Cheung, K.F.: Intercomparison of wind and wave data from the ECMWF Reanalysis Interim and the NCEP Climate Forecast System Reanalysis. *Ocean Modelling* 75, 65 – 83 (2014)
23. Tolman, H.: A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents. *Journal of Physical Oceanography* 21, 782–797 (1991)
24. The WAMDI Group: The WAM model-A third generation ocean wave prediction model. *Journal of Physical Oceanography* 18(12), 1775–1810 (1988)
25. Wu, M., Stefanakos, C., Gao, Z.: Prediction of short-term wind and wave conditions using Adaptive Network-based Fuzzy Inference System (ANFIS) for marine operations. In: 3rd International Conference on Renewable Energies Offshore, RENEW 2018. Lisbon, Portugal (October 8-10 2018)

A Measuring forecasting quality

Assuming that we have I steps of forecasts and actual values to be compared, there are three large categories of errors measuring the forecasting performance:

- (i) *Scaled-dependent measures*, that depend on the scale of the data. These are useful when comparing different methods applied to the same dataset, but should not be used, for example, when comparing across data sets that have different scales.
- (ii) *Measures based on percentage errors*. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance across different data sets.
- (iii) *Relative measures*, which are calculated relatively to the error from a benchmark method.

Popular representatives of the first two categories are the

- (a) Root Mean Square Error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{I} \sum_{i=1}^I |e(t_i)|^2} \quad (6)$$

- (b) Mean Absolute Percentage Error (MAPE) defined as

$$\text{MAPE} = \frac{1}{I} \sum_{i=1}^I \left| \frac{e(t_i)}{\text{actual}(t_i)} \right|, \quad (7)$$

where

$$e(t_i) = \text{actual}(t_i) - \text{forecast}(t_i) \quad (8)$$

denotes the forecasting error at time t_i .

- (c) Percentage Error (PE) defined as

$$\text{PE}(t) = \frac{e(t)}{\text{actual}(t)}. \quad (9)$$

Furthermore, scaled errors are defined as

$$q(t_i) = \frac{e(t_i)}{\frac{1}{N} \sum_{n=2}^N |X(t_n) - X(t_{n-1})|}, \quad (10)$$

where $\{X(t_n), n = 1, 2, \dots, N\}$ are the existing observations, used for training of the FTS model. Then, one can define various error measures in an analogous way. Let us consider, e.g., the

- (a) Mean Absolute Scaled Error (MASE) defined as

$$\text{MASE} = \frac{1}{I} \sum_{i=1}^I |q(t_i)|, \quad (11)$$

and the

(b) Root Mean Square Scaled Error (RMSSE) defined as

$$\text{RMSSE} = \sqrt{\frac{1}{I} \sum_{i=1}^I |q(t_i)|^2} \quad (12)$$

Also, the usual error measures Bias, Scatter Index (SI) and correlation coefficient R^2 are calculated:

(a) Bias:

$$\text{Bias} = \frac{1}{I} \sum_{i=1}^I [-e(t_i)], \quad (13)$$

(b) Scatter Index (SI) in %:

$$\text{SI} = \sqrt{\frac{\text{RMSE}}{\sum_{i=1}^I \text{actual}(t_i)}} \times 100, \quad (14)$$

(c) Correlation coefficient R^2 :

$$R^2 = \frac{\sum_{i=1}^I \left(\text{forecast}(t_i) - \overline{\text{actual}} \right) \left(\text{actual}(t_i) - \overline{\text{actual}} \right)}{\sqrt{\sum_{i=1}^I \left(\text{forecast}(t_i) - \overline{\text{actual}} \right)^2 \sum_{i=1}^I \left(\text{actual}(t_i) - \overline{\text{actual}} \right)^2}}, \quad (15)$$

where

$$\overline{\text{actual}} = \frac{1}{I} \sum_{i=1}^I \text{actual}(t_i). \quad (16)$$

Further, because most of the error measures are smoothed out due to averaging, a measure of the instantaneous error is calculated as the maximum of absolute values of the forecasting error (8), and is mentioned in the tables as “Bias (max)”.

Spatial distribution of climatic cycles in Andalusia (southern Spain)

J. Sánchez-Morales¹, E. Pardo-Igúzquiza² and F. J. Rodríguez-Tovar¹

¹ Universidad de Granada, Avd. Fuentenueva s/n, 18071 Granada (Spain)

² Instituto Geológico y Minero de España, Ríos Rosas 23, 28003 Madrid (Spain)
josesanmor@correo.ugr.es; e.pardo@igme.es; fjrtovar@ugr.es

Abstract. Several climatic cycles in Andalusia (southern Spain) have been identified by using precipitation and temperature data for the most part of 20th century and early 21st century at 707 meteorological stations. Some of these detected cycles have been recognized in previous studies, such as the 3-years cycle and the 7/8-years cycle, which have turned out to be very common across the study area. The statistical technique has been the spectral analysis. The power spectrum estimator that has been used is the smoothed Lomb-Scargle periodogram. The results reveal very interesting spatial patterns, not seen before in previous climatic studies, which illustrate a larger influence of the Atlantic Ocean in the West and a larger influence of the Mediterranean Sea in the East. In general, the studied precipitation record presents better results than the temperature one which shows less clarity on assessing the climatic variability on Andalusia. Nevertheless, most of the spotted cycles in the precipitation record have been detected in the temperature record as well.

Keywords: Power Spectrum, Climate, Andalusia.

1 Introduction

Andalusia (southern Spain) is a region characterized by huge climatic contrasts, i.e. the ‘Sierra de Grazalema’ in the South-West stands as the wettest place in the whole Iberian Peninsula with a rainfall of more than 2,000 mm/year on average, whereas the ‘Desierto de Tabernas’ in the South-East is considered the driest place in the Continental Europe with less than 150 mm/year on average. The influence of both Atlantic Ocean and Mediterranean Sea on this area of 87,597 km², plus the presence of the Betic Cordillera with altitudes above 3,000 m.s.l. (meters above sea level), makes this region unique from a climatic point of view and very interesting for analyzing the evolution of climate at past and present times, especially during most recent times. Thus, Andalusia could be considered as a natural laboratory for the study of past to present climatic changes.

Climatic studies from the region of Andalusia are frequent, focusing on variable aspects. Thus, several statistical techniques and methodologies have been used; i.e. principal component analysis (PCA)[1], empirical orthogonal function (EOF)[2],

innovative missing values estimator [3], non-instrumental climate reconstruction [4], and gridded dataset and combined indices evolution [5] amongst many others.

The causes of climate variability in Andalusia at annual, inter-annual, decadal and multi-decadal time scales are generally associated with diverse phenomena, involving several climatic subsystems, revealing the interactions between atmosphere and ocean [1-5]. In cases, a cyclic pattern in the climatic variability has been interpreted. The identification of climatic cycles in Andalusia by means of the spectral analysis has been carried out in previous studies, although the method differs from one to another [6-7]. On this base, here we present the study of climatic evolution during XX and XXI centuries at Andalusia, based on the spectral analysis of data from different climatic proxies, to evaluate the cyclic character, and to interpret the involved processes.

2 Methodology

This study uses the spectral analysis as the statistical technique to evaluate the importance of the frequencies associated to the precipitation and temperature time-series in Andalusia. The power spectrum estimator has been the smoothed Lomb-Scargle periodogram [8-10] which works directly with uneven time-series, such as annual precipitation and/or annual temperature series in the study area. The technique that has been used [10] evaluates the statistical significance of the peaks by the Monte Carlo permutation test as neighbouring frequencies are highly correlated, and then it adjusts the statistical significance by smoothing the periodogram. In this second case, linear smoothing with 3 terms was applied to the raw periodogram. The output is formed by the Lomb-Scargle spectrum, the achieved confidence level spectrum, the mean spectrum of permutations and the phase spectrum.

The required parameters have been optimized for dealing with the annual precipitation and/or temperature time-series in the study area, and for achieving the intended goal of capturing any climatic cycle having a duration just above the sampling interval (i.e. the biannual oscillation). Thus, 0.5 has been used as the highest frequency to evaluate, 200 as the number of frequencies in the interval, 2,000 as the number of permutations, 75,654 as the random seed, 3 as the number of smoothing terms and enabling linear smoothing. To illustrate the above process, we present the output of this methodology on one of the precipitation stations (Fig. 1).

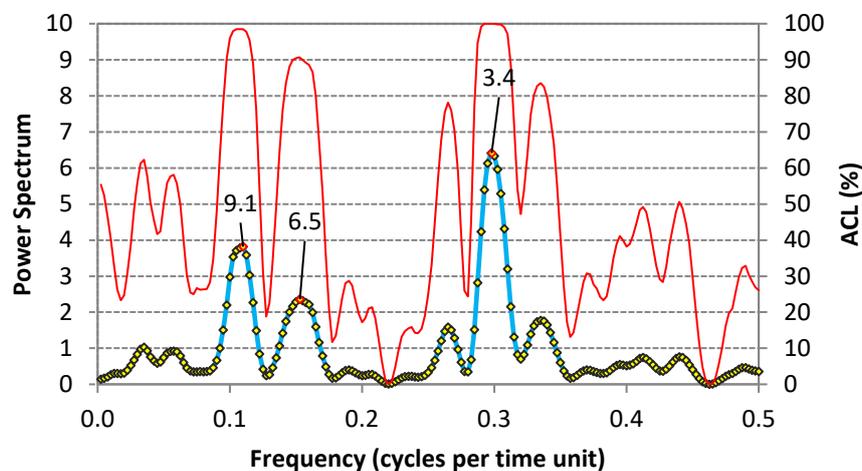


Fig. 1. Example of power spectrum of annual precipitation at station: P6289, and associated peaks (in years) above Achieved Confidence Level (ACL) of 90%: 9.1, 6.5 and 3.4 years, by using the smoothed Lomb–Scargle periodogram.

2.1 Meteorological datasets

The precipitation and temperature data were collected from two different sources and named chguadalquivir [11] and aemet [12]. The sources differ from each other at the sampling interval; datasets from chguadalquivir were available at monthly level and covering the period from 1951 to 1987, whereas datasets from aemet were available at daily level, and covering the period from 1901 to 2012. The number of datasets from chguadalquivir were 1,574 precipitation stations and 526 temperature stations, all spread across the study area. The number of datasets from aemet were 595 precipitation stations and 282 temperature stations, all located in the right half of the study area.

For a better comparison, all the datasets have been converted into annual datasets. As a first step, the daily datasets were summarized into monthly datasets; only months having a minimum of 25 days of precipitation measurements and/or a minimum of 20 days of temperature measurements were considered. Then, to convert the monthly datasets into annual datasets, only complete years were used and/or having 12 months of measurements. A second level of filtering came when excluding from the analysis those precipitation stations with less than a total of 20 years of records and those temperature stations with less than a total 10 years of records. Whether the total amount of years with records were consecutive years or not, has not been a requirement, as our spectral analysis methodology can deal with uneven series. At combining the two data sources, it was established a comparison method to see which one had more years of information and for the same station, as various stations were in both data

sources. If the length of the series in years was the same in both sources, we gave preference to the dataset from aemet which was originally compiled at daily level.

The filtering process produced 547 precipitation stations and 160 temperature stations; a total of 707 meteorological datasets to be analyzed (Fig. 2).

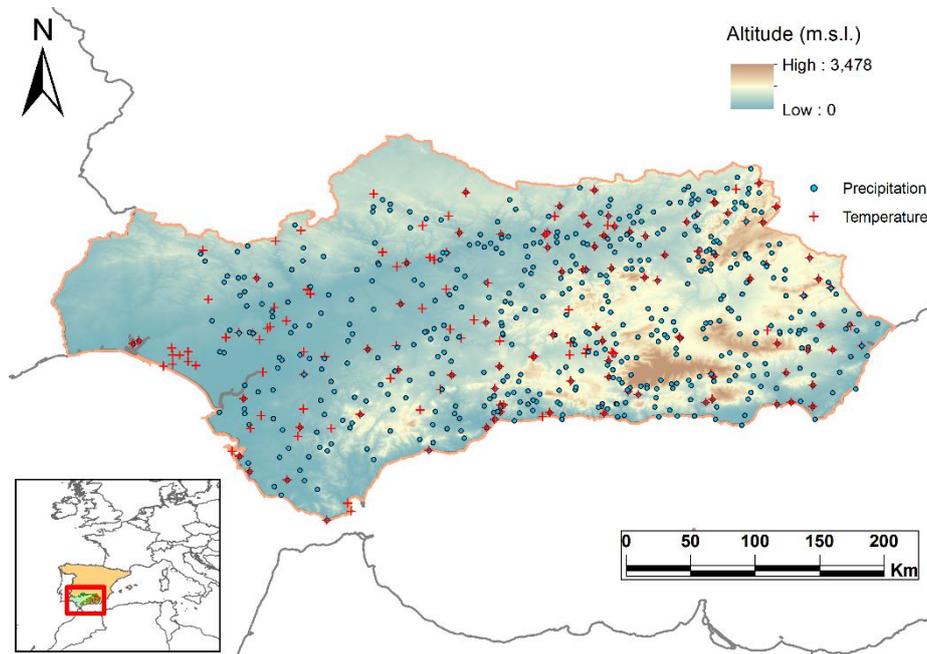


Fig. 2. Distribution of all meteorological stations selected for the spectral analysis.

3 Results

The spectral analysis carried out on the 707 datasets (547 precipitation stations and 160 temperature stations), has detected 1,751 significant peaks in the precipitation datasets (Fig. 3) and 466 significant peaks in the temperature datasets (Fig. 4), all above or equal to 90% of ACL. All the cycles detected for the same meteorological variable have been plotted in the same diagram.

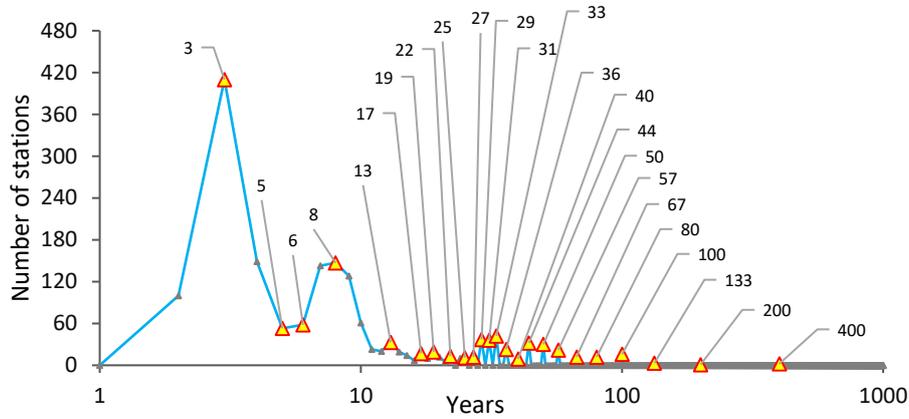


Fig. 3. Diagram containing all the precipitation cycles in the study area and the number of locations in which the cycles have been detected.

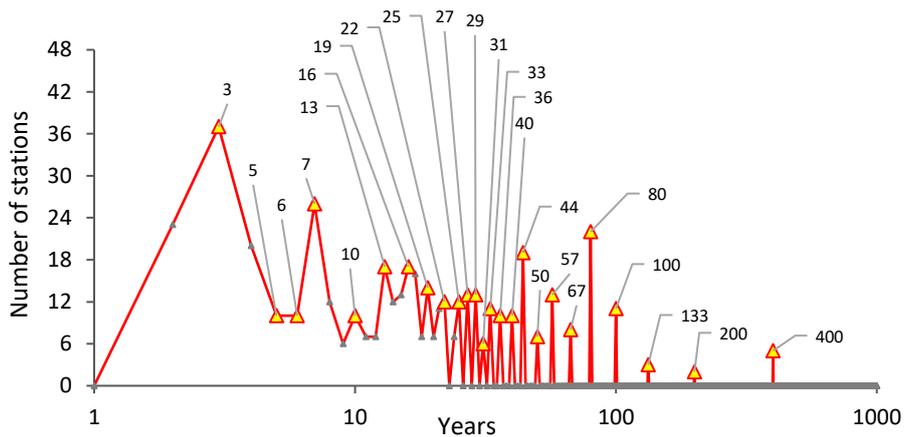


Fig. 4. Diagram containing all the temperature cycles in the study area and the number of locations in which the cycles have been detected.

The number of detected cycles is lower in the temperature record than in the precipitation one, but it is very remarkable that both temperature and precipitation variables show signals at the same frequencies, which speaks of common climatic processes affecting both variables. The 3-years cycle and the 7/8-years cycle have been the most detected ones, showing the highest significance. Thus, a detailed analysis of both cycles at 3 years and 7/8 years has been conducted.

All precipitation (344) and temperature (28) stations signalling peaks between frequency values of 0.4 (2.5 years) and 0.29 (3.5 years) have been collected, and plotted in a map to see their spatial distribution (see Figs 5 and 6).

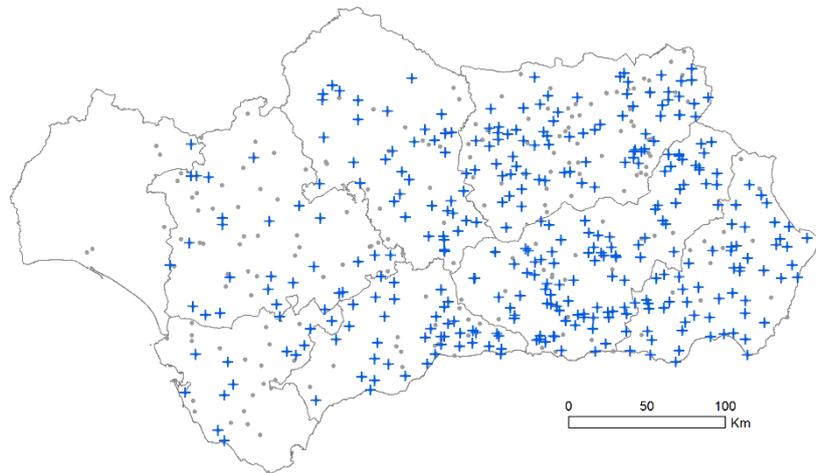


Fig. 5. Spatial distribution of all precipitation stations in which the 3-years cycle has been detected (crosses in blue) above 90% of ACL, and all the other stations used for the analysis (points in grey).

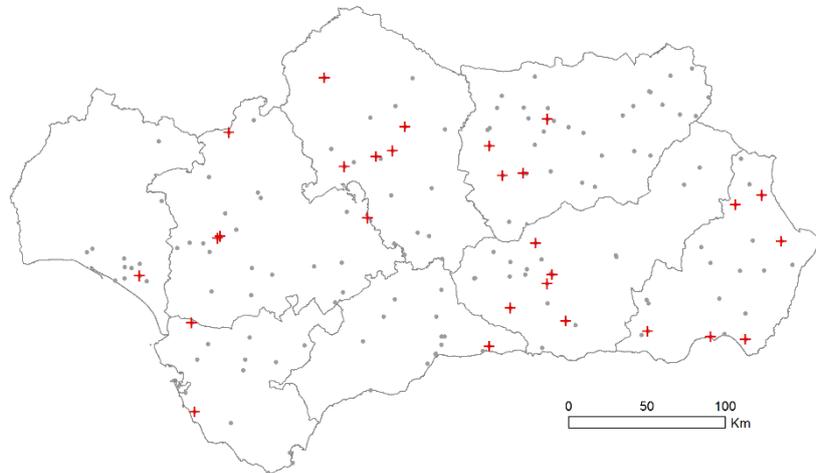


Fig. 6. Spatial distribution of all temperature stations in which the 3-years cycle has been detected (crosses in red) above 90% of ACL, and all the other stations used for the analysis (points in grey).

The same output has been conducted for the 7/8-years cycle by isolating all peaks between 0.154 (6.5 years) and 0.118 (8.5 years); 223 precipitation stations and 23 temperature stations (see Figs 7 and 8).

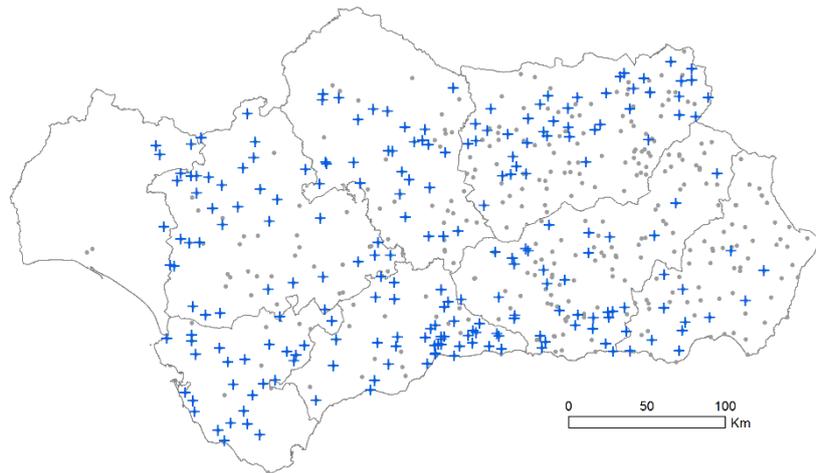


Fig. 7. Spatial distribution of all precipitation stations in which the 7/8-years cycle has been detected (crosses in blue) above 90% of ACL, and all the other precipitation stations used for the analysis (points in grey).

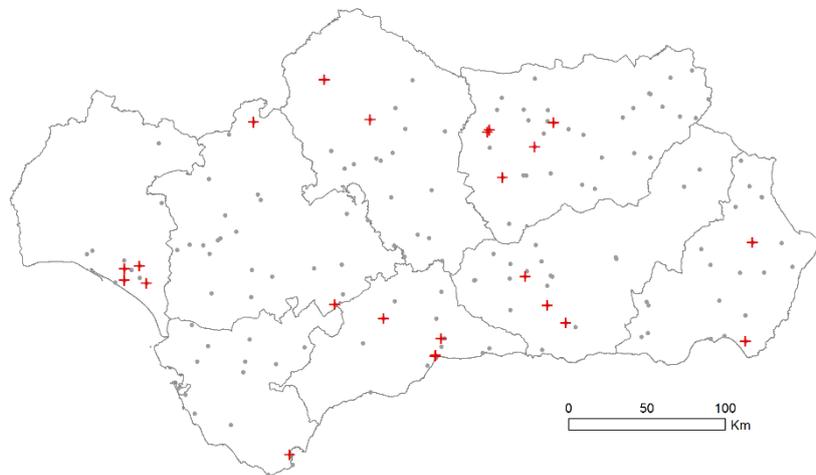


Fig. 8. Spatial distribution of all temperature stations in which the 7/8-years cycle has been detected (crosses in red) above 90% of ACL, and all the other temperature stations used for the analysis (points in grey).

Overall, the spatial distribution of spotted temperature cycles (see Figs 6 and 8) does not show any geographical predominance of one cycle over the other, as opposed to the maps showing spotted precipitation cycles where a general trend can be inferred (see Figs 5 and 7). In the later, there are more 3-years spotted cycles eastwards and more 7/8-years spotted cycles westwards. To confirm it, the two cycles combined into two maps have been plotted; one for precipitation and one for temperature (Figs 9 and 10). The process is based on a preliminary selection of stations in which for each station at least one of the two cycles must be present; if both cycles are detected in the same station, the cycle having more power spectrum is the one counted for that station. In this way, 423 precipitation stations derived from the imposed condition have been analysed. About 261 precipitation stations show predominance of the 3-years cycle and 162 precipitation stations show predominance of the 7/8-years cycle. Similarly, 47 stations were considered for the temperature, resulting in 27 stations having more importance of the 3-years cycle and 20 stations with more influence of the 7/8-years cycle. This process has been followed by an interpolation process in which the ‘Inverse Distance Weighting’ (IDW) technique has been applied. The parameters have been 2 for the power and 5 for the maximum number of neighbours (see Figs 9 and 10).

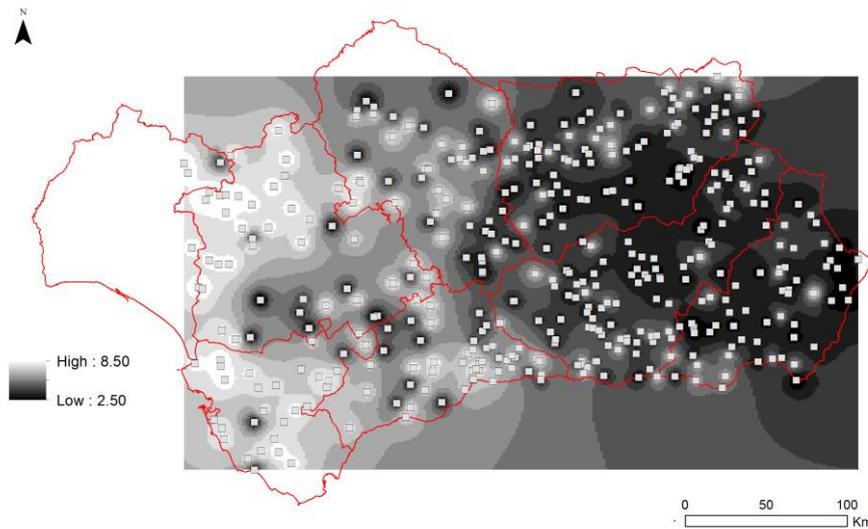


Fig. 9. Spatial interpolation on 423 precipitation stations showing the geographical influence in Andalusia of the 3-years cycle and the 7/8 years cycle. Values in years.

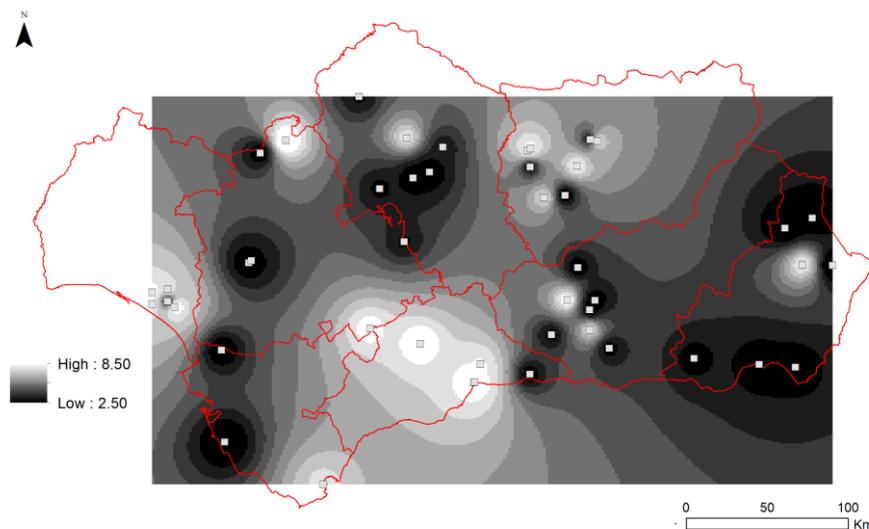


Fig. 10. Spatial interpolation on 47 temperature stations showing the geographical influence in Andalusia of the 3-years cycle and the 7/8 years cycle. Values in years.

4 Interpretation

As above commented, the causes of climate variability in Andalusia from annual to multi-decadal time scales are generally associated with phenomena across the Atlantic Ocean: large scale circulation features of Western Europe and Atlantic Ocean [1], alternation of zonal circulation and meridional circulation in the Atlantic producing shifts of the Azores High [2], persistency and displacement of the Azores High [3], and changes in the North Atlantic Oscillation NAO phases [4]. However, Eastern Andalusia is less influenced by Atlantic air masses and its variability is also influenced by the Mediterranean Sea dynamics [5].

Previous studies applying spectral analysis on climatic datasets reveal cyclic climatic variability in the range of 2 to 250 years, but mainly located in the range of 2 to 11 years. Thus, study on rainfall variability in Southern Spain [6] found peaks above 95% of significance level at 2.1, 3.5, 7-9, 16.7 and 250 years, and a previous study on hydraulic heads across the Vega de Granada' aquifer [7] found a decadal cycle (peaks between 8 and 11 years) and a 3.2-year cycle, amongst others.

In the case study, the registered climatological cycles at around 3 years and 7/8 years fit well with some phenomena mainly associated to the Sun and to oceanic activity. Other cycles have been associated with their most probable climatic origin (Table 1; [13] for a review).

Table 1. Association between detected cycles and their match with other well-known cycles, and the number of precipitation (P) and temperature (T) stations where detection has occurred.

Cycle value in years	Cycle	P	T
2	Quasi-Biennial Oscillation (QBO)	100	23
3	SST of the Mediterranean	410	37
5/6	Harmonic component of the sunspot cycle of 11 years or bear relation with El Niño Southern Oscillation (ENSO)	111	20
7/8	North Atlantic Oscillation (NAO)	290	38
10/11	Sunspot cycles of 11 years	84	17
17	Southern Oscillation Index (SOI)	17	16
19	Luni-solar cycle	19	14
20-25	Hale Cycle	51	49

5 Conclusions

Spectral analysis on a large number of meteorological stations (707 locations), distributed across southern Spain in the region of Andalusia, has allowed the characterization of the spatial variability of climate cycles. Many periodicities have been found. However, because the length of the stations is relatively short (in many cases smaller than 50 years, but always larger than 20 years for the precipitation and larger than 10 years for the temperature) the focus has been the high frequency cycles. It has been shown how the well-known 11 years sun spot cycle is badly represented, while a cycle of 3 years and a cycle in the range from 7 to 8 years are the best well represented by the meteorological stations. It has been seen how the 3-years cycle is better represented in the eastern part of the study area (with more Mediterranean influence), while the 7/8-years cycle is better represented in the western part of the study area (where the influence of the Atlantic is important). The explanation of this spatial pattern is outside the scope of this paper (research in progress). Nevertheless, these results can help meteorologists and climatologists to understand better how weather and climate works in the south of the Iberian Peninsula.

References

1. Esteban-Parra, M.J., Rodrigo, F.S., Castro-Díez, Y.: Spatial and temporal patterns of precipitation in Spain for the period 1880–1992. *International Journal of Climatology*, 18(14), 1557-1574 (1998).
2. Rodrigo, F.S., Esteban-Parra, M.J., Pozo-Vázquez, D., Castro-Díez, Y.: A 500-year precipitation record in Southern Spain. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 19(11), 1233-1253 (1999).

3. Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., Pita-López, M.F.: A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology* 28(11), 1525-1534 (2008).
4. Rodrigo, F.S., Gómez-Navarro, J.J., Montávez-Gómez, J.P.: Climate variability in Andalusia (southern Spain) during the period 1701–1850 based on documentary sources: evaluation and comparison with climate model simulations. *Climate of the Past* 8(1), 117-133 (2012).
5. Fernández-Montes, S., Rodrigo, F.S.: Trends in surface air temperatures, precipitation and combined indices in the southeastern Iberian Peninsula (1970-2007). *Climate Research* 63(1), 43-60 (2015).
6. Rodrigo, F.S., Esteban-Parra, M.J., Pozo-Vázquez, D., Castro-Díez, Y.: Rainfall variability in southern Spain on decadal to centennial time scales. *International Journal of Climatology* 20, 721-732 (2000).
7. Luque-Espinar, J.A., Chica-Olmo, M., Pardo-Igúzquiza, E., García-Soldado, M.J.: Influence of climatological cycles on hydraulic heads across a Spanish aquifer. *Journal of hydrology* 354(1-4), 33-52 (2008).
8. Lomb, N.R.: Least-squares frequency analysis of unequally spaced data. *Astrophysical Space Science* 39, 447–462 (1976).
9. Scargle, J.D.: Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal* 263, 835–853 (1982).
10. Pardo-Igúzquiza, E., Rodríguez-Tovar, F.J.: Spectral and cross-spectral analysis of uneven time series with the smoothed Lomb–Scargle periodogram and Monte Carlo evaluation of statistical significance. *Computers & Geosciences* 49, 207–216 (2012).
11. “Confederación Hidrográfica del Guadalquivir” (CHG), www.chguadalquivir.es.
12. State Meteorological Agency (AEMET), <http://www.aemet.es>.
13. Rodríguez-Tovar, F.J.: Orbital Climate Cycles in the Fossil Record: From Semidiurnal to Million-Year Biotic Responses. *Annual Review of Earth and Planetary Sciences* 42, 69–102 (2014).

Real time anomaly detection in network traffic time series

Sergio Martínez Tagliafico¹, Gastón García González¹, Alicia Fernández¹,
Gabriel Gómez Sena¹, and José Acuña^{1,2}

¹ Instituto de Ingeniería Eléctrica, Facultad de Ingeniería,
Universidad de la República, Uruguay

{sematag,gastong,alicia,ggomez,acuna}@fing.edu.uy

² Telefónica Móviles, Uruguay

jose.acuna@telefonica.com

Abstract. Anomaly detection is a relevant field of study for many applications and contexts. In this paper we focus in on-line anomaly detection on unidimensional time series provided by different network operator equipments. We have implemented two detection methods, we have optimized them for on-line processing and we have adapted them for integration into a testbed of a well known Hadoop big data platform. We have analyzed the behavior of both methods for the particular datasets available but we also have applied the methods to a publicly available labeled datasets obtaining good results.

Keywords: Anomaly Detection · Kalman Filter · Hadoop.

1 Introduction

Detecting anomalies in the behavior of multiple variables gathered from the network infrastructure is an essential task to be able to detect failures and also to react as soon as possible to solve the issues. Although common monitoring systems can be able to report failures in hardware equipment or software services, they do not normally provide alarms when the quality of one service is being degraded. Moreover, an efficient anomaly analysis can be useful to detect performance issues, attacks to network security and fraud attempts.

Although anomalies analysis in telecommunications traffic is a mature area [2][12][5][3][9] with approaches based on statistical methods [11][8], the emergence of big data platforms which enable the processing of massive and diverse data, poses new opportunities and challenges. Particularly the development of analytics for platforms that solve the detection of anomalies of large volumes of data simultaneously is a important issue [1][4][13] and determines the need of adapting the algorithms implementations for parallel processing.

All the variables considered in this work have some kind of periodic behavior so our methods will provide a prediction based on the statistical of the past samples of the variable. Also a decision process is needed to signal which samples should be considered as anomalies. As stated, the ability to detect anomalies in

real time is a value added feature because it enables a fast reaction to reduce the service unavailability or degradation time for the customer.

The main contribution of this work is the development of an anomaly detection strategy based on stochastic modeling and the implementation on a Hadoop³ big data platform testbed. We have also implemented a classification strategy based on Parzen Windows and we have compared both methods. For the validation process we used real data provided by a network operator and we also tested our method with publicly available labeled datasets.

In this document, we define in Section 2 the relevant types of anomalies for the specific application field and the proposed methods. In section 3 we depict some implementation details. In Section 4 we show some selected validation scenarios and finally in Section 5 we conclude and identify some possible future works.

2 Strategy for anomaly detection

2.1 Type of anomalies

We can define an “anomaly” as a set of values of a variable that are far from its normal or expected values. Therefore, we need to define a region of the feature space of the data that represents its normal behavior. Any data out of the normal region, will be consider as an anomaly.

The definition of the normal feature space can be a complex task. In some cases the normal behavior feature space may vary along the time and also it is sometimes difficult to have labeled traffic to aid the normal region definition.

Based on [5], anomalies can be classified as:

- Point Anomalies: A single value can be considered as anomalous with respect to the rest of the data.
- Contextual Anomalies: A single value is anomalous in the context of its neighbors values but in other cases can be considered normal.
- Collective Anomalies: A collection of related data values is anomalous with respect to the entire data set.

The data used for this work is non labeled unidimensional time series obtained from telecommunication infrastructure, for instance, the interface traffic from a router. The relevant anomaly types for this scenario are “Point Anomalies” and “Collective Anomalies”. When we find an abnormal change of the series value respect to the expected value for that time instant, we will be in presence of a point anomaly. Besides, we can find that some values have a slight but prolonged withdrawal in time so as to be considered a collective anomaly.

For this particular context it is also relevant the ability to perform anomaly detection in real time. The scenario is that a monitoring system will produce a stream of sample values for the chosen variable. The cadence of the variable samples will depend on the monitoring system configuration, typically in the

³ <http://hadoop.apache.org/>

order of minutes. The detection process will receive a streaming of values and it should produce an indication if an anomaly is detected. It is obvious that the detection process cannot take longer than the sample interval.

2.2 Point and Collective anomalies: ARIMA+Kalman models

There are a lot of techniques mainly used for anomaly detection in this scenarios [5] and our first approach is to adjust a stochastic model for the time series and define an anomaly based on whether an observation is suspicious of not being generated by this model.

Let be y_k a time series and $Y_k = (y_1, \dots, y_k)$ the vector of observations which represents its evolution up to time k , the detection process used is based on obtaining the distribution of the series in time $k + 1$ given its evolution Y_k . We will write this distribution as:

$$p(y_{k+1}|Y_k) = p(y_{k+1}|y_k, \dots, y_1)$$

In this approach we use ARIMA (Autoregressive Integrated Moving Average) models represented as a state space model [7].

$$y_k = \mathbf{Z}\mathbf{x}_k + \varepsilon_k, \quad \{\varepsilon_k\} \text{ iid} \sim N(0, \sigma_\varepsilon^2) \quad (1)$$

$$\mathbf{x}_{k+1} = \mathbf{T}\mathbf{x}_k + \mathbf{R}\boldsymbol{\eta}_k, \quad \{\boldsymbol{\eta}_k\} \text{ iid} \sim N(\mathbf{0}, \mathbf{Q}) \quad (2)$$

where \mathbf{Z} , \mathbf{T} and \mathbf{R} are fixed matrices. Matrices are represented by bold letters in our notation.

As every distributions in this model are Gaussian, the distributions $p(y_{k+1}|Y_k)$, $p(\mathbf{x}_k|Y_k)$ y $p(\mathbf{x}_{k+1}|Y_k)$ are also Gaussian. Lets call $\hat{\mathbf{x}}_{k|k} = \mathbf{E}[\mathbf{x}_k|Y_k]$, $\hat{\mathbf{x}}_{k+1|k} = \mathbf{E}[\mathbf{x}_{k+1}|Y_k]$, $\mathbf{P}_{k|k} = \text{Var}[\mathbf{x}_k|Y_k]$ y $\mathbf{P}_{k+1|k} = \text{Var}[\mathbf{x}_{k+1}|Y_k]$, then

$$p(\mathbf{x}_k|Y_k) = N(\hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}) \quad (3)$$

$$p(\mathbf{x}_{k+1}|Y_k) = N(\hat{\mathbf{x}}_{k+1|k}, \mathbf{P}_{k+1|k}) \quad (4)$$

$$p(y_{k+1}|Y_k) = N(\mathbf{Z}\hat{\mathbf{x}}_{k+1|k}, \mathbf{Z}\mathbf{P}_{k+1|k}\mathbf{Z}' + \sigma_\varepsilon^2) \quad (5)$$

where $\hat{\mathbf{x}}_{k|k}$, $\hat{\mathbf{x}}_{k+1|k}$, $\mathbf{Z}\hat{\mathbf{x}}_{k+1|k}$ are minimum variance unbiased estimators (MVLUE) of \mathbf{x}_k , \mathbf{x}_{k+1} and y_{k+1} respectively given Y_k .

The on-line estimation of this distribution can be done by the well known Kalman Filter equations.

$$\begin{aligned} e_k &= y_k - \mathbf{Z}\hat{\mathbf{x}}_{k|k-1}, & F_k &= \mathbf{Z}\mathbf{P}_{k|k-1}\mathbf{Z}' + \sigma_\varepsilon^2 \\ \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_k + \mathbf{P}_{k|k-1}\mathbf{Z}'F_k^{-1}e_k, & \mathbf{P}_{k|k} &= \mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1}\mathbf{Z}'F_k^{-1}\mathbf{Z}\mathbf{P}_{k|k-1} \\ \mathbf{x}_{k+1|k} &= \mathbf{T}\hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k e_k, & \mathbf{P}_{k+1|k} &= \mathbf{T}\mathbf{P}_{k|k-1}(\mathbf{T} - \mathbf{K}_k\mathbf{Z})' + \mathbf{R}\mathbf{Q}_k\mathbf{R}' \end{aligned}$$

Using this result, we can manage to efficiently update the distributions and the state as soon as each sample arrives to the system. As can be seen, only matrix sum and product are involved. This result will be very important for on-line anomaly detection.

Anomalies definitions The first stage in our anomaly detection process is to identify an abrupt change of the data values from the expected value for a given time instant. This condition can be defined as:

Definition (*Type A anomalies - Point Anomalies*): Let be y_k a Gaussian process modeling a time series, $Y_n = (y_1, \dots, y_n)$ the set of realizations representing the evolving up to the time instant n and $p(y_{n+1}|Y_n) = N(m_{n+1}, P_{n+1})$ the conditional distribution of y_{n+1} given its evolution up to time instant $k = n$. A point type anomaly occurs at $k = n + 1$ if

$$\|y_{n+1} - m_{n+1}\| > r * \sqrt{P_{n+1}}$$

that means at $k = n + 1$ the value is more than r standard deviations far from its expected value.

The second stage is intended to detect an slight but sustained in time shift from the expected series value.

Definition (*Type B anomalies - Collective Anomalies*): Let be y_k a Gaussian process modeling a time series, $\{y_{n-(l-1)}, \dots, y_n\}$ the last l observations of the series and $p(y_{k+1}|Y_k) = N(m_{k+1}, P_{k+1})$ the conditional distribution of y_{k+1} given its evolution up to the time instant k . A collective anomaly occurs at $k = n$ if

$$y_k - m_k < \sqrt{P_k} \quad \forall k = n - (l - 1) : n$$

or

$$y_k - m_k > \sqrt{P_k} \quad \forall k = n - (l - 1) : n$$

2.3 Point anomalies: Parzen windows

The analyzed data has an intrinsic weekly periodicity, so we have chosen the Parzen windows method⁴, to measure how near is a sample from its nearest neighbors.

For that purpose, we began constructing a circular buffer with the accumulated samples of the last week indexed by its collected time-stamp. When a new test sample arrives, a cluster of an hour history of samples centered in its time-stamp is created.

For each value in the cluster, we considered a gauss function with mean value the sample and a presetted variance value. The new test sample is then evaluated with each of the cluster functions and the accumulated value is compared with a presetted threshold value. If the sum is above the threshold, it means that in the environment of the current sample there are several samples of the cluster. Otherwise it means that the current sample is far from the values of the cluster and it is labeled as an out-lier. The outliers are discarded for updating the circular history buffer.

⁴ Inspired on the density estimation by Parzen windows.[6]

Method The circular buffer can be represented as $W = (x_t^n, \dots, x_{t+m}^n)$, where the index t indicates the temporal position in the buffer, and the index n indicates the week where the sample belongs. At the beginning all the samples of the buffer are from the same week.

When the samples begin to arrive in real time, the current sample x_k^n is taken and a cluster w_k is generated from the samples of the circular buffer taken half hour back and half hour forward referring to the time of the current sample.

$$w_k = (x_{k-h/2}^n, \dots, x_{k-1}^n, x_k^{n-1}, x_{k+1}^{n-1}, \dots, x_{k+h/2}^{n-1})$$

Then, for each cluster sample, a window function is used, where the value of the current sample is evaluated. In this case we will use a gauss function where the mean will be the value of the sample and the variance σ will be a presetted parameter.

$$p_k = (N(x_{k-h/2}^n, \sigma), \dots, N(x_{k-1}^n, \sigma), N(x_k^{n-1}, \sigma), N(x_{k+1}^{n-1}, \sigma), \dots, N(x_{k+h/2}^{n-1}, \sigma))$$

All the values obtained from the evaluated functions are added and the result is compared with a threshold value U . If the sum is above the threshold it means that in the environment of the current sample there are several samples of the cluster nearby. Otherwise it means that the current sample is very far from the values of the cluster, then it is labeled as out-lier. The outliers are discarded when updating the circular buffer.

$$\sum_{i=k-h/2}^{k+h/2} p_{ki}(x_k^n) < U. \quad (6)$$

3 Implementation highlights

The implementation was done in Python and has been integrated into a Hortonworks HDP⁵ platform testbed. The on-line data ingestion was done through Apache NiFi and Apache Kafka and then the data processing was done running the python code with pySpark. The implemented software was thought with a modular architectural design in mind so as to enable an easy change of the detection algorithms.

For the modeling phase, we use the module `statsmodels.tsa.statespace`⁶ and particularly the class `statsmodels.tsa.statespace.sarimax.SARIMAX`, which allow modeling by ARIMA space state models.

The anomaly detection algorithm is implemented in the module `anomaliasKF`⁷. The two main components of the anomaly detection process for real time detection are implemented in the class `AnomalyDetector.py` helped with the `pykalman`⁸ module for the Kalman Filter equations.

⁵ <https://hortonworks.com/products/data-platforms/hdp/>

⁶ (<http://www.statsmodels.org/dev/statespace.html>)

⁷ <https://iie.fing.edu.uy/sematag/anomalias/>

⁸ <https://pykalman.github.io/>

4 Experiments and results

4.1 Dataset characterization

The available time series for evaluating the proposed methods come from our partner, a mobile operator in Uruguay. Table 1 show the main characteristics of the series and figure 1 illustrate the general behavior of each one. For each series, we use the data of the first days to adjust the stochastic model and the rest of the data was used to test the method.

Series	Name	Sample frequency	Duration	Train set
1	Mobile_Data_Downlink_Bytes	1 sample each 5 minutes	28 days	First 7 days
2	Voice_Calls_Originated	1 sample each 1 hour	28 days	First 7 days
3	Accounting_Mobile_Data	1 sample each 5 minutes	18 days	First 4 days
4	SMS_Originated	1 sample each 5 minutes	28 days	First 7 days

Table 1: Time series used for validation

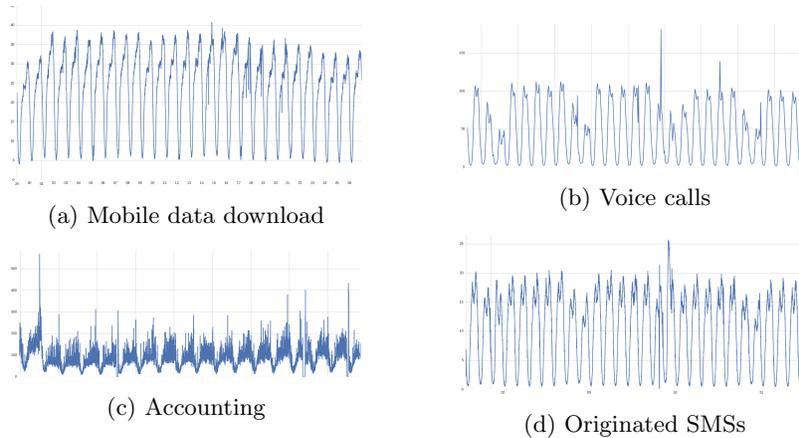


Fig. 1: Global view of the time series analyzed

Among the referred real datasets and to improve the evaluation of the proposed strategy in other types of time series, we have tested our approach in two time series from the Numenta Anomaly Benchmark (NAB)[10]: the hourly demand for New York City taxis and the real time traffic data (occupancy) from the Twin Cities Metro area in Minnesota.

4.2 ARIMA+Kalman results

The proposed approach was applied to detect anomalies for the four time series referred in Table 1. Figure 2 shows some anomalies (in red) detected on the series. In all cases anomalies were detected in zones where the series have an obvious abnormal behavior. Moreover, no relevant false alarms were generated, only some cases were generated immediately after a true anomaly was detected as shown in figure 2b.

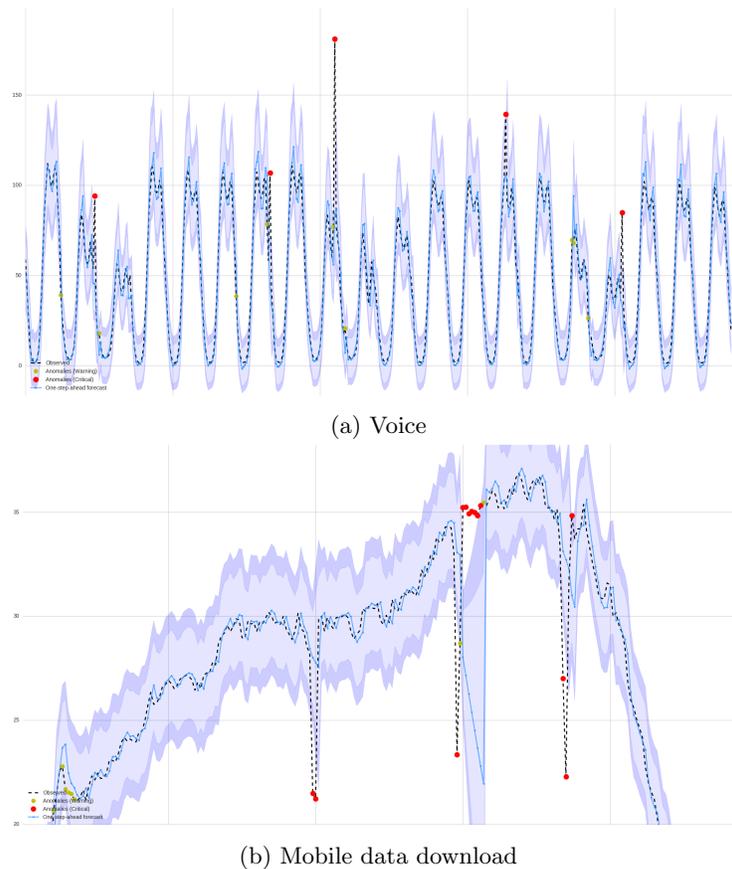
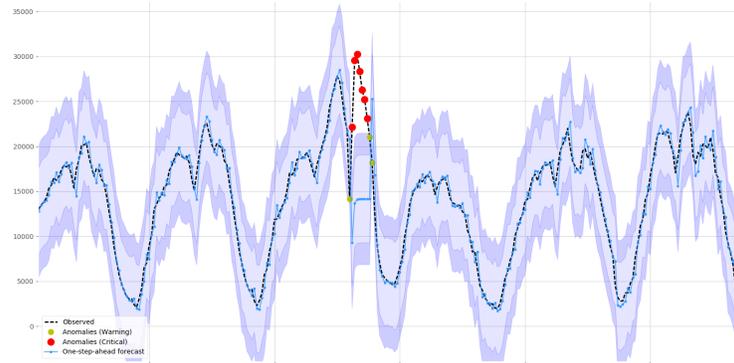
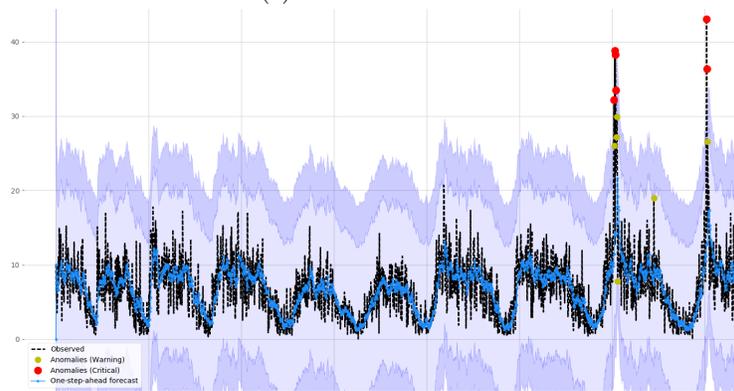


Fig. 2: Anomaly detection results for Mobile data download and Voice data

We have obtained similar results for the NAB time series, as can be seen in figure 3. In all cases, low false alarms were generated and labeled anomalies were completely detected. Also good performance is achieved for others NAB time series.



(a) NYX taxi demand



(b) Minnesota Metro occupancy

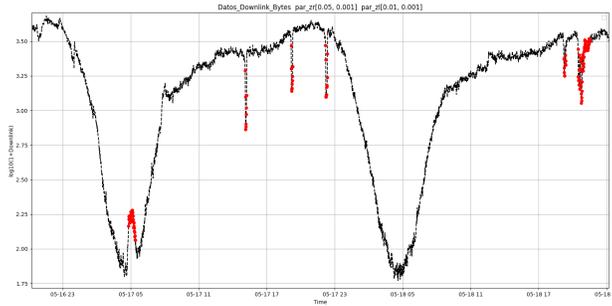
Fig. 3: Anomaly detection results for NYC taxi demand and Minnesota Metro occupancy

4.3 Parzen Windows results

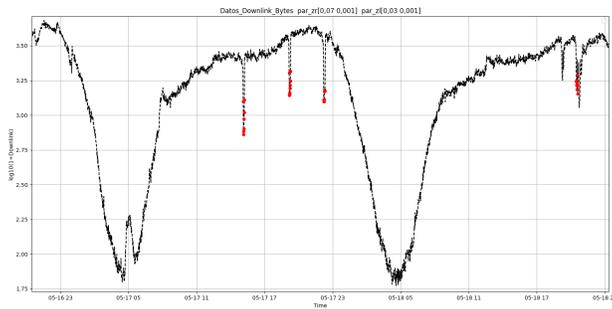
The implementation of the Parzen Windows method was also done in Python and in this section we are presenting the results for the mobile data download series. As explained in section 2.3 the method has two parameters, the window width σ and the decision threshold U . In this type of series during the night and part of the morning the difference in values between consecutive samples is larger than during the rest of the day, due to the rapid fall of the activity at the end of the day and the rapid reactivation of the activity by the morning. Then it is convenient to use different parameters for these two scenarios. When the difference in values between consecutive samples is large, it is better to use a larger window width (σ). For this test we have used this parameters:

- $\sigma_{night} = 0.05$, $U = 0.001$. For the night and the early morning.

– $\sigma_{day} = 0.01, U = 0.001$. For the rest of the day.



(a) Parzen method applied to mobile data download. Parameters: $\sigma_{night} = 0.05, \sigma_{day} = 0.01$ and $U = 0.001$.



(b) Parzen method applied to mobile data download with a wider window. Parameters: $\sigma_{night} = 0.07, \sigma_{day} = 0.03$ and $U = 0.001$.

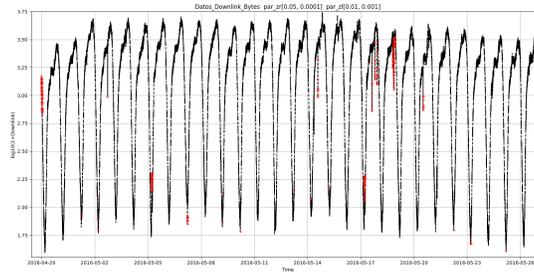
Fig. 4: Anomaly detection results for mobile data download

As shown in Figure 4a the outliers (in red), correspond to the detected point type anomalies. The parameters were adjusted until an acceptable result was obtained. In Figure 4b the effect of a wider window is shown ($\sigma_{night} = 0.07$, and $\sigma_{day} = 0.03$).

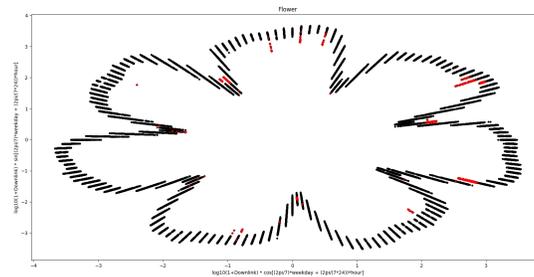
As can be seen, the amount of red samples decreased because the model is more tolerant when the window width increases.

Another way to visualize the data is the one shown in Figure 5b. This representation is quite helpful to find anomalies at a glance and to depict the weekly evolution of the series.

Figure 5a shows a series corresponding to one month of data collected and classified. In figure 5b you can see the same series represented as the flower. An



(a) Series corresponding to one month of data.



(b) The same series represented in a flower of a week.

Fig. 5: Flower representation

entire turn of the flower represents a week’s time, in figure 5b are the four weeks corresponding to the month of figure 5a. In figure 6 shows an example for three day of the week.

5 Conclusions and future work

We have implemented an algorithm for real time anomaly detection using ARIMA models and Kalman filtering, obtaining good performance results for our operator partner time series. Both point and contextual anomalies can be detected. The approach was also tested with publicly available labeled time series showing good results. The use of Kalman filtering enable us the use of the algorithm for real time anomalies detection.

We have also implemented a Parzen Windows oriented method for point anomalies which is simple and requires very few calculation resources. We have also obtained interesting detection results.

Both methods were implemented and integrated as a module into a hadoop platform sandbox, enabling the later integration to the operator production systems.

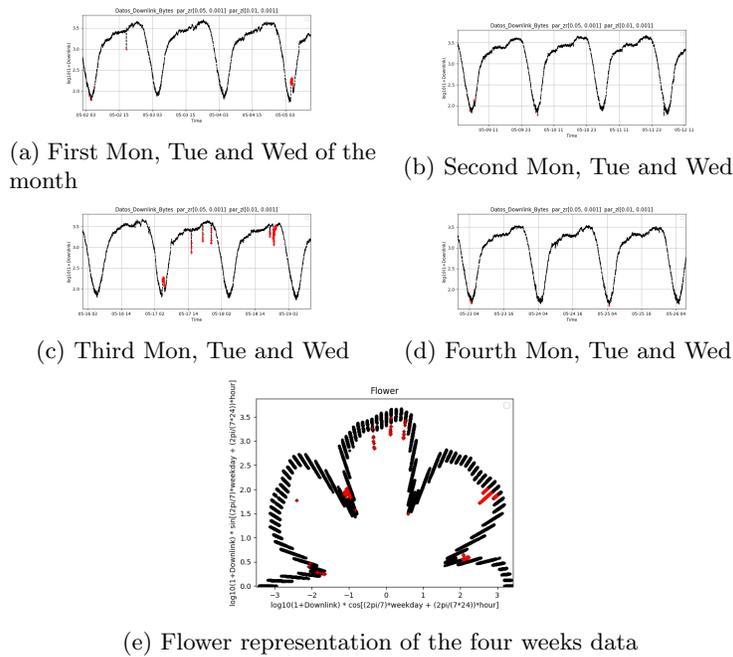


Fig. 6: An example of how the data is represented with the flower

Based on the experience and results of this work, we have found some relevant point to work on.

First of all, the anomaly definition is a relevant issue that condition the detection process. For now, we have worked with a Gaussian model of the series but an hypothesis test over the data distribution characteristics can be faced.

Regarding the ARIMA+Kalman Filter implementation we used ARIMA state space models but other state space model families can be explored, for instance structural models or dynamic factor models⁹. Moreover, we want to work on some kind of automatic learning for the model in the training phase. We have also shown that after an anomaly has been detected the subsequent prediction is not good, so we want to improve the detection for this anomaly stage, perhaps introducing robust Kalman filtering.

Regarding Parzen Windows method, we need to improve the parameters adjusting for the different stages of the series values. We want to introduce some kind of automatic adjusting depending on some statistical properties instead of setting them manually.

⁹ <http://www.statsmodels.org/dev/statespace.html>

Acknowledgements

This work was partially supported by Telefónica Móviles (Uruguay) and the Groups Program of the Comisión Sectorial de Investigación Científica, Universidad de la República (Uruguay). The authors are thankful to both institutions.

The authors would like to thank Pedro Casas for its good advice at the beginning of the project.

References

1. Bär, A., Finamore, A., Casas, P., Golab, L., Mellia, M.: Large-scale network traffic monitoring with dbstream, a system for rolling big data analysis. In: Big Data (Big Data), 2014 IEEE International Conference on. pp. 165–170. IEEE (2014)
2. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys & Tutorials* **16**(1), 303–336 (FebJan 2014). <https://doi.org/10.1109/surv.2013.052213.00046>, <http://dx.doi.org/10.1109/surv.2013.052213.00046>
3. Brutlag, J.D.: Aberrant behavior detection in time series for network monitoring. In: *LISA*. vol. 14, pp. 139–146 (2000)
4. Casas, P., Soro, F., Vanerio, J., Settanni, G., D’Alconzo, A.: Network security and anomaly detection with big-dama, a big data analytics framework (2017)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. *ACM Comput. Surv.* **41**(3), 15:1–15:58 (Jul 2009). <https://doi.org/10.1145/1541880.1541882>, <http://doi.acm.org/10.1145/1541880.1541882>
6. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. Wiley-Interscience, second edition edn. (2000)
7. Durbin, J., Koopman, S.J.: *Time series analysis by state space methods*, vol. 38. Oxford University Press (2012)
8. Knorn, F., Leith, D.J.: Adaptive kalman filtering for anomaly detection in software appliances. In: *INFOCOM Workshops 2008*, IEEE. pp. 1–6. IEEE (2008)
9. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. In: *ACM SIGCOMM Computer Communication Review*. vol. 34, pp. 219–230. ACM (2004)
10. Lavin, A., Ahmad, S.: Evaluating real-time anomaly detection algorithms—the numenta anomaly benchmark. In: *Machine Learning and Applications (ICMLA)*, 2015 IEEE 14th International Conference on. pp. 38–44. IEEE (2015)
11. Mazel, J., Casas, P., Labit, Y., Owezarski, P.: Sub-space clustering, inter-clustering results association & anomaly correlation for unsupervised network anomaly detection. In: *Proceedings of the 7th International Conference on Network and Services Management*. pp. 73–80. International Federation for Information Processing (2011)
12. Soule, A., Salamatian, K., Taft, N.: Combining filtering and statistical methods for anomaly detection. In: *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*. pp. 31–31. IMC ’05, USENIX Association, Berkeley, CA, USA (2005), <http://dl.acm.org/citation.cfm?id=1251086.1251117>
13. Vanerio, J., Casas, P.: Ensemble-learning approaches for network security and anomaly detection. In: *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*. pp. 1–6. ACM (2017)

Spacecraft Mission Control Center Resource State Estimation and Contingency Forecasting

^{1,2}Natalia Bakhtadze, ¹Denis Elpashev, ¹Alexey Lototsky,
¹Vladimir Lototsky,
¹Eddy Zakharov

¹V.A. Trapeznikov Institute of Control Sciences,
Moscow, Russia, sung7@yandex.ru

²Bauman Moscow State Technical University, Moscow, Russia

Abstract. The paper presents the concept of on-line system for estimating the state of the MCC resources and algorithms for forecasting of contingencies. Resource state prediction methods based on the development of a and a machine learning techniques called association rules are presented

Keywords: Binary Resource Model, Telemetric Data Analysis, Knowledge Base Prediction

The automated spacecraft control system (SC) is a complex of ground control and on-board control system. The system ensures the control of the required ballistic structure during system operation, and also the fulfillment of the required quality criteria. The effectiveness of the control complex is its ability to fulfill the plan and the quality of performing typical operations implemented by the system. The system should analyze plan and predict the ability to perform work plan generated when applying the authorized resources (control and monitoring ones). When the timing of launches of spacecraft and spacecraft (SC) flight programs programs is changed, situational analysis is carried out, and also the degree of availability of resources for use is analyzed. Further, the system should generate operational recommendations for use planning control and monitoring tools. Based on the overall results of the planning, the planning of the specific control and measurement tools and the development of recommendations for improving the efficiency of operational planning are analyzed. The basis for adjusting the current planning is the following changes in the initial data:

- occurrence of conflict situations on the use of resources as a result of non-standard malfunctions in the operation of equipment and software;
- changes in terrestrial and space conditions in the event of malfunction or emergency situation on board the spacecraft and ground control resources;
- change the launch date and launch time of the spacecraft;
- change in radio-electronic, meteorological, geo- and heliophysical conditions;
- change of maintenance plans for resources, time for other operational activities;
- and etc.

To fulfill these tasks, an information support system for operational control should be developed. Such a system may be a system of preventive control based on predictive models in real-time analysis of data, describing the processes as the work of on-board systems and the operation of communication systems, as well as the ground control

complex. Predictive models that support decision-making on the control of SC are developed on the base of data mining algorithms [1].

To determine the values of flight parameters and specific points in times of abnormal situations, telemetric information (TMI) is processed and analyzed - from the initial data from the launching complex and throughout the flight. System state is a vector whose components are defined by the various elements of the spacecraft subsystems, considered as a control system. In this case, we assume that the state of the spacecraft is regular, if at the same time: 1) all of its subsystems are in a "NORMAL" state and 2) the daily program is executed. One type of state parameters have a value of 0 or 1 ("works" - "not working"). Others, such as the values of certain characteristics of the onboard systems, taking values from a specific range of permissible values. As a result, the resource state can be encoded by a binary chain with the identifier. For the generated binary chain, prediction can be obtained, using data mining methods. It seems expedient to use methods that have been called the search for associative rules [3]. This method allows:

- Without preliminary statistical analysis, to reveal the latent statistical dependence between the states of various resources at different times
- Operate variable resource state, which differ in their characteristics and method of formalization.

One of the most common is the algorithm a priori [4], based on the notion of a popular (often occurring) set of zeros and ones. Under a common occurrence is understood such a set, the frequency of occurrence of which in the total aggregate of transactions exceeds a certain preset level. In our case, the associative rule consists of two sets of resource values that form, respectively, the condition (antecedent) and consequence (consequent) [5], related by the relation "from X follows Y". In particular, it is possible to consider X as the current set of values of resource states, and as Y - a set of these values at a subsequent moment of time, i.e. prediction. Next, the kits are analyzed at time points, separated by several positions. Thus, the prediction model is adjusted, depending on the values of the individual parameters, a few cycles back.

References

1. Kelly J.D., 2006. Logistics: the missing link in blend scheduling optimization / J.D. Kelly // *Hydrocarbon Processing*. June. P. 45-51
2. Bakhtadze N., Lototsky V., Maximov E., Pavlov B. Associative Search Models in Industrial systems / In: *Pr. of IFAC Workshop of Intelligent Manufacturing Systems*, Alicante, Spain, 2007.
3. Agrawal, R., Imielinski, T., Swami, A., 1993. Mining Associations between Sets of Items in Massive Databases. In *Proc. of the 1993. ACM-SIGMOD Int'l Conf. on Management of Data*. P. 207-216.
4. Agrawal, R., Srikant, R. Fast, 1994. Discovery of Association Rules / In *Proc. of the 20th International Conference on VLDB*, Santiago, Chile.
5. Ian H. Witten, Eibe Frank and Mark A. Hall, 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition / Morgan Kaufmann, P. 664. — [ISBN 9780123748560](https://doi.org/10.1002/9780123748560).

Towards Hybrid Prediction over Time Series with Non-Periodic External Factors

Xavier Fontes^{1,2*} and Daniel Silva^{1,2}

¹ Faculty of Engineering of the University of Porto, Portugal

² Artificial Intelligence and Computer Science Laboratory
{xavier.fontes, dcs}@fe.up.pt

Abstract. Currently we see prediction problems being tackled with many different types of models. More recently, artificial neural networks have come to gain a lot of popularity in time series forecasting. The focus of this work is to lay the foundation for an hybrid approach at time series forecasting. For that, we research methods that have been widely used for prediction in time series, and then choose five different methods to compare. By using a university student parking lot as a case study, we experiment with different parameter configuration for each method. Random Forest achieved the best results, with 89% accuracy, followed closely by Gradient Boosting and Decision Trees, with little over 86% each. These may be the most promising methods for a next step of hybridization, in an attempt to create a method with even better results.

Keywords: Time Series Forecasting · Prediction Models

1 Introduction

The last decade has brought forward several developments in machine learning. Areas like data mining, data science, big data and machine learning are hot topics for dealing with a recurrent problem, prediction. Prediction over time series comes with a set of added difficulties when compared to regular classification problems. Going a step further in this direction, time series with non-periodic stimuli are even more challenging, as they include other information on top of the time series.

This work's goal is to evaluate the performance of different classification methods for this kind of problem, using a University parking lot as case study. The work described herein is the first step towards a more ambitious goal of obtaining a hybrid approach for time series forecasting.

The remainder of this article is organized as follows. Section 2 presents a review of methodologies that have been used for prediction in time series, as well as some works that employed hybrid approaches. In section 3 we formulate the problem and approach to it. Section 4 outlines the experimental setup followed in this work. In section 5 we describe the obtained results and discuss them. Finally, section 6 presents the conclusions as well as remarks on future work.

* The first author was supported by the Calouste Gulbenkian Foundation, under a New Talents in Artificial Intelligence Program Grant.

2 State of The Art

The advances in Artificial Neural Networks (ANNs) has found a plethora of uses in prediction problems, in a variety of domains such as financial market [43], weather forecasting [2] or company bankruptcy [34]. Several types of ANNs have been used in these problems.

Radial Basis Function (RBF) Networks are a type of ANN used in supervised learning, in problems of classification, regression and time series prediction [35]. RBFs are advantageous for their influence in linear models, keeping the mathematics simple and calculations relatively light. However, they might not be as accurate when dealing with more complicated behaviors and patterns. RBFs have successfully been allied with clustering algorithms in [3] and with auto regressive methods in time series forecasting in [55].

Introduced in 1988, Probabilistic Neural Networks introduce a key difference in ANNs, replacing the commonly used sigmoid activation function with an exponential function [42]. It has been compared to other methods in predicting stock trends [40] and as a metric of performance in evaluating other models as seen on [36]. It is faster to train than back-propagation networks and is strong against noise examples [49]. However, it requires more memory than other models and performs slower on unseen cases.

Referenced in 1988 [12], and perhaps one of the most recognized ANN models today, Convolutional Neural Networks (CNNs) take advantage of current computational power to achieve excellent results in areas such as image pattern recognition. In [4] a CNN model was used in time series forecasting of financial data. CNNs are associated with deep learning, because of their capability of extracting low- and high-level features from data, and also with highly multi-layered networks. The effectiveness of this method depends on the solution's convergence to an acceptable value [20]. One other problem associated with CNNs (and ANNs in general) is overfitting, ie, creating a model that is exceptionally good on the trained data but not that good on generalization for new data, as discussed in [52]. In prediction problems, an interval discretization can be made, so as to associate the output with a few classes [29].

Recurrent Neural Networks (RNNs) are a class of ANNs encompassing multiple architectures, such as fully recurrent, echo state or long short-term memory, among others [15]. RNNs use nodes that keep a state, an impulse that keeps recurring and affecting the calculations even after the input has passed through. One disadvantage is that its recurrent nature makes RNNs harder to train, but advances in optimization methods and novel network architectures can mitigate this problem [32]. RNNs have been used in multivariate time series even when in the presence of missing values [6].

A Long Short-Term Memory (LSTM) unit is a basic component with an input gate, an output gate and a forget gate. Using LSTM units in a RNN gives origin to an LSTM network [25]. LSTM has some advantages such as resistance to noise and good performance without much fine-tuning. It carries, however, a larger number of weights. One example of its usage in time series is in traffic flow prediction in [11].

Introduced in 1990 [17], Stochastic Neural Networks are built introducing random fluctuations in a neural network. Their utility lies in the ability to help a problem escape from local solutions. Work on forecasting network loads has been done using this method [38] and most recently it was used in forecasting stock price fluctuations [44].

Echo State Networks are a specific architecture of RNNs that assume a type of reservoir to drive the input signal to the desired output. They have achieved good results at reproducing certain time series [30]. They are computationally efficient and easy to use [23]. A recent approach to time series using echo state networks is present in [50], where it is used in multivariate time series prediction. Likewise, in [9] this type of model is used in power systems load modeling.

Aside from neural networks, several other models have been thoroughly used in prediction problems such as in [45] with traffic flows, and thoroughly compared to other methods as in [51] and [22].

Naive Bayes classifiers are a family of probabilistic classifiers inspired on Bayes' theorem [5]. They appeared in the 1950s [39] and, as the theorem upon which they are based, assumes strong independence between features. An ideal Bayesian classifier is one that would classify a sample based on the probability of each class given the sample's features [18]. These classifiers are relatively easy to implement, also being used in classification and regression problems [22].

K-nearest neighbor has been thoroughly used in classification problems [49]. It assigns a given input the classification of the nearest classified points. One disadvantage is the choice of k : a small k increases sensitivity to sparse data and local noise, while a large k means including in the choice possible outliers. Another drawback is making sure the distance function is meaningful [31].

Going back to at least 1979 [46], Decision Trees (DTs) have shown good results in classification and forecasting problems. Being a white-box method, DTs allow users to understand the rationale behind a given decision. The complexity of the problem, however, can make these interpretations harder to understand. DTs have also been applied to time series [21], as well as in hybrid modeling, alongside neural networks [43].

Tree structures are also useful in ensemble models, like Random Forests (RFs). RFs combine several weak tree models to obtain better results by training several fully grown trees with the intuition to reduce variance and average out possible mistakes [19]. RFs have been used for time series regression of illnesses [48] and as a classifier for land cover [37].

Another way to simulate deep learning was proposed in 2017 with a model called deep forest [56]. It is based on decision tree ensembles and is claimed to rival deep ANNs in some of the latter's most successful domains. One of its key advantages is fewer hyper-parameters to tune than conventional ANNs.

First developed in 1995 [7], Support Vector Machines (SVMs) were introduced as a binary classification model, but can be extended to multiple-class problems. It is based on the idea of separating values of different classes by a hyper plane with its margins as big as possible. A problem that may arise with this

method is that points may not be linearly separable, even in an high dimensional data plane. SVMs have also been applied in time series, as in [47].

Like RFs, Gradient Boosting (GB) models fit the category of ensemble-type models. GB uses a forward stage-wise additive strategy to minimize a loss function. GB has been used in predicting waste patterns [24] and travel time [54].

Developed in 1958 [8], a Logistic Regression (LR) model predicts a two-valued variable based on independent features. LR analysis has the capability of separating distinct sets, when the dependent variable shows dichotomy and the independent variables are continuous or discrete. The distinction is performed through establishing the discrimination rules [1]. In LR where the output variable is not binary, a multinomial LR model can be used [16]. One caveat of this model is that it assumes inter-independence between the input variables, which may not be the case for most problems.

The Auto Regressive Moving Average model and its generalization Auto Regressive Integrated Moving Average (ARIMA) model fit time series, usually with the purpose of forecasting a given variable. It has been used in time series forecast problems [27] and in hybrid models [55]. ARIMA models might not deal with large amounts of non-linearity but can provide good accuracy.

3 Problem Formulation

After studying the problem and specific context to this particular university student parking lot, we came to concluded that some of the aspects that most influence the affluence of students to the parking lot are class schedules, exams, weather and national and regional holidays. Some of these factors present temporal repetitions with different periods (for instance, classes are repeated every week, but with different patterns within each semester), while others can be considered non-periodic (for instance, movable holidays).

We collected all recorded accesses to the park between 2012 and 2017. For each access, there is a timestamp, type of access (entry or exit) and the student code. For each student we know the courses (s)he is enrolled in at each given year and semester, and for each course we have the dates of the exams and the schedules of different classes. We retrieved weather information from a meteorological station distanced 100 meters from the park. The weather data spans the same time frame, and has many attributes related to rain, wind, temperature or humidity, among others. We also collected information from the school calendars (includes national and regional holidays) for those years.

When first plotting the data from park accesses, we noticed a number of problems that resulted in the park capacity appearing to be exceeded (upwards of 1000 for a park of 350 slots) and occasions when the number of cars inside the park took a (rather large) *negative* number. We learned that the park gate sometimes loses connection and not all accesses are kept. We attempted to use some missing data strategies to deal with this problem, namely deleting entries that didn't have a corresponding exit and vice-versa, but this resulted in a very large loss of information. An attempt to fill in missing entries and exists was

cogitated, but eventually it would introduce too much noise in the data, given that the temporal component is of the utmost importance here.

Given the data inconsistencies and the fact that any pre-processing would solve some problems, but also introduce a lot of noise in the data, always shifting it from real-world behavior, we decided to generate artificial access data using all knowledge we had learned about the park and try to come up with a model complex enough that accurately portrays the time series problem we are trying to solve, while at the same time providing with a dataset that can be used for benchmarking. This approach of using artificial data is supported by other works such as [10] and [13].

For each student with access to the park we created a class schedule for any given (*year*, *semester*) pair, based on class enrollments. For each day of the pair, we give each student a probability p_{takes_car} of taking the car to the park. For each student who takes the car to the park, we simulate the arrival time by using a normal distribution with a standard deviation of v and an initial mean of u . The exits were simulated in a similar manner. Based on the weekday we give each student a different delay in the entries d_{entry} and exits d_{exit} .

Depending on the type of class (lectures, with no mandatory attendance, practical classes, with mandatory attendance, and exams) there is also a chance to skip class and thus u would shift to the start of the next class. For a selected number of weather conditions, we introduce changes to p_{takes_car} , u and v . For example, if it was raining 30 minutes before classes start, the probability of taking car is higher, but there is also a higher chance of arriving late. In case of a holiday we add a probability of skipping the whole day. Besides this, we also add different delays $d_{holiday_entry}$ and $d_{holiday_exit}$ to the entries and exits, respectively, according to what was observed in the access data from the park.

4 Experimental Setup

We selected five models to compare – Convolution Neural Network, Long Short-Term Memory Network, Random Forest Ensemble, Decision Tree and Gradient Boosting Model. These were chosen among all possibilities based on their performances in similar problems, as seen in section 2. The CNN and LSTM models were implemented with *Keras*, a Python deep learning library, while the other ones were implemented with *scikit-learn*, a Python machine learning library [41].

The data was prepared so that it contained information for every 10 minute interval. The classification was configured to target ten classes, each corresponding to 10% of the parking lot's total capacity.

The LSTM consists of an LSTM layer parameterized with a given number of neurons, an activation function and a recurrent activation function. The next layer is a Dense layer, where every input neuron is connected to all the output neurons from the previous layer, containing ten neurons, one for each output class. We use hard sigmoid as the recurrent activation function, the categorical cross-entropy as loss function and RMSProp as the optimizer strategy. We can

train the model for a given number of epochs and with the data packed in batches. The tested combinations for these parameters are shown in Table 1.

Table 1. LSTM Parameters and tested values

Parameter	Values
Training examples (%)	70, 80, 90
Batch sizes	32, 64, 128, 256
Epochs	4, 8, 12, 16
Hidden Neurons	50, 100, 150, 200
Activation	relu, softmax, tanh, sigmoid, hard sigmoid, elu, selu, softplus, softsign

The *training examples* represents the percentage of the data set used for training. The *batch size* is how many examples the model will be fed before updating its internal state. The *epochs* is how many passes we will go through the data fed for training. Usually a higher number of epochs increases the chances of over fitting. The *hidden neurons* represent how many neurons the LSTM layer will have inside. The *activation function* is a function that is called on the output of a layer before passing the value for the next layer.

The CNN consists of a convolutional 2D layer with a given *activation function*, *kernel size* and *convolutional filters*. This is followed by a Flatten layer and a Dropout layer with a dropout probability of 12.5%. Lastly a Dense layer serves as the output, with the number of neurons equal to the number of classes. We use the categorical cross-entropy as loss function and Adadelta as the optimizer. Table 2 presents the combinations of parameters tested.

Table 2. CNN Combinations

Parameter	Values
Training examples (%)	70, 80, 90
Batch size	64, 128, 256
Epochs	8, 12, 16
Convolution filters	16, 32
Kernel size	2, 3
Activation	relu, softmax

Convolution filters represents how many output filters we have in the convolution. The *kernel size* is the side of the squared 2D convolution window.

For the Decision Tree model, we experimented many configurations to try and find a good structure for the model. Table 3 presents the combinations of parameters tested.

Table 3. Decision Tree Combinations

Parameter	Values
Training examples (%)	10, 20, 30, 40, 50, 55, 60, 65, 70, 75, 80, 85, 90
Criterion	Gini Impurity, Information Gain
Splitter	random, best
Max Depth	None, 10, 50, 100, 500, 1000
Max Features	All Features, Sqrt(Features), Log ₂ (Features)

The *criterion* is the measurement of a split’s quality. The *splitter* parameter is the strategy when choosing between possible splits at a given state. *Max depth* is the maximum depth allowed for the tree. *None* means we expand the tree until all leaves are pure or contain less than 2 samples. Note that this number could be changed but we opted for not doing so. *Max features* represents the maximum number of features from the input to consider when searching for the best split. *None* means we set the max to the number of available features.

Table 4 shows the combinations of parameters that were tested with the Random Forest model. Estimators are the number of trees to include in our random forest ensemble.

Table 4. Random Forest Combinations

Parameter	Values
Training examples (%)	50, 60, 70, 80, 90
Estimators	10, 20, 50, 100, 200, 500, 1000, 2000, 5000
Criterion	Gini Impurity, Information Gain
Max Depth	None, 10, 20, 40, 80, 160
Max Features	All Features, Sqrt(Features), Log ₂ (Features)

Finally, with the Gradient Boosting model the tested combinations are presented in Table 5. The learning rate represents how strongly the model is changed at each iteration. It is related to the speed at which the model learns.

Table 5. Gradient Boosting Combinations

Parameter	Values
Training examples (%)	70, 80, 90
Loss	Deviance, Exponential
Learning Rate	0.01, 0.1
Estimators	100, 200, 400
Max Depth	5, 10
e Criterion	Mean Squared Error, Mean Absolute Error
Max Features	All Features, Sqrt(Features)

5 Results

The LSTM model performed worse than expected, especially when compared to the other models. From our understanding this might have happened because of a less than optimal choice of parameters. Nevertheless, the configuration with the best accuracy, 54.6 %, had a sigmoid activation, batch size of 128 with 12 epochs, 5 hidden neurons and a 90% percentage training.

The CNN model was less disappointing, with an accuracy of 70.9%. With a batch size of 64, a softmax activation, 16 convolution filters, 16 epochs, a kernel size of 3 and trained on 80% of the data set.

Achieving better results, the DT model managed an accuracy of 86.4%. It was set to use the maximum available features up to a depth of 1000, trained on 85% of the dataset, using the best split at each state and evaluating how good the split was through the Information Gain strategy.

Even better than the previous one, the RF model managed an impressive 89.4% accuracy. Trained on 90% of the dataset, using the Gini Impurity measure, a max depth of 100 and using all the features when deciding the split. This model was an ensemble of 100 estimators.

The gradient boosting modeled achieved also good results with 86.8% accuracy, trained on 90% of the dataset, a deviance loss function, 400 estimators, and a max depth of 10. It also used every available feature, a mean squared error criterion and a learning rate set to 0.01.

Table 6 summarizes the obtained results for all five methods, and using several known metrics. The *Off by 1 average accuracy* evaluates the accuracy considering that prediction might be off by one class (ie. the park is predicted to be at 60% but is actually at 70%). We can observe that again RF presents the best results, followed closely by DT and GB. Evaluating each of the corresponding confusion matrices (not shown here) we can see that the LSTM classified every input with the same class and the CNN classified only with the two extreme classes, this is, the park being completely full and completely empty. The *Time for 1 Prediction* metric averages the time for 100 predictions, thus diluting possible model loading time.

Our insight to justify these results is that the configurations of the LSTM and CNN were not the most suitable. Another problem to address is the fact that the dataset is unbalanced – it has 20448 elements and considering the 10 output classes used, where the first class contains 54.36% of the data and the tenth 20.59%, with the remaining 25.05% divided among the other 8 classes.

6 Conclusions and Future Work

This work started with the hope of finding models capable of producing a synergy in time series prediction. Our results produced mixed feelings about which models to choose for hybridization but we must take into consideration the problems mentioned in 5. For that, we studied different methods that have been applied to this context over time, and selected five methods to test. As a case study, we

Table 6. Results Table

Metric	LSTM	CNN	DT	RF	GB
Average Accuracy	0.55	0.71	0.86	0.89	0.87
Off by 1 Avg. Accuracy	0.55	0.77	0.94	0.96	0.94
Precision	0.30	0.60	0.86	0.89	0.86
Recall	0.55	0.71	0.86	0.89	0.87
F1-Score	0.39	0.62	0.86	0.89	0.86
Log Loss	15.67	10.06	4.68	3.66	4.56
Hamming Loss	0.45	0.29	0.14	0.11	0.13
Time to Train (s)	21.43	318.22	0.65	10.83	2401.73
Time for 1 Prediction (μ s)	464.79	368.38	3.04	1023.71	352.08
Model Space (MiB)	0.02	1.78	0.54	44.82	246.50

use a semi-artificial dataset of a university parking lot, together with classes and weather information. The five chosen methods were tested with several combinations of parameters, in order to determine the best configuration for each. From the tested configurations, RF has proven to be the best method, followed closely by DT and GB.

Obviously, this study can be extended with the use of additional methods and the testing of additional parameter setting for the used methods. Also, we intend to try these methods with different temporal granularity and also with more output classes (for instance, 20 classes representing intervals of 5% of the parking lot capacity).

This work was the first step of a larger and more ambitious endeavour that aims at implementing hybrid models that can result in better results than these individual models. Hybrid approaches are typically capable of building on the advantages of the single models they are based on, at the same time mitigating their complementary disadvantages. Several approaches at hybrid models have already been practised, such as in photo-voltaic panels [28] and known data sets like in [55], [53] with the Canadian Lynx and Dollar versus Pound data set. Focusing specifically on time series, there is also some recent work on the forecasting of financial markets, as in [33]. Our approach will focus on three types of models:

- **Superficial** hybrids, stemming from works such as [20], [26] and [56]. These combine models with complementary prediction strengths, achieving an ensemble of methods capable of higher global performance.
- **Middle-Deep/Limb-Deep** hybrids, where parts of two different models are exchanged, in an approach similar to transfer learning.
- **Chimera/Full-Deep** hybrids, where two models are deeply integrated in one another. A similar example to this is shown in [14].

References

1. Abdolmaleki, P., Yarmohammadi, M., Gity, M.: Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings. *Intl. Journal of Radiation Research* **1**(4), 217–228 (2004)
2. Abhishek, K., Singh, M., Ghosh, S., Anand, A.: Weather Forecasting Model using Artificial Neural Network. *Procedia Technology* **4**, 311–318 (jan 2012)
3. Awad, M., Pomares, H., Rojas, I., Salameh, O., Hamdon, M.: Prediction of Time Series Using RBF Neural Networks: A New Approach of Clustering. *The International Arab Journal of Information Technology* **6**(2), 138–143 (2009)
4. Borovykh, A., Bohte, S., Oosterlee, C.W.: Conditional Time Series Forecasting with Convolutional Neural Networks. arXiv:1703.04691 (2017)
5. Bruss, F.T.: 250 years of An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S.. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **115**(3-4), 129–133 (2014). <https://doi.org/10.1365/s13291-013-0069-z>
6. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent Neural Networks for Multivariate Time Series with Missing Values. arXiv:1606.01865v1 (2016)
7. Cortes, C., Vapnik, V.: Support-Vector Networks. *Machine Learning* **20**(3), 273–297 (1995). <https://doi.org/10.1023/A:1022627411411>
8. Cox, D.R.: The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society, Series B (Methodological)* **20**(2), 215–242 (1958)
9. Dai, J., Venayagamoorthy, G.K., Harley, R.G.: An Introduction to the Echo State Network and its Applications in Power System. In: *Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems (ISAP '09)* (2009). <https://doi.org/10.1109/ISAP.2009.5352913>
10. Dwiputrantanto, T.H., Setiawan, N.A., Aji, T.B.: Machinery equipment early fault detection using Artificial Neural Network based Autoencoder. In: *Proceedings of the 3rd International Conference on Science and Technology - Computer (ICST 2017)*. pp. 66–69 (2017). <https://doi.org/10.1109/ICSTC.2017.8011854>
11. Fu, R., Zhang, Z., Li, L.: Using LSTM and GRU neural network methods for traffic flow prediction. In: *Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, YAC 2016* (2017). <https://doi.org/10.1109/YAC.2016.7804912>
12. Fukushima, K.: Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks* (1988). [https://doi.org/10.1016/0893-6080\(88\)90014-7](https://doi.org/10.1016/0893-6080(88)90014-7)
13. Gallicchio, C., Micheli, A., Pedrelli, L.: Deep reservoir computing: A critical experimental analysis. *Neurocomputing* **268**(April), 87–99 (2017). <https://doi.org/10.1016/j.neucom.2016.12.089>
14. Gil-Begue, S., Bielza, C., Larraaga, P.: Multi-dimensional Bayesian network classifier trees. Submitted. In: *The 9th International Conference on Probabilistic Graphical Models, September 11–14, 2018, Prague* (2018)
15. Goel, H., Melnyk, I., Banerjee, A.: R2N2: Residual Recurrent Neural Networks for Multivariate Time Series Forecasting. arXiv Computing Research Repository (2017). <https://doi.org/arXiv:1709.03159>
16. Greene, W.H.: *Econometric analysis*. Prentice Hall (2002)
17. Gullapalli, V.: A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks* (1990). [https://doi.org/10.1016/0893-6080\(90\)90056-Q](https://doi.org/10.1016/0893-6080(90)90056-Q)

18. Hand, D.J., Yu, K.: Idiot's Bayes—Not So Stupid After All? *International Statistical Review* **69**(3), 385–398 (2001). <https://doi.org/10.1111/j.1751-5823.2001.tb00465.x>
19. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning – Data Mining, Inference and Prediction*. Springer (2009)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016)
21. Hills, J., Lines, J., Baranauskas, E., Mapp, J., Bagnall, A.: Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery* **28**(4), 851–881 (2014). <https://doi.org/10.1007/s10618-013-0322-1>
22. Horton, P., Nakai, K.: Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier. In: *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*. pp. 147–152 (1997)
23. Jaeger, H., Haas, H.: Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science* **304**(5667), 78–80 (2004)
24. Johnson, N.E., Ianiuk, O., Cazap, D., Liu, L., Starobin, D., Dobler, G., Ghandehari, M.: Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. *Waste Management* **62**, 3–11 (2017)
25. Karim, F., Majumdar, S., Darabi, H., Chen, S.: LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access* **6**, 1662–1669 (2017)
26. Khashei, M., Bijari, M.: A new class of hybrid models for time series forecasting. *Expert Systems with Applications* **39**(4), 4344–4357 (2012)
27. Kumar, S.V., Vanajakshi, L.: Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review* **7**(3) (2015)
28. Laudani, A., Lozito, G.M., Riganti Fulginei, F., Salvini, A.: Hybrid neural network approach based tool for the modelling of photovoltaic panels (2015). <https://doi.org/10.1155/2015/413654>
29. LeCun, Y., Bengio, Y.: *The handbook of brain theory and neural networks*, chap. Convolutional networks for images, speech, and time series, pp. 255–258. MIT Press (1995)
30. Li, D., Han, M., Wang, J.: Chaotic Time Series Prediction Based on a Novel Robust Echo State Network. *Transactions on Neural Networks and Learning Systems* **23**(5) (2012). <https://doi.org/10.1109/TNNLS.2012.2188414>
31. Li, P., Gou, J., Yang, H.: The Distance-Weighted K -nearest Centroid Neighbor Classification. *Journal of Intelligent Information Hiding and Multimedia Signal Processing (JIH-MSP)* **8**(3), 611–622 (2017)
32. Lipton, Z.C., Berkowitz, J., Elkan, C.: A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv Computing Research Repository* (2015). <https://doi.org/arXiv:1506.00019>
33. Marwin Züfle, A.B., Herbs, N., Curtef, V., Kounev, S.: Telescope: A Hybrid Forecast Method for Univariate Time Series. In: *Proceedings of the International work-conference on Time Series (ITISE 2017)* (July 2017)
34. Odom, M., Sharda, R.: A neural network model for bankruptcy prediction. In: *Proceedings of the 1990 International Joint Conference on Neural Networks (IJCNN)*. pp. 163–168 (vol. 2) (1990). <https://doi.org/10.1109/IJCNN.1990.137710>
35. Orr, M.J.L.: *Introduction to radial basis function networks*. University of Edinburg (1996)
36. Pai, P.F., Lin, C.S.: A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* **33**(6), 497–505 (2005)

37. Pal, M.: Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* **26**(1), 217–222 (2005)
38. Prevost, J.J., Nagothu, K., Kelley, B., Jamshidi, M.: Prediction of cloud data center networks loads using stochastic and neural models. In: *Proceedings of the 2011 6th International Conference on System of Systems Engineering*. pp. 276–281 (2011)
39. Russell, S.J., Norvig, P.: *Artificial Intelligence: A Modern Approach*. Prentice Hall (1995). [https://doi.org/10.1016/0925-2312\(95\)90020-9](https://doi.org/10.1016/0925-2312(95)90020-9)
40. Saad, E.W., Prokhorov, D.V., Wunsch, D.C.: Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks. *IEEE Transactions on Neural Networks* **9**(6), 1456–1470 (1998)
41. Scikit-Learn: *Scikit-Learn User Guide*, 0.19.1 edn. (2017)
42. Specht, D.F.: Probabilistic neural networks. *Neural networks* **3**(1), 109–118 (1990)
43. Tsai, C.F., Wang, S.P.: Stock price forecasting by hybrid machine learning techniques. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2009)*, March 18–20, 2009, Hong Kong. vol. 1 (2009)
44. Wang, J., Wang, J.: Forecasting stochastic neural network based on financial empirical mode decomposition. *Neural Networks* **90**, 8–20 (2017)
45. Williams, B.M., Hoel, L.A.: Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering* **129**(6), 664–672 (2003)
46. Winston, P.: A heuristic program that constructs decision trees. *Artificial Intelligence Memo 173*, MIT (1969)
47. Wu, C.H., Wei, C.C., Su, D.C., Chang, M.H., Ho, J.M.: Travel time prediction with support vector regression. In: *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*. vol. 2, pp. 1438–1442 (2003)
48. Wu, H., Cai, Y., Wu, Y., Zhong, R., Li, Q., Zheng, J., Lin, D., Li, Y.: Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression. *Bioscience trends* **11**(3), 292–296 (2017)
49. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y.X., Chang, Y.F., Xiang, Q.L.: A leaf recognition algorithm for plant classification using probabilistic neural network. In: *Proceedings of the 2007 IEEE International Symposium on Signal Processing and Information Technology*. pp. 11–16. IEEE (2007)
50. Xu, M., Han, M.: Adaptive elastic echo state network for multivariate time series prediction. *IEEE transactions on cybernetics* **46**(10), 2173–2183 (2016)
51. Xue, J.H., Titterton, D.M.: Comment on on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Neural processing letters* **28**(3), 169 (2008)
52. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *European conference on computer vision*. pp. 818–833. Springer (2014)
53. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **50**, 159–175 (2003)
54. Zhang, Y., Haghani, A.: A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* **58**, 308–324 (2015)
55. Zheng, F., Zhong, S.: Time series forecasting using a hybrid rbf neural network and ar model based on binomial smoothing. *World Academy of Science, Engineering and Technology* **75**, 1471–1475 (2011)
56. Zhou, Z., Feng, J.: Deep forest: Towards an alternative to deep neural networks. *CoRR abs/1702.08835* (2017), <http://arxiv.org/abs/1702.08835>

A Forecasting Methodology based on growth models, for assessing performance: Application on the Moroccan Railway.

Karima SELMANI BOUAYOUNE

Mohammadia Engineering School, University Mohamed V Agdal, Rabat, Morocco

karimaselmani@yahoo.fr

Abstract. Forecasting methods is highly developed in machine learning and all the areas of predictive modelling. They are useful for a wide range of phenomena using simulation. The most popular of these approaches are the structural models and the time series forecasting. These methods are used to correct errors and they are recognized by their reliability and efficiency in prevision. However, in our study, a new forecasting approach is tried for predicting. It is based on growth models, which mostly used in testing software. Our approach aims to make prediction of accidents in the Moroccan Railway system, using two growth models and to determine the best performance for them. This is will be executed by the compute of the estimator of each model and to conclude the model that gives the best simulation, so the model with the best performance. Results of the application on Moroccan Railway show that the best model estimator vary during the years of the study.

Keywords: Forecasting, Growth Models, Performance, Predictive modeling

1 Introduction

Forecasting of accidents has become necessary for safety in the transport sector. Many frequent and continuous studies are required to ensure the human or, mainly, the travelers security.

Hence, based on these predictions, the decision on the traffic management will be made as well as the actions to ensure greater safety for travelers. The decision is taken from the interpretation of modeling results. This kind of models is part of the predictive modeling (which is usually used in machine learning).

By browsing different revues about forecasting [1, 2] [5], we can see that the methods of forecasting are diverse, but the time series is the most popular and increasing in the forecasting approaches. [3,4][6-8].

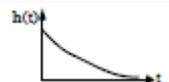
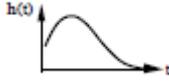
In this article, we choose to study the accidents in the transport sector and we the application will be on Moroccan railway system. And for the predictive modeling we suggest a new methodology for predicting. It is based on growth models. In fact, these model categories are usually used in testing software, but we will use them because the prevention, in our study, is founded on a small sample.

2. Forecasting using Growth models

Forecasting is used for predicting. Usually, in the method like ARMA, AFIRMA, VAR,..it is defined variables and number of observations. The simulation is given from the experiment parameters. The prediction is estimated by the results that the model formulation gives. The models chosen for forecasting in our study are based on growth models, that are used in Reliability and trend analysis.

The trend analysis allows predicting the characteristics measures of software reliability, over the time [9-12]. There are many software reliability growth models, the most popular are presented in the Table 1.

Table 1, Examples of Models of reliability growth [13]

Model Type	Graphic
Hyper exponential (HE) $h(t) = \frac{\alpha\phi_{sup}e^{-\phi_{sup}t} + \bar{\alpha}\phi_{inf}e^{-\phi_{inf}t}}{\alpha\phi_{sup}e^{-\phi_{sup}t} + \bar{\alpha}\phi_{inf}e^{-\phi_{inf}t}}$	
Exponential (Exp) $h(t) = N\phi exp(-\phi t)$	
S Type (Gompertz distribution) $h(t) = N\phi^2 t exp(-\phi t)$	

For searching the performance of the estimation, the best estimator is the model that gives approximations, which tends the most towards the real values.

3. Moroccan Railway Statistics

In [14] some statistics are collected, as a part of risk management activity, from the year 2000 to the year 2008, for the Moroccan railway. We choose those statistics obtained on accidents (Table 2) and rate failure for three components: rail, human and vehicle (Table 3). We obtain, by this way, a sample of 9 periods (years).

Table2, Statistics collected for accidents in Moroccan Railway

Number of years	Cumulative number of accidents
1	11
2	17
3	35
4	48
5	63
6	84
7	96
8	103
9	118

By presenting graphically values in this table, we obtain the following figure.

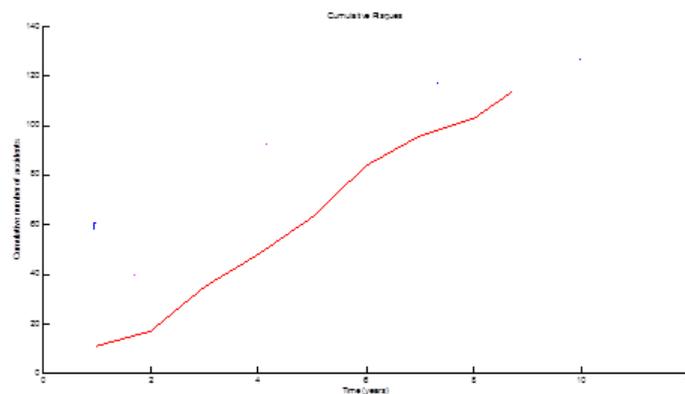


Fig.1 Cumulative number of accidents in the Moroccan Railway, during the period (2000-2008)

4. Estimation by predictive modeling

We suggest, then, to calculate parameters of these models and to use them for verifying if our methodology leads to conclude good prediction or not.

For our example, we choose only the two models Exponential and S-Type. The model Hyper exponential contains many parameters and it is exhausting to work with it, for just the validation of results and interpretations.

4.1 The Calculation of the model parameters

Thus, to calculate the parameters N and \emptyset , in the table 1, we use the maximum likelihood method. So, for an observation interval $[0, t_n]$, that is divided into a set of intervals $[0, t_1]$, $[0, t_2]$, ..., $[0, t_m]$ and the number of failures per subinterval is recorded as n_k ($k = 1, \dots, m$), respectively to the number of failures in $[t_{k-1}, t_k]$.

We note $H(t) = \int_0^t h(s) ds$, so, we have :

– For the Exponential model :

$$H_1(t) = N(1 - \exp(-\emptyset t)) \quad (1)$$

– And for the S Type model:

$$H_2(t) = N(1 - (1 + \emptyset t)\exp(-\emptyset t)) \quad (2)$$

The likelihood function is :

$$L(n_1, n_2, \dots, n_m) = \prod_{k=1}^m \frac{\{H(t_k) - H(t_{k-1})\}^{n_k}}{n_k!} \exp\{[H(t_k) - H(t_{k-1})]\} \quad (3)$$

By taking the natural logarithm of both sides, we have,

$$\begin{cases} \frac{\partial \ln L}{\partial N} = 0 \\ \frac{\partial \ln L}{\partial \emptyset} = 0 \end{cases} \quad (4)$$

$$\ln L(n_1, n_2, \dots, n_m) = \ln \prod_{k=1}^m \frac{\{H(t_k) - H(t_{k-1})\}^{n_k}}{n_k!} \exp\{[H(t_k) - H(t_{k-1})]\} =$$

$$\sum_{k=1}^m \ln \frac{\{H(t_k) - H(t_{k-1})\}^{n_k}}{n_k!} \exp\{[H(t_k) - H(t_{k-1})]\} = \sum_{k=1}^m \{n_k \ln[H(t_k) -$$

$$H(t_{k-1})] - [H(t_k) - H(t_{k-1})] - \ln n_k!\}$$
(5)

And the Likelihood equations are :

4.2 Numerical results of parameters

If we consider the Exponential model, from this equation we can calculate the value of N , by solving the equations below,

$$\begin{cases} N = \frac{\sum_{k=1}^m n_k}{1 - e^{-\phi t_m}} \\ \sum_{k=1}^m \left(\frac{n_k}{e^{-\phi t_{k-1}} - e^{-\phi t_k}} - N \right) (t_{k-1} e^{-\phi t_{k-1}} - t_{k-1} e^{-\phi t_{k-1}}) \end{cases}$$
(6)

Consequently, N can be calculated directly, however, it is very difficult to calculate ϕ , so, it can be obtained, numerically.

The calculation using data in the Table 1 leads to obtain that $N = 118,11$ and $\phi = 0,7$. Thus, calculating H_1 and H_2 becomes easy and it is represented in table below,

Table 2, Values of H_1 and H_2 for the time between 1 and 9 year

Number of years	Cumulative number of accidents	H_1 (Exponential model)	H_2 (S Type model)
1	11	59,458	18,402
2	17	88,984	48,209
3	35	103,647	73,274
4	48	110,928	90,817
5	63	114,543	102,060
6	84	116,339	108,900
7	96	117,230	112,921
8	103	117,673	115,227
9	118	117,893	116,527

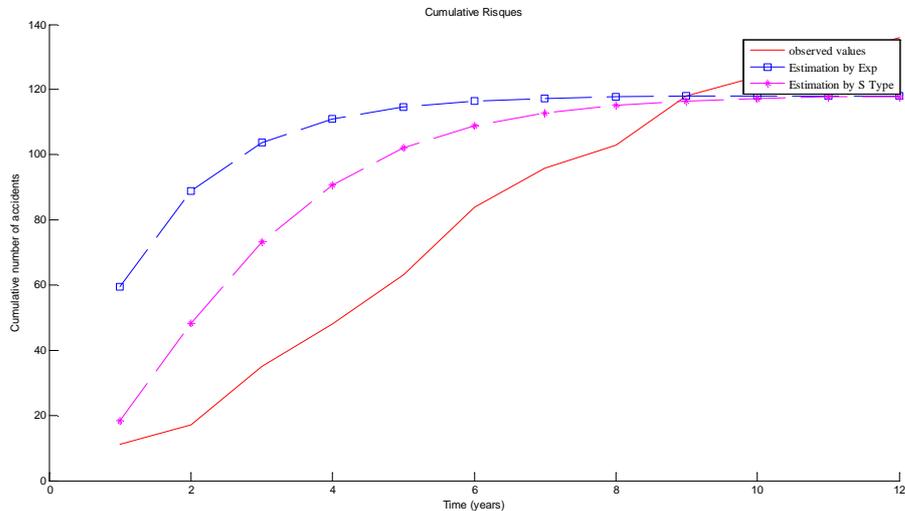


Fig.2 , Test for detecting the best model estimator, by calculating cumulative number of accidents. The graphic shows that the model S Type approximates, mostly, more best than the exponential the real values.

As we can notice from the table 2 , that the model S Type approximates the values of cumulative number of accidents observed, more best than the model of exponential, except from the eight year. This is can be simulated in the graphic, in the figure 2.

In this figure, we can notice that the model S Type is the best model for the estimation, specially for the first years. This is confirms that, the model exponential becomes the most best for estimation, which is concluded that our assessment indicate the performance in the model S Type.

5. Conclusion

This article provides a new methodology for forecasting , by using Reliability Growth Models, which are mostly applied on software reliability.

However, we have tried to apply these models on a railway system. It should to make experiences to validate the best estimation results that we will obtain. And conclude the performance in this system, for this estimation.

The advantage of this methodology that it provides results in order to assure if the system is safe and secure in the considered period or no. However, if we need to know what are the necessary actions that we should make to maintain the required level of reliability, we must monitor rate failures at this case.

References

1. Fildes, R.; Nikolopoulos, K.; Crone, S.F.; Syntetos, A.A. Forecasting and operational research: A review. *J. Oper. Res. Soc.* 2008, 59, 1150–1172.
2. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast.* 2014, 30, 1030–1081.
3. Bontempi, G.; Taieb, S.B.; Le Borgne, Y.A. Machine learning strategies for time series forecasting. In *Business Intelligence (Lecture Notes in Business Information Processing)*; Aufaure, M.A., Zimányi, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 138, pp. 62–77.
4. De Gooijer, J.G.; Hyndman, R.J. 25 years of time series forecasting. *Int. J. Forecast.* 2006, 22, 443–473.
5. Hong, T.; Fan, S. Probabilistic electric load forecasting: A tutorial review. *Int. J. Forecast.* 2016, 32, 914–938.
6. Mei-Ying, Y.; Xiao-Dong, W. Chaotic time series prediction using least squares support vector machines. *Chin. Phys.* 2004, 13, 454–458.
7. Faraway, J.; Chatfield, C. Time series forecasting with neural networks: A comparative study using the air line data. *J. R. Stat. Soc. C Appl. Stat.* 1998, 47, 231–250.
8. Zou, H.; Yang, Y. Combining time series models for forecasting. *Int. J. Forecast.* 2004, 20, 69–84.
9. M. Feizabadi and A. Eshraghniaie Jahromi, *Reliability Engineering & System Safety*, 157, 101–112 (2017).
10. R. Billinton, and R. N. Allon, *Reliability Evaluation of Power Systems* (Plenum Press, New York, 1986).
11. D. J. Smith, *Reliability, Maintainability and Risk* (Eighth Edition, 2010).
12. M. Xie, and C.D. Lai, *Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function*, *Reliability Engineering & System Safety*, 52,1, 87-93, (1996).
13. K. Kanoun, M. Kaâniche and J. Laprie, *Fiabilité du logiciel : de la collecte des données à l'évaluation probabiliste*, *RAIRO, Technique et Science Informatiques*, Hermès, 16, 7, 865-895 (1997)
14. J. BOUDNAYA and A. MKHIDA, *Comparative study of the unreliability of a Moroccan Level Crossing using stochastic Petri Nets approach and fault tree analysis*, 11ème CONGRES INTERNATIONAL DE GENIE INDUSTRIEL, (2015).

Pereira Market Scan

Leandro Pereira¹; Carlos Jerónimo¹; José Santos² ∞

¹ Business Research Unit – BRU-IUL, ISCTE, Portugal

² Winning LAB, Winning Scientific Management, Portugal

∞ jose.santos@winning.pt

Abstract— The increasing competitiveness of market search related with information needs led to several companies to invest in new forms of decision making and models applied to scientific methods to position them as market leaders and to extract added value. This study was conducted with the propose of developing a market tool to allow a deeper and more pragmatic analysis to clients value and clients life cycle in order to contribute to the competitiveness of companies with relevant outputs and increase the market share of his company in a more conscious, objective way. The findings of this study allowed to quickly identify and to mitigate the type of problems which affect the organization and thus can provide benefits to all the stakeholders that are in interaction with organizations.

Keywords — Marketing, Strategy, Management, Decision Making

1 Introduction

The increasing competitiveness of the domestic market and the existence of scarce resources related with the need for organizations on having business information on time and ongoing to support management decision-making has increased the demand for process-based solutions/models and scientific methods [1]. However, nowadays there are no systematized, reliable and permanent techniques

for the predictive control of market shares despite the literature review and techniques mentioned on the new chapter, that is, systems for collecting critical and anticipated information of the business variation in a continuous way. This would empower managers of the organization to have evidence on the market and thus enable them to make management decisions in order to create competitive advantages. This competitive advantage has been conducted through the testing of processes and/or techniques in different contexts and organizations where the existence of a cause-effect relation allowing the conclusion of a universal management existence with rigor and validity.

Taking into account the principle of scientific management [2] and the growing demand from national organizations to have timely and relevant business information for the control and management of their market shares (through the decision on strategies and other types of decisions), the “Pereira Market Scan” has been implemented to provide companies with greater knowledge of their business in real time and on a continuous basis, so that managers can make decisions that allow them to achieve several important objectives such as:

I - Increase **Client Value**, i.e., boost business value by increasing current customer billing (up and cross

selling) and capturing new customers (in segment, geography, product or win-back);

II - Increase **Client Life Cycle**, i.e., the time current customers remain when they choose to continue buying instead of changing to a competitor, thus avoiding its exit and reducing the current customer abandonment rate.

In this sense, the "Pereira Market Scan" model is analytical in providing information on variables that influence market share and dynamic by giving directions for the search for improvement actions or good practices. At the same time, the model allows the estimation of the market share according to possible changes on the critical variables.

2 Technical and Scientific Objectives

The technical and scientifically objectives of the current model are developed to enhance and to reply to several manager's needs.

The main needs and gains from "Pereira Market Scan" are to provide a method of preventive control of market share variation, easy to apply, in order to avoid future losses, thus guaranteeing stability and financial security to the organization with positive consequences for the final consumer; Justify the positioning of the market share in a predictive way by estimating quantitatively which variables have a direct impact on the same market share and to provide the competitive market with tools based on the principle of scientific management that enables greater international competitiveness to the Portuguese business environment. The market share indicator, mentioned before, is the slice/percentage of an industry or the total market sales of a company over a given period of time. This metric is used to give a general idea of the proportional dimension of the company in the market in comparison with its competitors.

3 Literature Review and State of the Art

To develop "Pereira Market Scan" it has been conducted an extensive literature review on the state of art in order to provide a more efficiency tool and not overcome and offer the same solutions to the current existing tools. From the available literature describing the available tools/models, some of them could be highlighted such as:

-Boston Consulting Matrix [3] it's a model based on product lifecycle, created by Bruce Henderson in the early 1970s, to determine what priorities should be given in a product portfolio of a business unit. The model implies that to ensure long-term value creation, a company must have a product portfolio that contains both high-growth products requiring cash inputs as well as low-growth products that generate a significant amount of cash. It has 2 dimensions: market share (the proxy for competitive advantage) VS market growth (serves as a proxy for industry attractiveness). This matrix maps the position of the business units in these two dimensions that generate profit. The higher the market share of the product or the faster the product market growth, the better it is for the company. This tool allows to verify the business portfolio of a company and can serve as a starting point for a discussion of resource allocation in strategic business units.

Limitations: high market share is not the only success factor; market growth is not the only factor/indicator of market attractiveness; static model, not providing the market information in a continuous and detailed way for making management decisions. The measurement is made on a qualitative scale that is defined in the perspective of those who analyze and elaborate the matrix, opening space for subjectivity. In comparison with the "Pereira Market Scan" model, it does not allow the constant and predictive

measurement of market shares and it is based on a qualitative market analysis.

-Internal-External (IE) Matrix is used to analyze the strategic position of one or more businesses [4]. It is based upon the IFE Matrix and the EFE Matrix. Through the scores assigned in these two matrices it is possible to draw the final matrix. The axis of the XX's with the score attributed in the matrix IFE (Internal Factors) and the YY's axis with the score obtained in the matrix EFE (External Factors). The result of the total score of each of the factors, external and internal, should be inserted in a matrix. This matrix is divided into three major areas (Grow and Build, Hold and Maintain and Harvest or Divest) each one of these areas have a strategy depending on the location of the endpoint. The Grow and build means the company strategy should focus on market penetration as well as market and product development. The Hold and Maintain the company strategy should focus on market penetration and market development. Finally, the Harvest or Divest the company strategy should focus on reducing costs in these areas or opting for the elimination of this area/business.

Limitations: It only reflects the current position of the company/business, not taking into account a future image of it. The measurement is made on a qualitative scale that is defined in the perspective of those who analyze and elaborate the matrix, opening space for subjectivity. In comparison with the "Pereira Market Scan" model, it does not allow the constant and predictive measurement of market shares in an analytical way.

-GE/Mckinsey Matrix [5] is an analysis model for portfolio and its business units. On the one hand, the ideal business portfolio helps to explore the most attractive industries and markets and, on the other hand, is it embedded in the company's main

strengths. The goal regarding portfolio management is to define in which businesses to invest in, where to develop growth strategies and in which businesses to disinvest. This matrix is based on two axes, competitive strength and market attractiveness. The Competitive Strength analyzes the internal factors of the business units (e.g. strength of assets and competences, quality, patents, access to financial resources and investments, cost versus competition, market share, growth of market share, among others). The Market Attractiveness measures the factors external to the business (e.g. market size, market growth rate, market profitability, entry barriers, competition, demand, market segmentation, among others).

Limitations: It does not consider the links/interactions between the different business areas. It does not consider core competencies that lead to value creation. The measurement is made on a qualitative scale that is defined in the perspective of those who analyze and elaborate the matrix, opening space for subjectivity. In comparison with the "Pereira Market Scan" model, it does not allow the constant and predictive measurement of market shares in an analytical way.

-Space Matrix Strategic Management Method [6] is a management tool used in the formulation of business strategies considering the position of the company in the market against the competition. The acronym SPACE refers to Strategic Position and Action Evaluation. This matrix suggests, depending on the results obtained, the nature of the strategy to be used from among 4 options (Aggressive, Conservative, Defensive and Competitive). To obtain these results it is needed to analyze first 4 strategic dimensions, 2 internal and 2 external, which are: Internal (Financial Strength and Competitive Advantage) and External (Environmental Stability and Industry Strength).

Limitations: As in the other models, it requires an assessment by the user for each of the variables, which is not always the most accurate. A predefined universal ranking could mitigate this problem. Static model. It represents, just like in other models, just an image of the present with the actual conditions.

In comparison with the "Pereira Market Scan" model, it does not allow the constant and predictive measurement of market shares in an analytical way.

In addition to these main models there are others of added value from the marketing branch, namely the Attraction Models (Kotler's Fundamental Theorem) [7] with a parallel study from Bell et al [8], the Market Share Theorem [9] and Choice Modelling, which deserve to be mentioned.

4 Methodology

The "Pereira Market Scan" tool sought to respond to the scientific/technological uncertainty of being able to implement a market share estimation, analysis and management model that, besides supporting the necessary strategic definition, allows for a permanent determination and estimation of the market share through a set of metrics that represent Client Value and Client Life Cycle.

The aim of the tool was to solve the problems related to the fact that the models that exist are static models, which do not anticipate the need for project priority [10] changes and decisions on new investments. These models are limited because they are based on the classification of factors based on a subjective process (e.g. classification of market attractiveness, classification of competitive strength), although based on the use of scales (albeit qualitative). In addition, they do not address the depth of critical, measurable, and monitorable variables that are the source of power to generate and increase a company's market share.

The main uncertainty this tool sought to solve was to be able to project the market share expected due to a set of business variables associated with a high degree of confidence.

In order to carry out a permanent "Scan" Model to respond to the current problem in organizations, the project activities and methodology were based on the literature review (chapter 3), analysis of existing models in the market, survey of limitations of the existing solutions, diagnosis of the current need and necessary requirements, model development, sample testing and validation of the model.

4.1 Model Development

The model presents a set of explicit variables in the vertices of a diamond that represent:

- New clients, which focuses on the rate of new customers in a certain period of time (usually monthly) with the possibility of analyzing the composition of the customer by geography, segment, product and also win-back customers. For this it is necessary to identify what a customer is (quantitatively framed about what is the minimum value to be considered one) and what causes or sources exist to be able to attract new customers. The calculation is obtained through the behavior of the variable over time, that is, the past is compared to the present;
- Up-selling, frequently assumed as a sale of more product quantity or product upgrade to the current customer or even as the increase of the profit margin by increasing the product price, expressed in the average customer value (average items by basket). To understand its measurement it is essential to be aware of the products that are more requested by the customers and the reasons why a customer leaves without buying from the store. In this way, it is

possible to understand and determine the leverage of sales in up-selling;

- Cross-selling, understood as complementary sales or impulse sales to the same customers or even sales within the sphere of influence in B2B or B2C. It is also reflected in the average customer value (average of different items in the basket). To better perceive its measurement it is important to understand for the products more requested by the customers which complementary products are added to the purchase. In this way, it is possible to understand and determine the leverage of cross-selling sales;

-Retention, is based on the customer abandonment rate, that is, it focuses on clients who leave. To determine it, it is crucial to know the average annual purchase frequency per customer. In order to act on it, understanding the causes that lead clients to leave the organization determines the variation of the abandonment rate, and the variation is calculated in comparison with the previous measurement. At the same time, the action can focus on the current customer's life cycle (therefore it is crucial to determine which is the average duration of this cycle). Once again, understanding the reasons that allow to extend the relationship with the customer enables the organization to act on the current customer's life cycle.

4.2 Model Validation

The typical process of market share analysis is in itself a process with a low degree of confidence, as the tools normally used are poorly measurable and monitorable given the limitations listed above. At the same time, its definition is not trivial given the ambiguity of the market concept. In the present case, this tendency was countered by exhaustively delimiting the variables addressed. Thus, the market share is given by a fraction of the actual gross sales (in monetary value) of the company in relation to the

competitors, for a monthly period, in a given set of products and a given geographic area. The total market consists of the company and the competitors with specific characteristics depending on the industry.

During the reference period, a set of tests and analysis were performed on the results obtained at each stage of the tool development. In this sense, a sample of 20 companies from the retail industry were used and, for the present study, the data used was encrypted in order to protect the source that provided it. At the same time, it was ensured that all parties did not become aware of the study as well as who did the double-blind comparative analysis. During the same period, the KPIs associated with the four variables of the "Pereira Market Scan" model were measured, as well as the evolution of the market share. Each company was asked to record, in a previously distributed collection matrix, the management measures they were taking over the reference period to increase market share. In this way, it would be possible to understand the relationship of these measures with the variables of the model.

In the scope of the present study, and in order to validate the "Pereira Market Scan" model, it was analyzed the relationship between the variables that the model exposes and the market share, that is, the attraction of new customers, the retention of current customers, up-selling and cross-selling sales, and market share.

For the previous validation, two methodologies with a high degree of confidence were used: the Pearson correlation coefficient and the linear regression (through the coefficient of determination advanced by r^2), thus establishing the aforementioned principle of scientific management. Pearson's correlation coefficient indicates the direction of correlation [12

1], if any, between each variable and the market share, whether positive or negative, of metric scale (ratio or interval). The linear regression through r^2 allows to obtain the estimated value (conditional) of a variable in relation to data of another variable that one wants to test, that is, how much of the variation of a variable (in this case, market share) is matched by variation of the other variable (each one presented in the "Pereira Market Scan" model) [12].

It should be noted that each variable is measurable by specific indicators. The new clients variable is obtained through the attraction rate of new customers and presents a positive correlation (more detailed in the presentation of results) with the market share. The Retention variable is quantified by the abandonment rate. In this case the correlation is negative, i.e. if our customers abandon us our market share is negatively affected. The Up-selling and Cross-selling variables are measured by the up-selling and cross-selling sales rates, respectively, that are positively correlated with the market share.

Finally, it was intended to obtain an equation that would explain the relation of the four variables of the model with the market share with the intention of being able to establish a predictive model on the market share. For this verification, a multiple linear regression, shown below, was applied and represented by the equation

$$Y = M_1.X_1 + M_2.X_2 + M_3.X_3 + M_4.X_4 + C.$$

The uniqueness of the model developed is not only justified by the knowledge/skills in the technical fields of Business Case and Project Management, but also in the experience and knowledge acquired over time.

5 Results

The application of methodologies that incorporate scientificity, confidence and rigor to the validation of the tool leads to believe that all the foundation described here is solid.

It was in this logic that, after collecting the data from the sample of 20 companies, a set of tests was applied, as already mentioned. Pierson's correlation indicated that all model variables are strongly correlated with market share. The correlation with the New Clients' variable presents a $p = 0.9379$, for the Retention variable a $p = -0.8968$, for the Up-selling variable a $p = 0.8971$ and, finally, for the variable Cross-selling $p = 0.9123$.

Thus, the correlation between the variables New Clients, Up-selling and Cross-selling with the market share is positive indicating that they are heavily dependent. At the same time, the Retention variable correlates negatively with the market share, also showing a high dependence.

In parallel, a linear regression was made through r^2 in order to understand how much of the market share variation is matched by the variation of each "Pereira Market Scan" model variable. The results are quite illuminating of their relationship. Between the New Clients variable and the market share, r^2 was equal to 0.8797 (87.97%). For the Retention variable, a $r^2 = 0.8043$ (80.43%) was obtained. The Up-selling and Cross-selling variables presented r^2 equal to 0.8049 (80.49%) and 0.8322 (83.22%), respectively.

The results of this linear regression show that the variation of the market share of the companies in the sample is explained with great confidence by the variation of the four variables of the "Pereira Market Scan" model. This finding allows us to proceed to the description of the result of the next step that involved the realization of a multiple linear regression.

The goal was to obtain a parametric equation that explained the predictive model by relating the oscillation of its variables with the variation of the market share, in a predictive way. Thus the objective was an equation $Y = M_1.X_1 + M_2.X_2 + M_3.X_3 + M_4.X_4 + C$, obtained through the use of a multiple linear regression, where Y affects the market share variation and the factors $M_1.X_1$ until $M_4.X_4$ refer to the variation of the four variables of the model plus the constant. The creation of this model allows to know the marginal effect of each variable of the tool.

For the present case the expression obtained was reflected in this way:

$$\Delta QM = M_1 \times \Delta NC - M_2 \times \Delta R + M_3 \times \Delta US + M_4 \times \Delta CS + C$$

in which:

ΔQM – Variation of the market share

ΔNC – Variation of new clients

ΔR – Variation of retention of actual clients

ΔUS – Variation of sales in Up-Selling

ΔCS – Variation of sales in Cross-Selling

It can be also emphasized that the degree of confidence obtained associated with the multiple linear regression in question is reflected in 96.31%, which provides good indicators on the reliability of the equation obtained. Thus, the equation used was the following:

$$\Delta QM = 0,5759\Delta NC - 0,4152\Delta R + 0,1952\Delta US + 0,3087\Delta CS - 0,0035$$

In this way, the presented results infer that the variables of the model are strongly correlated with the market share of the sample of companies

analyzed and that explain it with a high degree of confidence. At the same time, an equation was obtained that expresses the predictive relationship between the variables and the market share, thus helping to estimate the market share, indicating the weight of each variable over the market share and indicating where to act in case it is intended to increase market share.

Regarding this results, it can be said that, for the tool to return the results reliably, it is necessary to ensure that organizations are able and have the capacity to carry out an assiduous and computerized collection based on the history (performed) so that the variables of the model can be permanently updated. Only in this way, it will be possible to meet the objectives of the model effectively.

At the same time, the current paradigm conspires to use the data collected in an intelligent way. The well-known Big Data that so many companies apply to self-diagnose is giving way to Smart Data through which companies understand themselves, understand the market, and adapt to their vicissitudes in order to survive.

The variables of the “Pereira Market Scan” model are data-driven, proposing an intelligent analysis in order to provide the company with strategic information that indicates directions coherent with its challenges. There is then a need to place the sensors in the As-is so that the performance delta of the organization can be analyzed in a smart way in order to improve the value proposition to the customer.

The study focused on the retail sector comprising a sample composed of 20 companies. It is understood that the equation obtained is adequate for this sector and lacks validity over other sectors. Thus, as a future necessity, the replication of the study on other sectors in order to obtain accurate and adapted equations for each one, as well as the exploration of a holistic

equation, remains open and is assumed as a challenge to be achieved in the near future.

6 Conclusion

The tool developed responds to the needs of any manager, who operates in the retail sector, to make a management decision about his business and to increase the market share of his company in a more conscious, objective and quantifiable way.

The market will thus be endowed with a solution that was derived from a mathematical model based on business metrics, thus allowing managers to make decisions considering reliable data that depict reality as well as visualize the behavior of these metrics over time.

The model allows, therefore, to ensure that the causes of the real problems that affect the organizations are quickly identified so that solutions can be secured to mitigate the negative impacts and thus can provide benefits to all the stakeholders that are in interaction with organizations. The most obvious consequence is the leverage of market share.

Finally, it is important to underline the predictive characteristic of the "Pereira Market Scan" that innovates the current paradigm of obtaining or calculating the market share. It is believed that this model stands out from the other existing tools because it presents a dynamic and predictive characteristic, with a low cost of associated and without space for the subjectivity since it has been tested and it is understood on a quantitative manner.

7 References

[1] Knight, B. and McGee, J. (2015). Market Structure: The Analysis of Markets and Competition. In Wiley Encyclopedia of Management (eds C. L.

Cooper, J. McGee and T. Sammut - Bonnici). doi:10.1002/9781118785317.weom120079

[2] Carol Carlson Dean, (1997) "The Principles of Scientific Management by Fred Taylor: Exposures in print beyond the private printing", Journal of Management History, Vol 3 Issue: 1, pp 4-17, <https://doi.org/10.1108/13552529710168834>

[3] Stern C, Deimler M (2012) The Boston Consulting Group on Strategy: Classic Concepts and New Perspectives, 2nd Edition, pp 432

[4] Christopher M, Glissmeyer M, Capps C (2013) Mapping An Internal-External (I-E) Matrix Using Traditional And Extended Matrix Concepts, The Journal of Applied Business Research, Vol 29

[5] Amatulli C, Caputo T, Guido G (2011) Strategic Analysis through the General Electric/McKinsey Matrix: An Application to the Italian Fashion Industry. International Journal of Business and Management. Vol 6. pp 61-75

[6] Ghochani S, Mohamadreza F, Alavije K (2012) Application of Space Matrix. In: Developing Country Studies. Vol 2

[7] Kotler P (1984) Marketing Management: Analysis, Planning and Control. In: Prentice-Hall. pp 1-18

[8] Akiva B, Bierlaire M (1999) Discrete Choice Methods and their Applications to Short Term Travel Decisions. In: Handbook of Transportation Science. Kluwer. Springer. pp 5-33. Doi: 10.1007/978-1-4615-5203-1{ }2

[9] Cooper L, Nakanishi M (1988) Market-Share Analysis: Evaluating Competitive Marketing Effectiveness, Kluwer Academic Publishers, Boston

[10] Pereira L, Teixeira C, Salgado A (2017) Pereira Diamond: Projects Economic and Social Impacts. In: International Conference on Engineering, Technology and Innovation (ICE/ITMC), Funchal, pp. 6-14. doi: 10.1109/ICE.2017.8279862

[11] Pearson K (1920) Notes on the history of correlation. Biometrika 13: 25-45.

[12] Groemping U (2006) Relative Importance for Linear Regression in R: The Package relaimpo. Journal of Statistical Software, vol 17, pp 1-27. Doi: 10.18637/jss.v017.i01

Forecasting health of complex IT systems using system log data

Shivshanker Singh Patel

Institute of Rural Management Anand, INDIA
{shivshanker@irma.ac.in}

Abstract. Predicting the health of digital infrastructure is a vital issue to keep minimum downtime to maintain a high service level. This research work is an applied predictive analytics to forecast future health of complex IT (Information Technology) based infrastructure. Every subset of significant complex IT infrastructure at the single machine level, it tracks run-time status for generating system messages, error events, and log files. This research has suggested *3-steps* method building a novel predictive analytics model using text mining algorithm for extract features from the log. Further, it provides a model for critical device selection. At last steps, it suggests a forecasting model that can be used to predict the health of infrastructure for given time stamp. The models are built using a different algorithm to transform data for a time series modeling. The Time-series models are built using GL-ARMA and support vector machine. This research also used the selection of optimal model parameters, the selection of the most useful features to obtain a high-performance prediction model. This approach can be readily applied to many other types of information technology-based medical and energy infrastructure, and other applications also.

Keywords: Machine log, text mining, Time-Series data

1 Introduction

In a complex IT-based infrastructure (*see Fig.1*), to keep a high service level for the customer the organizations follow a preventive maintenance approach. In which scheduled maintenance with a big team of technicians. This approach leads to be high-cost operations as service level goes up. However, the predictive maintenance approach is undoubtedly a less costly alternative due to its ability to make predictions concerning failure or error when and what servicing is required. To build a predictive model, it requires past data that can provide information about the state of every component at any given time in the infrastructure. For example, in the case of bank ATMs (automatic teller machines), every device operations e.g., card reader, cassette, cash dispenser, and generating a receipt are stored in its machine logs. Also, every errors and warning occurring in an ATM gets recorded in logs. These records carry the errors contain informative messages, and it helps to build a predictive model. This model

essentially for monitoring the status of a machine and gaining knowledge to be able to discover potential faults in advance of their actual occurrence. In the following sections, related literature, data description, and a three-step predictive model are discussed.

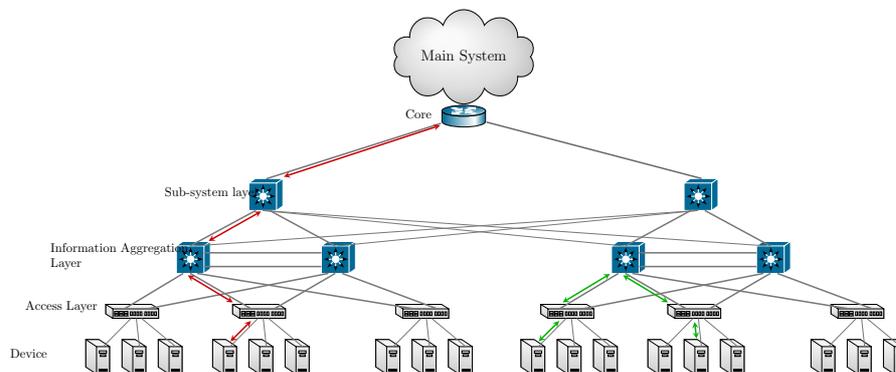


Fig. 1. Complex IT Infrastructure

2 Related Literature

There is the model based on pattern recognition [1] and using principal component analysis (PCA) to detect anomalies out of parsed textual log messages [2–4]. Another approach for log-base using a Multi-Instance Learning (MIL) [5,6] technique to build predictive models from log data. Some other models based on sequential pattern mining [8] statistically relevant patterns. Rule-based expert systems [9] other similar models to sequential pattern mining in that rules are defined based on preconditions, which can be considered to be patterns that can be matched to input data. Further, the survival model [10] and Cox model [11] are also related to this work. There other works in the literature but in isolation they can not solve the problem of complex IT-based infrastructure, some of the above methods may be able to solve failure prediction problems for a specific class of equipment to predict failures for log-based equipment. However, some concepts and insights from these articles are useful and important for building a general solution. Further, the methods for binary time series analysis are available in the literature [12–14] and the methods based on SVR have gained much more importance in recent times [15].

3 Problem statement

In a complex IT-based infrastructure once a device fails, the equipment owner issues a maintenance service request (or a ticket), that is sent to a technician

to repair the failed device if it cannot be fixed remotely (e.g., by applying a software update). In the Predictive Maintenance scenario, devices generate log data, which is the property of the equipment owner (e.g., the bank), whereas the maintenance support system owns the ticket data. To support the model, the collection of both log and ticket data apply machine learning techniques to build a prediction model. So then deploy the model (including a runtime engine for running the model) which receives real-time log/system messages, and generates failure alerts and reports those alerts to the system.

4 Data description

In a complex IT-based infrastructure once a device fails, the equipment owner issues a maintenance service request (or a ticket), that is sent to a technician to repair the failed device if it cannot be fixed remotely (e.g., by applying a software update). In the Predictive Maintenance scenario, devices generate log data, which is the property of the equipment owner (e.g., the bank), whereas the maintenance support system owns the ticket data. To support the model, the collection of log message or switch data for applying machine learning techniques to build a prediction model. So then deploy the model (including a runtime engine for running the model) which receives real-time log/system messages, and generates failure alerts and reports those alerts to the system. Two types of instances can be possible of interest which can define the health of the system. Namely, 1) Warning, 2) fault, Warning is the sign of malfunctioning but not the system is not down. However, the fault translates to the system breakdown. This type of data is an essential part of any predictive model. Notably, for Complex IT-based systems it becomes more, and that requires the data from different categories such as maintenance records, systems log, messages, inventory data, utilization data. Each of this data sources carries a specific type of information as explains below.

- log messages provided by equipment owner
- maintenance records provided by maintenance service provider

The maintenance records contains information related to service requests, it is also called as ticket. which includes the specific information such date and time of ticket generation, resolution, call backs and textual information about the devices, error. Information about the team that rectified the error. The log message (i.e. “switch data”) data on real time basis captures time stamp, message content, priority, machine type, code, serial number, location and installation date. Some of the rich data can be even captured such temperature, humidity, pollution level. For example a simple system of ATM is used and shown in Fig. 2, and 3.

The Switch data is machine generated message, Example of switch log message could be "The print operation has been completed, MDS Status is 4000PR01:3F:40:0B MDS Description is There is no receipt for the receipt printer to retain.The consumer might have taken the receipt before the

3537528	SBI	SBIR20006	Platinum	H1	FFES00155	Rajasthan KOTA	Shop No-J	Offsite	Urban	Switch Fei	Monitorin	Dispenser	ATM Hard	f
3508787	SBI	SBIK10000	Platinum	H1	FBEX0008	Karnataka Bijapur	Bijapur m	Onsite	Urban	Switch Fei	Monitorin	JP Hardw	ATM Hard	f
3505716	SBI	SBIK10064	Platinum	H1	FBEX0007	Karnataka Gangavati	Gangavati	Onsite	Urban	Switch Fei	Monitorin	ATM Dowi	NULL	f
3502728	SBI	SBIR20068	Platinum	H1	FFES00155	Rajasthan KOTA	2-N-7 Ma	Offsite	Urban	Switch Fei	Monitorin	ATM Dowi	NULL	f
3529802	SBI	SBIK20007	Gold	H2	FBEX0009	Karnataka Bangalore	No 1, 15th	Offsite	Urban	Switch Fei	Monitorin	ATM Dowi	Connectiv	f
3505575	SBI	SBIK10008	Platinum	H1	FBEX0011	Karnataka Bangalore	L K PLAZA,	Onsite	Urban	Switch Fei	Monitorin	ATM Dowi	NULL	f
3553838	SBI	SBIO2000	Platinum	H1	FFER0000	Odisha BARIPADA	AT/PO-BA	Offsite	Urban	Switch Fei	Monitorin	Superviso	ATM Hard	f
3533090	SBI	SBII10004	Silver Plus	H5	FBET0026	Gujarat BHADALI	Bhadali Br	Onsite	Semi Urbi	Switch Fei	Monitorin	ATM Dowi	Connectiv	f
3510254	SBI	SBIR20068	Platinum	H1	FFES00155	Rajasthan KOTA	2-N-7 Ma	Offsite	Urban	Switch Fei	Monitorin	ATM Dowi	NULL	f
3513344	SBI	SBII10000	Gold	H2	FBET0152	Gujarat DOLVAN	DOLDAN-	Onsite	Rural	Switch Fei	Monitorin	ATM Dowi	NULL	f
3541637	SBI	SBIK20010	Gold	H2	FBEX0007	Karnataka Bangalore	3436, 1st r	Offsite	Urban	Switch Fei	Monitorin	ATM Dowi	ATM Hard	f
3503772	SBI	SBIG20001	Silver Plus	H5	FBEW0005	Goa TISK USGA	Shop No.-	Offsite	Semi Urbi	Switch Fei	Monitorin	ATM Dowi	NULL	f

Fig. 2. Log Message from Ticket

Call Type	Category	Sub Categ	Status	ATM Status
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down	Investigati	Bulk Down	Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
JP Hardware Fatal	ATM Hardw	JP Fatal	Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
ATM Down			Closed	Machine Down
RP Hardware Fatal	ATM Hardw	RP Fatal	Closed	Machine Not Down
ATM Down			Closed	Machine Down
Cash Out -Critical			Closed	Machine Down

Fig. 3. Log Message from Switch

receipt printer could retain it, or the receipt printer". This message will appear with the with time stamp, Machine ID and other details.

5 Solution

A big complex, IT system would be made of many components. Every component and subsequently every component produce logs. A prediction model build based on a log message for consumers and equipment owners. The primary concern for crucial performance evaluation metrics of the prediction model is the high cost associated with the wrong prediction. Due to the cost associated with moving the maintenance team. Thus, the critical requirement is that all failures should be predicted and prevented in due time. The false alarms raised by the model will lead to the undesired cost of parts replacement and human effort.

A *Three step predictive model* is of integrated in nature. This research work is an applied predictive analytics to forecast future health of digital infrastructure. Every Machine, track run-time status by generating system messages, error events, and log files. Three steps are as follows that is explained in detail modeling section. As depicted in Fig.4 *step-1* Using text mining algorithm for extract features from the log regarding devices of IT systems. *step-2* an algorithm is devices to help and find the vital device selection from the huge number of devices that explain the vulnerability of the complex system. Moreover, *step-3* it provides a forecasting model that can be used to predict the health of infrastructure. The models are built for transformed data obtained from *step-1* augmented information from the *step-2* to get a time series data using some specific machine learning algorithm to predict the future fault. The detail methodology is explained as follows.

5.1 Step-1: feature selections

In order to select the feature from the a log message generated by switch as given in Fig. 2, and Fig. 3. A **n-gram** algorithm based Text mining approach is adapted for selection devices and indexing of 1 and 0 based the error message that maps to machine up or down. This step is a preprocessing of the data and assigning a category to message. The basic statistical model of feature selection,

$\mathbf{B} = \{c_{ij}, i \in [1, T], j \in [1, X]\}$, where T is the number of error types, X is the number of sub-windows and c_{ij} is the error instance number of th i -th error type in the j th sub-windows. Thus the length of \mathbf{B} is $X \times T$.

5.2 Step-2: Device selection and prediction

Association rule based First of all, we can use logistic regression to identify the high likelihood of the devices that lead to the failure of the whole system or sub-system. Once that is done that we can identify that the what is the combination at this their occurrence lead to failure to one device or sub-system. The association relationships among **error types**. Using the apriori algorithm,

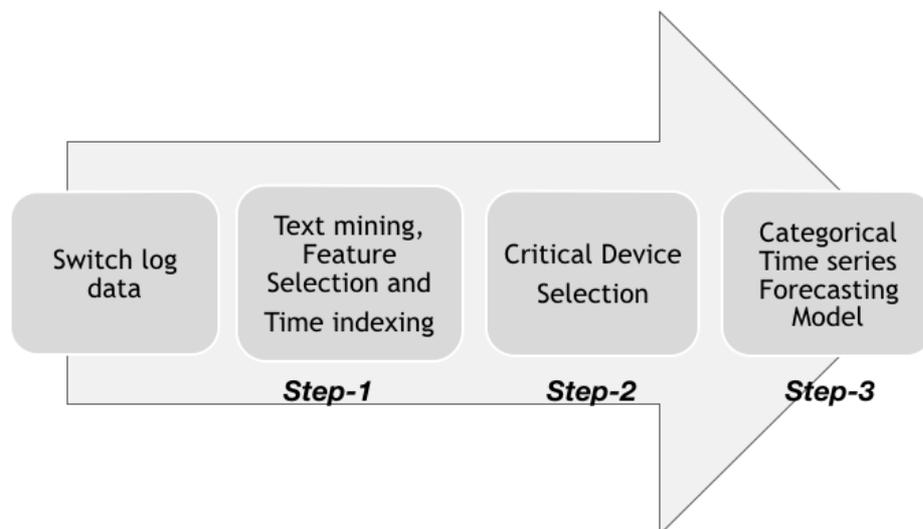


Fig. 4. Integrated Approach of Forecasting Model

we expect the model to learn the predictive insights regarding relationships between error patterns and failures. In our method, we define a **combination of devices** those possess error together types that repeat in **different instances**. The firstly it identifies all pattern candidates from all instances. For example, if three devices comes together and possess faults (d_1, d_2 , and d_3) occur in an observation window, then all combinations of these errors, without consideration of occurrence order, e.g., ($\langle e1 \rangle, \langle e2 \rangle, \langle e3 \rangle, \langle e1, e2 \rangle, \langle e1, e3 \rangle, \langle e2, e3 \rangle, \langle e1, e2, e3 \rangle$), are candidate patterns. The evaluation of each patterns confidence in predicting failures. We define the confidence of a pattern as the ratio of the count of pattern instances in the instance where the whole system shows the error to the count of pattern instances in all instances. We select a pattern as a feature if its confidence exceeds a predefined threshold. Support and confidence used appropriately for selecting the device.

Rank based If any set of devices that are failing in sequence then a rank algorithm predicts or suggests which is going to be next device that is highly likely to fail. This prediction is not actually time dependent, it is incidence based calculated based on transition probability from rank **rank algorithm** to select critical devices from a complex system.

5.3 step-3: Forecasting models

Binary Time series based model Preparation of binary time series i.e. The simplest example of a time series model is a model where the dependent variable is binary as system failure as error = 1 and no-error = 0. For simplicity, and

without loss of generality, typically, the value 1 indicates that some event occurs, and 0 that it does not occur. Now, let $y_t, t = 1, 2, \dots, T$ a univariate binary time series. From the *step-1* text mining and *step-2* select critical device that actually affect the whole system or sub-system. The value of critical devices or status played as predictor variables in the model. thus *step-3* preparing the binary time series sequence of critical devices and support vector regression based model predicts the future error.

Inter arrival time based model : every device that it fails or shows error it has some inter-arrival time this inter-arrival time is be modeled with the help of assuming either on inter-arrival time as continuous variable of interest and time-series modeling, there may be some kind of autoregressive behavior that is captured with the help of ARIMA or using any other machine learning model.

6 Application:

A data set of ATMs are used to showcase integrated 3 step model, and at every step in three steps model the different **performance criterion**, is used. In *step-1* the lift, support, and confidence. *Step-2* and *step-3* the confusion matrix and AUC and ROC, RMSE.

7 Conclusion:

In this paper we have proposed a predictive model based on machine learning method which is in an integrated manner gives us the power to predict the status of any complex IT-based infrastructure that can help us take preventive measure, or it can prepare to shoot a computer routine to run once the error occurs in an automated environment.

References

1. Li, T., Liang, F., Ma, S., Peng, W.: An integrated framework on mining logs files for computing system management. in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, (1981) 776-781.
2. Xu, W.: Detecting large scale system problems by mining console logs, Ph.D. dissertation, Univ. California, Berkeley, CA, USA, (2010).
3. Xu, W., Huang, L., Fox, A., Patterson, D., Jordan, M.: Mining console logs for large-scale system problem detection. in Proc. 3rd Workshop Tackling Comput. Syst. Problems Mach. Learn. Techn., San Diego, CA, USA, (2008) 4.
4. Xu, W., Huang, L., Fox, A., Patterson, D., Jordan, M.: Online system problem detection by mining patterns of console logs. in Proc. 9th IEEE Int. Conf. Data Mining, Miami, FL, USA, (2009) 588-597.
5. Sipos, R., Fradkin, D., Moerchen, F., Wang, Z.: Log-based predictive maintenance. in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, (2014)18671876.

6. Maron, O., Lozano-Pemerez, T.: A framework for multiple- instance learning. in Proc. Adv. Neural Inf. Process. Syst., (1998) 570576.
7. Mabroukeh, N. R., Ezeife, C. I.: A taxonomy of sequential pattern mining algorithms. ACM Comput. Surveys (CSUR), 43, 1, (2010) Art. no. 3.
8. Zaki, M., J., Lesh, N., Ogihara, M.: PlanMine: Sequence mining for plan failures. in Proc. 4th Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, (1998) 369373.
9. Butler, K., L.: An expert system based framework for an incipient failure detection and predictive maintenance system. in Proc. Int. Conf. Intell. Syst. Appl. Power Syst., Orlando, FL, USA, (1996) 321326.
10. Richards, S., J.: A handbook of parametric survival models for actuarial use. Scand. Actuarial J., 4, (2012) 233257.
11. Li, Z., Zhou, S., Choubey, S., Sievenpiper, C.: Failure event prediction using the cox proportional hazard model driven by frequent failure signatures. IIE Trans., 39, 3,(2007) 303315.
12. Benjamin, M. A., Robert, A. R., Stasinopoulos, D. M.: Generalized Autoregressive Moving Average Models. Journal of the American Statistical Association, 98, (2003) 214223.
13. Li, W. K.: Time Series Models Based on Generalized Linear Models: Some Further Results. Biometrics. 50, (1994) 506511.
14. McCulloch, C. E.: Maximum Likelihood Algorithms for Generalized Linear Mixed Models. Journal of the American Statistical Association. 92, (1997) 162170.
15. Cao, L. and Tay, F., E., H.: Support vector machine with adaptive parameters in financial time series forecasting, IEEE Transactions on Neural Networks. 14(6), (2003) 15061518.

Comparing linear and non-linear dynamic factor models for large macroeconomic datasets

Alessandro Giovannelli*

Marina Khoroshiltseva[†]

June 29, 2018

Abstract

This paper proposes a non-linear extensions for macroeconomic forecasting using a large dataset based on dynamic factor model (DFM). The main idea is to allow the factors to have a non-linear relationship to the input variables using the methods of (i) kernel and (ii) neural networks principal components analysis. We compare the empirical performances of these methods with (iii) the standard principal-component model introduced by Stock and Watson in 2002, conducting a pseudo forecasting exercises based on a Euro Area macroeconomic dataset composed by 834 monthly variables spanning the period January 1996 - September 2017. Using a rolling window for estimation and prediction, the results obtained from the empirical study suggest that (i) and (ii) have the same forecasting performances of (iii) for both Industrial Production and Inflation, but (i) significantly outperforms (iii) for the Unemployment Rate. Moreover, there is no difference with respect to previous results if we consider the pre-crisis period. However during the crisis and subsequent recovery we observe a slight improvement of (ii) with respect to (i) for Industrial Production and Inflation while (i) is the best model for Unemployment Rate.

Keywords: Kernel Methods; Principal Component Analysis; Large Dataset; Artificial Neural Networks; Forecasting.

JEL Classification Numbers: C45, C53, C13, C33.

*Department of Economics and Finance, University of Roma Tor Vergata. Corresponding address: Via Columbia 2, 00133 Rome - Italy. e-mail: alessandro.giovannelli@uniroma2.it

[†]European Centre for Living Technology (ECLT) San Marco 2940 - 30124 Venezia, ITALY. e-mail: marina.khoroshiltseva@unive.it

Simultaneous Multi-Response Multi-Covariate Best Subset Selection- with application to fault modelling¹

We employ a two-step algorithm for an automated approach to fault modelling. The first step is an extension of the work by Bertsimas et al. (2016), implementing a multivariate version of best subset selection, formulating the problem as a Mixed-Integer Quadratic Optimization (MIQO) problem. The second step is to determine the autocorrelation structure of the regression residuals. Results have shown the multivariate covariate selection to be much more accurate and able to deal with much lower signal-to-noise ratios. The MIQO framework gives us some guarantee to obtain interpretable models.

There is a complex relationship between weather variables and faults in the telecommunications network. One example is the *aggregated* effect of persistent rain over a number of days. Other complications include delays in the development of faults and of fault reporting. We seek a model to explain the behaviour between weather and faults.

We apply a number of transformations to the weather variables to emulate the lag and aggregated effects, but the parameters of these transformations are not obvious. Our method selects which weather variables, and the parameters for their transformations that best explain the faults. Further, we select the variables simultaneously across a number of regions.

We seek to fit linear models of the form

$$y_{m,t} = \sum_{p=1}^P x_{m,p,t} \beta_{m,p} + \eta_{m,t} \quad (1)$$

where,

- m : identifies the individual, we assume M such individuals,
- p : indexes the covariate, we assume P such covariates,
- t : is the time index, we assume T time points,
- $y_{m,t}$: The response variable for individual m ,
- $\{x_{m,p,t} : p = 1, \dots, P\}$: The set of covariates for region m ,

¹Aaron Lowther, July 8, 2018, e: a.lowther1@lancaster.ac.uk

- $\eta_{m,t}$ is the error term for region m at time t where we assume an

$$\text{ARIMA}(p, d, q)(P, D, Q)_s$$

structure.

We implement the multivariate best subset selection approach using the Gaussian likelihood function for the objective in the MIQO problem, using the formulation,

$$\min \sum_{m=1}^M \sum_{t=1}^T (y_{m,t} - \sum_{p=1}^P x_{m,p,t} \beta_{m,p})^2 \quad \text{subject to,} \quad (2)$$

$$(\eta_p \beta_{m,p}) \in \mathcal{SO}S - 1, \quad (3)$$

$$\beta_{m,p} \leq \mathcal{M}_u \quad (4)$$

$$\|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_l \quad (5)$$

$$-\sum_{p=1}^P \eta_p \leq k - P \quad (6)$$

$$\eta_p \in \{0, 1\} \quad (7)$$

where (3) ensures that only one of η_p or $\{\beta_{1,p}, \dots, \beta_{M,p}\}$ are different from 0, (5) and (4), bound the model coefficients and (6) allows only k coefficients for an individual to be non zero. This formulation ensures the same *family* of covariates is present in the model for each region, but allows the coefficients to vary. Once the best subsets are found the errors for each region are modelled.

We have modified discrete first order methods to produce good warmstarts for the optimization solver. The warmstarts can also be used to obtain the formulation parameters, producing a tight formulation to improve problem solve time. We are able to solve problems with up to 40 covariates and 10 regions to proven optimality. Very good solutions can be obtained if the solver is terminated early. Quite often the solutions produced are optimal but the certificate of optimality is not available.

A number of extensions to this flexible approach are possibly and are under investigation. Firstly, the model coefficients can be restricted to values in the positive half line as negative coefficients do not always make sense for the covariates. This has the effect of draining explanatory power from the other, highly correlated covariates and we do not obtain models with many highly correlated predictors with both positive and negative coefficients which is common in many other covariate selection procedures. Further, this restriction improves optimisation solve time dramatically. We are also interested in the effect of penalising *dissimilar* coefficients across regions and generalising the framework do deal with grouped time series.

References

Bertsimas, D., King, A., and Rahul, M. (2016). Best subset selection under a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

A comparison of statistical methods for estimating individual location densities from smartphone data

Francesco Finazzi and Lucia Paci

Department of Management, Information and Production Engineering,
University of Bergamo, Italy

and

Department of Statistical Sciences,
Università Cattolica, Milan, Italy

`francesco.finazzi@unibg.it`

`lucia.paci@unicatt.it`

Abstract. In this work, the focus is on location data collected by smartphone applications. Specifically, we propose and compare a set of models of increasing complexity to estimate individual location at any time, uncertainty included. Unlike classic tracking for high spatio-temporal resolution data, the approaches are suitable when location data are sparse in time and are affected by non negligible errors. The approaches build upon mixtures of densities that describe past and future locations; the model parameters are estimated by maximum likelihood. The approaches are applied to smartphone location data collected by the Earthquake Network citizen science project.

Keywords: location-based applications; maximum likelihood; normal mixtures; spatio-temporal patterns

1 Introduction

The advent of GPS technology had a huge impact across a variety of fields. The main limit of GPS tracking is that the object to be tracked must carry a GPS receiver. When it comes to people, the limit is even more relevant as people are not supposed to carry around a GPS receiver. However, things have changed with the smartphone revolution. Nowadays, smartphones have a large number of built-in sensors and they are usually equipped with a GPS receiver. While this may suggest that tracking smartphones (and thus people) is an easy task, there are some issues to discuss.

Assuming that the smartphone user gave the permission to be tracked, the first problem is that the GPS receiver is likely off. Indeed, the receiver has an impact on battery consumption and it is usually on only when the smartphone is used for navigation purposes. Therefore, the smartphone location is usually given by the service provider's network infrastructure or wi-fi networks at lower accuracy than GPS. Additionally, independently of the way the location is obtained,

acquiring and storing the smartphone location at high temporal frequency is considered a malpractice in smartphone application programming. Last but not least, the smartphone may stay off for long periods of times (from hours to days) and then no information on its location is collected. The above considerations suggests that, if the tracking does not have to negatively affect or alter the smartphone user experience, then the information on the smartphone location is available at low temporal resolution and with a lower precision if compared with GPS-based tracking.

This paper addresses the problem of estimating the smartphone/person location at a given point in time considering all the locations collected. Such problem is increasing relevant in many fields. For instance, if an area is affected by a natural disaster, it is useful to estimate the last location of missing people [1]. In epidemiological studies, pollution exposure can be dynamically assessed at population level if people locations are known at high temporal resolution [2]. Also, from a commercial perspective, individual recommendation or advertising can be provided to people whose location is known [3].

The analysis of individual location data has been explored by several authors to predict short-term trajectories. Customary, approaches used for human/animal tracking are based on interpolation methods, such as splines. [4] offered a comparison between interpolations methods at different temporal resolutions for animal tracking. Alternatively, [5] employed a Bayesian dynamic network for learning transportation routines between locations where the person spends a given amount of time and building personal map based on their behavior.

However, it is hard to recover the actual trajectory of a smartphone when the temporal resolution of the available locations is low and in the absence of additional information such as speed and acceleration in space. Secondly, when the smartphone user moves from a point \mathbf{s} to another point \mathbf{s}' in space, the followed path is rarely the shortest path in terms of Euclidean distance. This is because people are constrained by both natural topography (e.g., mountains) and man-made artifacts (streets, roads, etc.). For these reasons, tracking methods based on dynamic modeling are not suitable in this context.

A different approach to predict individual's location relies on the reproducibility of human patterns. Indeed, daily and weekly routines are well-established in human societies such that human activities are characterized by a certain degree of regularity and predictability [6, 7]. In this framework, [8] provided a spatio-temporal approach to predict arrival and residence times of users in their relevant places. [9] analyzed functional mobile data to identify subregions of the metropolitan area of Milan (Italy) sharing a similar pattern along time, and possibly related to activities taking place in specific locations and/or times within the city.

In principle, if we assume that people spend most of their time at few spatial locations (home, work, gym, etc.) than it is possible to group all the observed spatial locations into a small number of spatial clusters. As common in literature, clusters might be identified using finite mixture modeling. For instance, [10]

employed a two-state mixture of Gaussian distributions centered at “home” and “work” locations to understand human motion from cell phone data and social networks. Also, [11] modeled human geolocation data from social networks by means of mixtures of kernel densities that allow to smooth individual’s models towards an aggregate population model. However, the authors focused on user’s spatial patterns ignoring the temporal dimension, i.e., assuming a time-invariant location density.

Mixture models can be extended to allow the model parameters to vary over time in order to represent complex dynamic distributions [12, 13]. For instance, we may want to give more weight to a given mixture component depending say on the time of the day or the day of the week. Then, a possible form for the mixture weights may depend on time-varying covariates or even dynamic processes. However, the resulting parameter space would be high dimensional and model estimation would become very challenging with existing algorithms. Moreover, the number of clusters should be estimated for each smartphone user, increasing the computational burden.

Our contribution is to propose flexible approaches that build upon mixtures of Gaussian distributions that describe past and future observed locations (with respect to each time). Time-varying mixing weights are introduced to estimate the location density at any time by exploiting temporal dependence as well as the fact that people follow reproducible patterns, such as daily and weekly patterns. Moreover, the precision of the mixture components depends upon the precision associated with each location; in other words, we also account for the positional error [14] arising in smartphone data. As a result, the number of model parameters is small and the number of clusters does not need to be estimated. A comparison among the proposed models shows the benefit of exploiting cyclical spatio-temporal patterns of people.

Our motivating data consist of smartphone locations collected by the Earthquake Network citizen science project (www.earthquakenetwork.it), which implements a world-wide early warning system based on smartphones. This is an instance of location data collected by a smartphone application (app) which makes use of geolocation but the primary role of which is not tracking. Clearly, the approach can be applied for modeling location data gathered by any location-aware app, including social networks.

The remainder of the paper is organized as follows. Section 2 introduces smartphone data and describes the inferential problem. Section 3 specifies the models proposed to describe the location density function, with estimation details provided in Section 3.4. An application to real data is illustrated in Section 4. We conclude with a brief discussion in Section 5.

2 Smartphone data and general set-up

Given a smartphone, a location-aware app is usually needed to acquire, store and send the smartphone location to a central server. Here, we assume that a smartphone app periodically acquires the smartphone location. If the Internet

connection is available, the location is immediately sent to a central server otherwise it is stored and sent when the connection becomes available. The location provided by the smartphone at the generic time t is given as the probability density function of a bivariate Normal distribution centered on $\tilde{\mathbf{s}}_t$ and with variance $\boldsymbol{\Sigma}_t$, i.e., $\varphi_t = \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_t, \boldsymbol{\Sigma}_t)$, $\mathbf{s} \in \mathbb{R}^2$. In particular, $\tilde{\mathbf{s}}_t$ is given by the easting and northing coordinates within Universal Transverse Mercator (UTM) zone, while $\boldsymbol{\Sigma}_t = \sigma_t^2 \mathbf{I}_2$ is a diagonal matrix, where $1/\sigma_t$ is the location precision and \mathbf{I}_2 is the 2×2 identity matrix. Here, we assume to have a collection of locations with associated precisions observed at irregularly times over a fixed period. In fact, the smartphone may stay off for long periods or the location may not be available when requested by the app.

In this work, we employ a classical measurement error model (MEM; [15]) by assuming that the location provided by the smartphone is a noisy version of the “true” unknown location \mathbf{s}_t , that is:

$$\tilde{\mathbf{s}}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t, \quad (1)$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_t)$. Therefore, the goal is to estimate the probability density function of the true location \mathbf{s}_t over the projected geographic region $\mathcal{D} \in \mathbb{R}^2$ at the generic time t for any smartphone, that is

$$f_t \equiv f_t(\mathbf{s}; \mathcal{X}, \boldsymbol{\theta}), \quad (2)$$

where \mathcal{X} denotes the information set and $\boldsymbol{\theta}$ is a vector of parameters. Given m observed locations for a generic smartphone, \mathcal{X} contains all the observed locations with associated variances.

Using the density in (2), we can also provide a point estimate of the smartphone location at any time t . To accomplish that, different solutions can be considered such as the mean of f_t , the median, the mode and so on. Since, in general, f_t is multimodal, we estimate the smartphone location by the mode

$$\hat{\mathbf{s}}_t = \arg \max_{\mathbf{s} \in \mathcal{D}} f_t(\mathbf{s}; \mathcal{X}, \hat{\boldsymbol{\theta}}) \quad (3)$$

where $f_t(\mathbf{s}; \mathcal{X}, \hat{\boldsymbol{\theta}})$ denotes the density function of the smartphone location corresponding to parameter estimate $\hat{\boldsymbol{\theta}}$. By product, point estimator in (3) can be employed for online tracking when data are available at high temporal resolution. However, we stress that smartphone tracking is beyond our goal. Rather, we are interested in location density (2) and, potentially, its multiple modes. For instance, it might be useful to know all the locations where people are more likely to be if they are missing, say after a disaster.

3 Location density functions

In this section, parametric models of increasing complexity are introduced in order to describe f_t .

3.1 Tracking-like

Let $t' \leq t$ and $t'' \geq t$ be the nearest sampling times to a generic time t observed in the past and in the future, respectively. Hence, $\{\tilde{\mathbf{s}}_{t'}, \boldsymbol{\Sigma}_{t'}\}$ and $\{\tilde{\mathbf{s}}_{t''}, \boldsymbol{\Sigma}_{t''}\}$ denote the locations and associated variances at time t' and t'' , respectively. The tracking-like model relies on the Gaussian assumption and assumes a Markovian structure such that the conditioning set \mathcal{X} in (2) reduces to locations and variances of the nearest times from t . The resulting density is,

$$f_t = k_t \mathcal{N}_2(\mathbf{s}; \boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t) I(\mathbf{s} \in \mathcal{D}), \quad (4)$$

where k_t is the normalizing constant and

$$\boldsymbol{\mu}_t = w_t \tilde{\mathbf{s}}_{t'} + (1 - w_t) \tilde{\mathbf{s}}_{t''}, \quad (5)$$

$$\boldsymbol{\Lambda}_t = g(t; \alpha) [w_t \boldsymbol{\Sigma}_{t'} + (1 - w_t) \boldsymbol{\Sigma}_{t''}] \quad (6)$$

and

$$w_t = 1 - \frac{t - t'}{\Delta_t}, \quad (7)$$

$$g(t; \alpha) = (1 + \exp(\alpha) \Delta_t)^{w_t(1-w_t)}, \quad (8)$$

with $\Delta_t = t'' - t'$ the sampling interval, $I(\mathbf{s} \in \mathcal{D})$ the indicator function equal to 1 if $\mathbf{s} \in \mathcal{D}$ and 0 otherwise. The density f_t has a maximum located over a straight line that connects $\tilde{\mathbf{s}}_{t'}$ to $\tilde{\mathbf{s}}_{t''}$ and depends on the temporal distance $t - t'$. The linear combinations in (5) - (6) are such that

$$f_{t=t'} = k_t \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_{t'}, \boldsymbol{\Sigma}_{t'}) I(\mathbf{s} \in \mathcal{D}), \quad (9)$$

$$f_{t=t''} = k_t \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_{t''}, \boldsymbol{\Sigma}_{t''}) I(\mathbf{s} \in \mathcal{D}), \quad (10)$$

namely, at the sampling times t' and t'' , the density f_t is equal to the density φ_t sent by the smartphone and normalized over \mathcal{D} .

Assuming $\alpha > 0$, the term $g(t; \alpha)$ in (8) is equal to 1 when $t = t'$ and $t = t''$, while it is maximum when $t = (t' + t'')/2$. This model is suitable to describe a smartphone/person that moves from $\tilde{\mathbf{s}}_{t'}$ to $\tilde{\mathbf{s}}_{t''}$ on a straight path and at a constant speed, with precision on the starting and ending locations given by $1/\sigma_{t'}$ and $1/\sigma_{t''}$, respectively. The term $g(t; \alpha)$ modulates the location uncertainty along the path and it is high if Δ_t is high. Moreover, for $\Delta_t \rightarrow \infty$, the density f_t converges to the density of the uniform distribution over the geographic region \mathcal{D} , since $f_t = 0$ outside \mathcal{D} . Finally, the higher α the more the straight path hypothesis is violated. In fact, a high value implies that f_t is more spread over \mathcal{D} and thus the potential smartphone location is not restricted to lie on a straight path.

3.2 Bimodal density

The tracking-like model discussed above may be suitable when the sampling interval Δ_t is relatively small, say less than 5 minutes. In practice, if the consecutive sampling times are far apart (hours or days), the straight path and

constant speed assumption would be unrealistic. Thus, when Δ_t tends to be large, we can assume that, for any $t' < t < t''$, there are no reasons to prefer any other location different from $\tilde{\mathbf{s}}_{t'}$ or $\tilde{\mathbf{s}}_{t''}$. Indeed, people tend to stay in the same location for long periods rather than constantly moving across space. Thus, assuming that the Markovian structure still holds, the following mixture model is proposed:

$$f_t = k_t \left[w_t \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_{t'}, g(t; \alpha) \boldsymbol{\Sigma}_{t'}) + (1 - w_t) \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_{t''}, g(t; \alpha) \boldsymbol{\Sigma}_{t''}) \right] I(\mathbf{s} \in \mathcal{D}), \quad (11)$$

where w_t and $g(t; \alpha)$ are defined in (7) and (8), respectively and k_t is the normalizing constant.

Contrary to (4) which is unimodal, for any $t' < t < t''$ the density f_t in (11) is bimodal with modes in $\tilde{\mathbf{s}}_{t'}$ and $\tilde{\mathbf{s}}_{t''}$. In this case, the parameter α describes how “fast” the locations $\tilde{\mathbf{s}}_{t'}$ and $\tilde{\mathbf{s}}_{t''}$ become unreliable when moving far in time from the sampling times t' and t'' , respectively. Conditions (9) and (10) still hold for model (11).

3.3 Full-history mixture

In order to exploit all the available information on the smartphone location, we relax the Markovian structure and assume, at each time, a mixture of densities describing all past and future (with respect to t) locations. Therefore, the proposed density f_t is

$$f_t = k_t \left[\sum_{t' \in \mathcal{T}} v(t, t'; \phi) \mathcal{N}_2(\mathbf{s}; \tilde{\mathbf{s}}_{t'}, \alpha \boldsymbol{\Sigma}_{t'}) \right] I(\mathbf{s} \in \mathcal{D}), \quad (12)$$

where, again, k_t is the normalizing constant, \mathcal{T} is the set of observed times and $v(t, t'; \phi)$ are the mixture weights defined as

$$v(t, t'; \phi) = \frac{l(t, t'; \phi)}{\sum_{t' \in \mathcal{T}} l(t, t'; \phi)}, \quad (13)$$

with

$$l(t, t'; \phi) = \exp\left(-\frac{|t - t'|}{\phi_1}\right) \exp\left(-\frac{|h(t, t')|}{\phi_2}\right) \exp\left(-\frac{1 - d(t, t')}{\phi_3}\right). \quad (14)$$

The parameter α in (12) affects the variance of the components of the mixture such that the higher α , the more f_t is spread over the geographic region \mathcal{D} . When $\alpha \rightarrow \infty$ the density converges to the uniform density over the region.

The mixture weights (13) describe the daily and weekly cycle of smartphone users. Indeed, the function $h(t, t')$ returns the difference in time between t and t' , independently of the calendar day, namely $|h(t, t')|$ is always less than 12 hours even when t and t' are more than 12 hours far apart. On the other hand, the

function $d(t, t')$ is equal to 1 if t and t' are both working days or both weekends, otherwise it is equal to 0. In practice, $v(t, t'; \phi)$ tends to be high when t and t' are close in time and/or when they are characterized by a similar time within the day and/or when they are days of the same type. The unknown parameters $\phi > 0$ in the exponential terms of (14) describe the temporal persistence of the information carried by each mixture component. In particular, the persistence over time increases as ϕ_1 increases, while ϕ_2 modulates the intra-day persistence, i.e., the higher ϕ_2 the less the time within the day matters. Finally, the higher ϕ_3 the higher the weekend effect.

When $t \ll t_1$ or $t \gg t_m$, namely when f_t is used to predict the smartphone location far in time from the first or the last available observed location, the first exponential term in (14) approaches 0. However, the constraint on $v(t, t'; \phi)$ implies that f_t still reflects the daily and the weekly cycles estimated from the information set. In this sense, the daily and weekly cycles characterize the smartphone location regardless of the position of t along the time line, and f_t is informative even if the smartphone has not sent its location for a long time.

3.4 Model estimation

For a smartphone i ($i = 1, \dots, n$) at time $t_{i,j}$ ($j = 1, \dots, m_i$), we observe its location $\tilde{\mathbf{s}}_{i,j}$ with associated variance $\Sigma_{i,j}$. Note that we suppressed the index t to simplify the notation. Let $\mathcal{X}_i = \{\tilde{\mathbf{s}}_{i,1}, \dots, \tilde{\mathbf{s}}_{i,m_i}, \Sigma_{i,1}, \dots, \Sigma_{i,m_i}\}$ denote the information set of smartphone i ; hence, $\mathcal{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_n\}$ is the full information set. Recall the MEM in (1); since $\varepsilon_{i,j}$ is normally distributed and also $\mathbf{s}_{i,j}$ is assumed normally distributed or a mixture of normal distributions, then the likelihood function for smartphone i is given by the product of m_i normal densities or by the product of mixtures of normal density functions.

The information set \mathcal{X} is exploited to estimate θ_i for each smartphone i . Let \mathcal{L}_h denote the log likelihood for model $h = 1, \dots, 3$; therefore, the maximum likelihood estimation is provided by

$$\hat{\theta}_i = \arg \max_{\theta_i} \log \mathcal{L}_h(\theta_i; \mathcal{X}), \quad (15)$$

under the constraint that all the elements of θ_i are positive.

4 Analysis of Earthquake Network data

Data used to illustrate models described above, come from the Earthquake Network project [16] which implements a world-wide early warning system based on smartphones. The earthquake detection is based on the signals sent by the smartphones every about 30 minutes to a central server, with information on the smartphone location and its precision. Additionally, the smartphone sends a signal when an acceleration above a threshold is detected. This implies that the sampling interval of the smartphone location is not regular, plus the smartphone does not send signals when it is switched off or Internet is not available.

Here, smartphones located in the metropolitan area of Rome, Italy, are considered. The dataset consists of 4.1 millions signals sent by 1336 smartphones to the server in the period January, 1st - April, 30th 2017 over the geographic box (41.75°N, 42.00°N, 12.35°E, 12.65°E). As an example, Figure 1 shows the locations observed for three different smartphones with associated standard deviations represented by the disks.

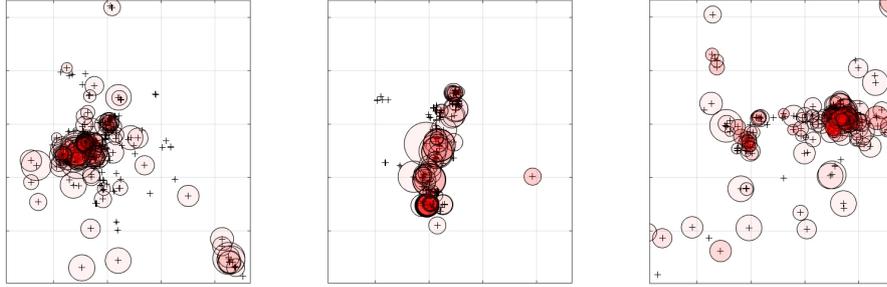


Fig. 1. Locations observed for three users; the center of the disk is the observed location while the radius of the disk is the associated standard deviation.

For each of the 1336 smartphones, the three models introduced in Section 3 are estimated adopting the maximum likelihood approach described in Section 3.4. Since the likelihood function may be characterized by multiple local maxima and the maximization in (15) is not guaranteed to converge to the global maximum, the maximization is carried out multiple times with different starting values for the parameter vector θ_i . The estimation $\hat{\theta}_i$ with the highest likelihood is then retained.

Before presenting the estimation results, a flavor is given of the density f_t involved in the three models. Figure 2 shows $\log f_{t^*}(\mathbf{s}; \mathcal{X}, \hat{\theta}_i)$ for smartphone $i = 1$ and t^* equal to February 5th, 2017 14:42:14 local time. Time t^* is a sampling time for the location of smartphone i but $f_{t^*}(\mathbf{s}; \mathcal{X}, \hat{\theta}_i)$ is evaluated assuming that the smartphone location at t^* is unknown. In each panel of Figure 2, the diamond and the circular markers are the known locations of smartphone i at the nearest time from t^* in the past, say t_{-1}^* , and in the future, say t_{+1}^* , respectively. In particular, time t_{-1}^* is 13:09:10 while t_{+1}^* is 15:10:32. The dotted circles around each marker depict the smartphone location uncertainties, $\sigma_{t_{-1}^*} = 20.1 m$ and $\sigma_{t_{+1}^*} = 900.0 m$ (the circle around the diamond smartphone may not be visible since very small with respect to geographic area). The star marker is $\tilde{\mathbf{s}}_{t^*}$, which is supposed to be unknown. Note that $f_{t^*}(\mathbf{s}; \mathcal{X}, \hat{\theta}_i)$ has a different number of modes depending on the density model. In particular, only one mode for the tracking-like density, two modes for the bimodal model and multiple

modes for the full-history. In all the cases, $f_{t^*}(\mathbf{s}; \mathcal{X}, \hat{\theta}_i)$ is high at the locations \mathbf{s} where the smartphone may be located at time t^* . The square marker is the maximum $\hat{\mathbf{s}}_{t^*}$ (see equation (3)). If we were to find the smartphone at time t^* , then $\hat{\mathbf{s}}_{t^*}$ would be the location we would search first.

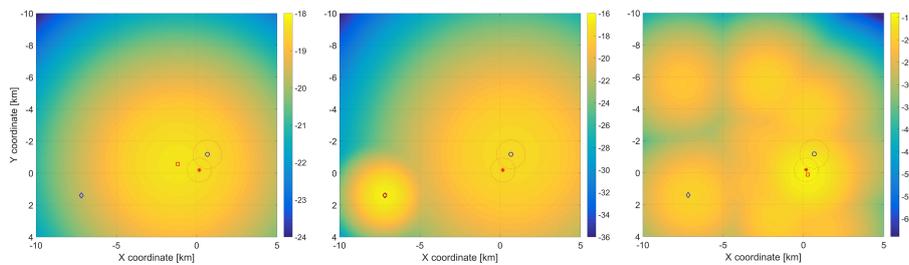


Fig. 2. Location densities under the three models at time t^* for a given smartphone: tracking-like density (left panel); bimodal density (mid panel) and full-history mixture (right panel). The star is the observed location at time t^* . Diamond and the circular markers are the locations at the nearest times from t^* ; the dotted circles around each marker depicts their standard deviations.

The histograms in Figure 3 depict the estimated $\hat{\alpha}_i$ values for the 1336 smartphones when f_t is modeled using the tracking-like and the bimodal density. Note that an high $\hat{\alpha}_i$ value implies a large uncertainty on the smartphone location at time t . In this case, the average $\hat{\alpha}_i$ is equal to 48.4 and 32.9 for the tracking-like and the bimodal model, respectively. On the other hand, Figure 4 shows the parameter estimation results for the full-history density model. Note that $\hat{\alpha}_i$ is smaller since the model exploits all the past and future information on the smartphone behavior in space and time. Figure 5 shows the estimates of weights $v(t, t'; \phi)$ for the three smartphones the locations of which are plotted in Figure 1. These are the actual weights used to estimate the location density corresponding to February 14th, 2017 12:00:00 for the three smartphones.

The three density models are then compared using the AIC criterion. In Table 1, it is reported the percentage of times a model (row) is better than the other (column) when the comparison is based on AIC. For instance, the tracking-like density outperforms the other models only less than 1% of the times. The bimodal model is better than the tracking-like model around 99% of the times. The full-history model outperforms both the tracking-like and the bimodal model showing the benefit of accounting for the cyclical patterns.

5 Summary and concluding remarks

In this work, we addressed the problem of location density estimation over time using location data collected by smartphone apps. Specifically, smartphone loca-

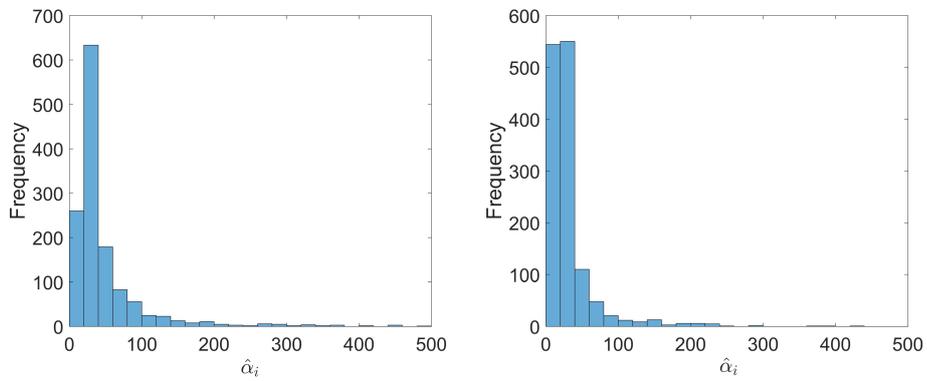


Fig. 3. Estimates of α_i for all smartphones when f_t is the tracking-like density (left panel) and the bimodal density (right panel), respectively.

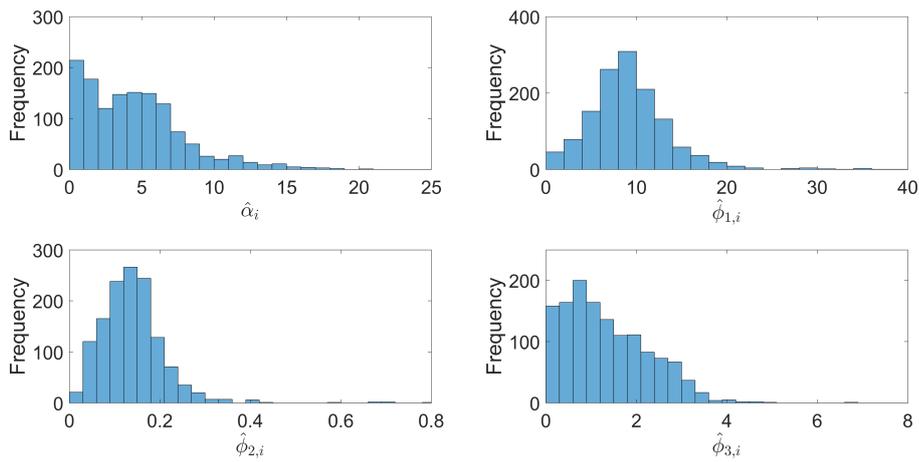


Fig. 4. Parameter estimates for all smartphones when f_t is a full-history mixture density.

Table 1. Percentage of times that a model (row) is better than the other (column) when the comparison is based on AIC.

	tracking-like	bimodal	full-history
tracking-like		0.67%	0.52%
bimodal	98.95%		2.40%
full-history	99.10%	97.60%	

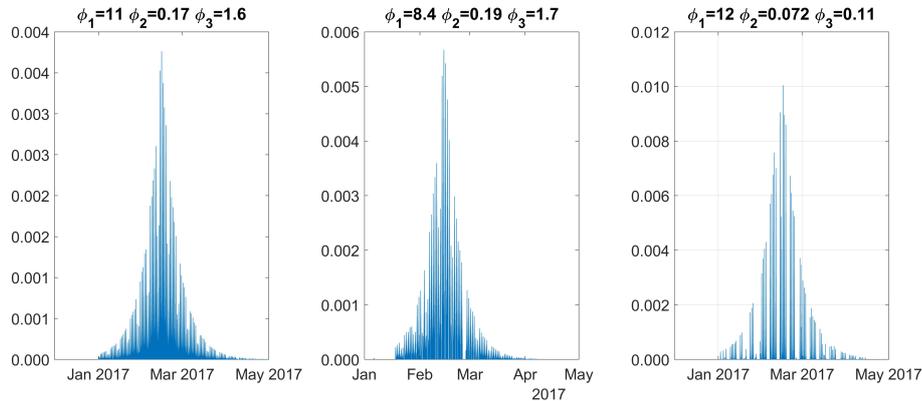


Fig. 5. Estimates of weights $v(t, t'; \hat{\phi})$ when t is February 14th, 2017 12:00:00 for the smartphones shown in Figure 1.

tions and associated precisions are jointly used to estimate a spatial density for the smartphone location at any given time. The approaches are particularly suitable when the smartphone location is not observed at high temporal frequency and the sampling intervals are irregular, potentially with gaps of days or weeks. This is the case of smartphone locations collected by smartphone apps which make use of geolocation but the primary role of which is not tracking. The approaches are flexible and can be applied to analyze location data collected from any location-based app, including social networks.

Computationally, the model estimation time is linear in the number of observed locations and it is feasible on a laptop computer up to 10'000 locations. On the other hand, location prediction at any given point in time is almost real-time given the estimated model. The analysis is carried out using a MATLAB code.

References

1. Zorn, S., Rose, R., Goetz, A., Weigel, R.: A novel technique for mobile phone localization for search and rescue applications. In: 2010 International Conference on Indoor Positioning and Indoor Navigation. (2010) 1–4, DOI: 10.1109/IPIN.2010.5647107
2. Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S.R.H., Ratti, C.: Exposure track: The impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental Science & Technology* **50** (2016) 9671–9681
3. Do, T.M.T., Gatica-Perez, D.: Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* **12** (2014) 79 – 91
4. Tremblay, Y., Shaffer, S.A., Fowler, S.L., Kuhn, C.E., McDonald, B.I., Weise, M.J., Bost, C.A., Weimerskirch, H., Crocker, D.E., Goebel, M.E., Costa, D.P.: Inter-

- polation of animal tracking data in a fluid environment. *Journal of Experimental Biology* **209** (2005) 128–140
5. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Building personal maps from GPS data. *Annals of the New York Academy of Sciences* **1093** (2006) 249–265
 6. González, M.C., Hidalgo, C.A., Barabási, A.: Understanding individual human mobility patterns. *Nature* **453** (2008) 779–782
 7. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. *Science* **327** (2010) 1018–1021
 8. Scellato, S., Musolesi, M., Mascolo, C., Latora, V., Campbell, A.T.: Nextplace: A spatio-temporal prediction framework for pervasive systems. In Lyons, K., Hightower, J., Huang, E.M., eds.: *Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings.* Springer, Berlin, Heidelberg (2011) 152–169
 9. Secchi, P., Vantini, S., Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods & Applications* **24** (2015) 279–300
 10. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM* (2011) 1082–1090
 11. Lichman, M., Smyth, P.: Modeling human location data with mixtures of kernel densities. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '14, New York, NY, USA* (2014) 35–44
 12. Lawlor, S., Rabbat, M.G.: Time-varying mixtures of Markov chains: An application to road traffic modeling. *Signal Processing IEEE Transactions* **65** (2017) 3152–3167
 13. Paci, L., Finazzi, F.: Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, **28** (2017) 359–374
 14. Cressie, N., Kornak, J.: Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* **18** (2003) 436–456
 15. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC, Boca Raton (2006)
 16. Finazzi, F.: The Earthquake Network project: Toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America* **106** (2016) 1088–1099

Forecasting of Multiple Yield Curves Based on Machine Learning

Christoph Gerhart^a, Eva Lütkebohmert^{a,*}, Marc Weber^b

^a*Department of Quantitative Finance, Institute for Economic Research, University of Freiburg,
Platz der Alten Synagoge 1, KG II, 79098 Freiburg i. Br., Germany*

^b*Department for Mathematical Stochastics, Mathematical Institute, University of Freiburg,
Ernst-Zermelo-Straße 1, 79104 Freiburg i. Br., Germany*

Abstract

In this paper we develop robust methods for forecasting term structures of interest rates. We implement a deep long short-term memory (LSTM) neural network based on keras. Our input data is based on the bootstrapped bid, mid and ask multiple (tenor-dependent) yield curves reflecting different risk categories over the period 2005-2018. Furthermore, we use the bid-ask spreads as an additional input factor modelling the market depth. Since there is only a limited amount of data available there is a lack of a sufficiently large training data set. We cope with that difficulty by generating data based on fitted time series models in order to enlarge the training data. We also apply support vector machines to predict trends in the term structures. For this approach we include different market variables to investigate the relationship of these quantities to future yields.

Keywords: forecasting of yield curves, multiple term structures, machine learning, neural networks, support vector machines

JEL: G1, E4, C5, C3

1. Introduction

The term structure of interest rates (or yield curve), which describes the interest rate as a function of maturity, represents an important tool for derivative pricing, risk management and monetary policy and hence has been intensively studied in the literature. In the last years following the financial crisis 2007/2008, however, there has been a major change in interest rate markets. While interest rates of the same maturity showed certain consistencies before the crisis, giving rise

*Corresponding author

Email address: eva.luetkebohmert@finance.uni-freiburg.de (Eva Lütkebohmert)

to a single term structure, this no longer holds in post-crisis markets. Instead e.g. rates on swaps with the same maturity can differ quite substantially depending on the tenor of the underlying reference rate. Thus, in the post-crisis setting the term structure of interest rates becomes tenor-dependent reflecting different risk categories.

In this paper we take these new characteristics of interest rate markets into account and develop robust methods for forecasting of multiple (tenor-dependent) yield curves based on neural networks. Our input data consists of bootstrapped bid, mid and ask yields over the time period 2005-2018. More specifically, we consider the historical discount, three-month and six-month yield curves for maturities up to ten years. We implement a deep long short-term memory (LSTM) neural network based on keras to predict future yields. Our results show extremely accurate predictions of 1-day, 1-month, 3-month and 6-month ahead yields across all maturities and curves. In particular, our results point out the importance of cross-tenor dependencies for predicting future yields, a feature which is naturally missing in existing single-curve approaches. Moreover, we show that using bid-ask spreads as additional inputs reflecting a measure of market depth increases the accuracy of yield curve predictions. Additionally we apply support vector machines (SVMs) to predict trends in the term structure over a time horizon of up to 15 weeks. Here our results for the SVM classifier show a very accurate and extremely robust performance compared to various classification benchmark methods.

2. Data

Our data set consists of daily bootstrapped multiple yield curves over the period 2005-2018. More specifically, we consider historical discount, three-month and six-month curves for maturities up to ten years. The discount curves have been bootstrapped from market data on overnight indexed swap (OIS) rates for maturities ranging from one week to 10 years.¹ For the construction of the risky (tenor-dependent) yield curves, market quotes of deposit rates, rates from forward rate agreements (FRA) and swap rates were used. More explicitly, the short-end is constructed from deposit rates for contracts where the tenor agrees with the maturity, i.e. 3 month deposits for the 3 month curve etc. Besides, FRAs for periods starting in one month ending in 4 months are used

¹The European market publishes the swap rate of an OIS at every business date for maturities ranging from 1 week to 60 years. The floating leg is indexed on the EONIA rate and the payments are based on annual frequency.

for the construction of the 3 month curve while the 6 month curves are constructed from market data on FRAs for time periods [1m, 7m] and [3m, 9m]. Furthermore, the medium to long-term rates are derived from market data on tenor-dependent interest rate swaps with maturities ranging from 6 months up to 10 years. For each term structure we derived the bid, mid and ask curves from the corresponding quotes of the bootstrapping instruments. We use European market data provided by Bloomberg for the time period from September, 2005, until May, 2018, on a daily basis. The (tenor-dependent) term structures have been constructed by an exact fit bootstrapping methodology as explained in Gerhart and Lütkebohmert (2018). Additionally, the term structures were fitted to the parametric model of Nelson and Siegel (1987). Below we forecast yields over different time horizons and we work with the convention, that one week consists of 5 trading days and one months of 21 trading days.

3. Forecasting of Multiple Yield Curves via Neural Networks

In order to predict future yields of various maturities and tenors and over different forecasting horizons we implemented a neural network which we present in the following subsection. Thereafter we discuss how we dealt with the problem of limited available data, present our forecasting results and elaborate on alternative approaches.

3.1. Neural Network Design and Methodology

The neural network is based on the `keras` package with `tensorflow` backend. The input layer consists of the mid, bid and ask yields of a specific maturity, the corresponding bid-ask spreads as well as the basis spreads calculated as the difference between the mid values of the risky tenor-dependent yields and the risk-free discount yields for the specified maturity. The output consists of the yields themselves. As we follow a multiple curve approach, this results in a rather large input and output dimension. We use the eminent relationship between the risky and the discount curve in order to fine tune the training process of the neural network. Furthermore, the bid-ask spread is included in order to teach the network about market depth.

We ran a hyperparameter optimization code using the `R` package `flags` which searches for the best hyperparameter combination on a discrete grid. Afterwards we also implemented the whole code in `Python` and used the hyperparameter optimization procedure based on the `hyperas` package, which is more advanced. For a deeper insight on grid searches versus random searches we refer to

Bergstra and Bengio (2012). The network architecture based on the hyperparameter optimization consists of three layers and two dropout layers in between. The first layer is an LSTM layer with 128 neurons. The next layers are given by a second LSTM layer with 32 neurons and a third dense layer. The drop out rates are set to 0.195327 and 0.056875 while the learning rate is chosen as $8.777485 \cdot 10^{-4}$ and is allowed to decay over time.

We chose 3 neurons in the output layer since we analyse only one maturity at a time for three different curves (discount, three-month and six-month curve) and the last layer is supposed to directly correspond to the number of outputs. For a deeper insight on Long-Short-Term-Memory neural networks we refer the reader to Hochreiter and Schmidhuber (1997) and emphasize the ability of store and forget important respectively redundant information of those layers.

The LSTM layer in keras needs a 3-dimensional input shape. It is therefore necessary to reshape the given dataset in tensor form. The first dimension corresponds to the number of historical dates used as a training, resp. test data set. The last dimension agrees with the number of input, resp. output parameters. In our case this amounts to 14 parameters – one fixed maturity times three curves and three different values (mid, ask, bid) already yields 9 dimensions. When adding three bid-ask spreads and 2 basis spreads, one ends up with 14 input parameters. The second dimension specifies how much data is used to be fitted to one output unit. Here we chose chunks of 300 data points meaning that the last 300 values are used to find a connection to the next value of interest. For monthly forecasts this would e.g. be the 322nd value.

Hence, the input shape in the LSTM layer needs to be of the form `c(300,14)` in our case. As explained above, a three dimensional array needs to be passed to the network. As one can see in the keras documentation, the three dimensions coincide with (`batch_size`, `sequence_length`, `features`). The first dimension does not need to be passed along. It only defines how long the training process is going to be. The reshaping procedure of the two dimensional data is accomplished via a loop. Starting with the matrix consisting of the first 300×14 data points, one overwrites the first entry in the first dimension of a dummy tensor. In the next step, one starts at the second value and proceeds to the 301st value. The procedure is repeated over the whole training data set. The output data then consists of one-day-ahead forecasts. If more than one forecast shall be constructed with each 300 data point chunk, then the output needs to be reshaped correspondingly.

As activation function in the first layers we chose the tangent hyperbolic as the sigmoid function produced larger errors. In the output layer we use the linear activation function. Usually a relu

function is chosen in the last layer. However, that does not correspond well with our chosen normalizing procedure as we transformed the data set into one which values range from -1 to 1. Therefore, a relu function would just cut off roughly half of the values and set them to 0. A linear activation is the natural choice in that context.

Based on the above specified architecture of the network, we predict yields of different maturities and tenors. As in Gerhart and Lütkebohmert (2018) we use the root-mean-squared-error as loss function.² The optimizer chosen in our model is the Adam optimizer (see Kingma and Ba (2014)). Furthermore we chose a validation split of 0.2 and a batch size of 32. The calculation is done via 100 epochs.

3.2. Dealing with Limited Data

As there is only a limited data set available, we face the problem that running a small amount of epochs yields bad results but choosing a large amount of epochs often results in over-fitting. We cope with that problem by generating new data based on the historical yields via the following procedure. We use the first 500 historical data points of the training set on the 9 yields (bid, mid and ask yields for three tenors and fixed maturity) and fit an ARIMA(1,0,1) model to these time series. Based on the time series model, we then predict intervals for the next ten values and save the interval borders. Then we proceed with the points eleven up to 510 to predict the next ten intervals and so on. We end up with 1000 confidence intervals in which our yields are likely to lie in. In this way, we can easily generate new paths of the 9 time series available by randomly choosing values in these intervals. Now, one has an arbitrary large training data set to which we can fit the neural network. We still have to cope with the risk of over-fitting, though, since the generated paths are highly dependent on the true evolution.

3.3. Results

The neural network performs very well with highly accurate predictions of future yields across all tenors and maturities over various forecasting horizons. In Tables 1 – 4 we list the results of that approach for 1-day, 1-month, 3-month and 6-month ahead predictions. The first row in each table shows the RMSEs of the predicted yields for the discount, 3-months and 6-months curve for

²We did not use a pre-implemented loss of keras as most of these functions are applicable only for categorization problems and hence fail to be of any value for our application.

maturities in 1 year and 3 years, resp, over the specified forecasting horizons for the neural network (NN). The second row lists the corresponding RMSE for the random walk (RW) approach. Rows 3 and 4 show the correlation between the predicted values and the actual values which indicates the robustness of a forecasting method.

Our results show that the neural network produces extremely accurate predicitions of future yields across all maturities and curves and for all forecasting horizons. In particular, for longer term predictions as 3- or 6-month ahead forecasts, our results clearly outperform the benchmark random walk approach which has been proven to be hard to beat in the related literature. For 1-month ahead predictions, the neural network still beats the random walk procedure for yields with longer maturities. For very short-term predictions such as 1-day ahead forecasts, the random walk approach performs better than the neural network. However, it should be mentioned here that the RMSEs of both methods are extremely small for all maturities and curves in this case so that both approaches actually produce very precise forecasts of next day yields. The correlation values verify the robustness of our forecasting methodology.

Method	Discount - 1y	3m - 1y	6m - 1y	Discount - 3y	3m - 3y	6m - 3y
NN	0.002106980	0.002221245	0.002320099	0.002087555	0.002096353	0.002189963
RW	0.002455532	0.002646123	0.002898431	0.002351593	0.002598185	0.002784662
Correlation NN	0.873266651	0.879881245	0.880366621	0.879653312	0.874569823	0.890000451
Correlation RW	0.851333642	0.861110326	0.871005612	0.862343540	0.864706349	0.876448767

Table 1: Comparison of neural network six month forecast results with random walk results as well as the correlation of the forecasts to the true values.

Method	Discount - 1y	3m - 1y	6m - 1y	Discount - 3y	3m - 3y	6m - 3y
NN	0.001256644	0.001198885	0.001400062	0.001075632	0.001101326	0.001178669
RW	0.001344568	0.001399984	0.001506844	0.001266431	0.001395539	0.001470158
Correlation NN	0.938855486	0.93713645	0.93941100	0.94000325	0.93972566	0.93910442
Correlation RW	0.93225647	0.93089991	0.93222241	0.93772375	0.93596972	0.93643050

Table 2: Comparison of neural network three month forecast results with random walk results as well as the correlation of the forecasts to the true values.

Method	Discount - 1y	3m - 1y	6m - 1y	Discount - 3y	3m - 3y	6m - 3y
NN	0.000745689	0.000678933	0.000580112	0.000889645	0.000898752	0.000911564
RW	0.000587419	0.000582809	0.000583411	0.000913224	0.000955568	0.000990582
Correlation NN	0.841000726	0.866559841	0.880076524	0.902354446	0.920068596	0.919167822
Correlation RW	0.838753128	0.866618380	0.877812199	0.894634916	0.908987169	0.909525220

Table 3: Comparison of neural network one month forecast results with random walk results as well as the correlation of the forecasts to the true values.

Method	Discount - 1y	3m - 1y	6m - 1y	Discount - 3y	3m - 3y	6m - 3y
NN	0.0003088099	0.00017444	0.000386521	0.000333298	0.000376663	0.000361211
RW	0.000215763	0.000225456	0.000215041	0.000215763	0.000225456	0.000215041
Correlation NN	0.988765231	0.990015202	0.991645551	0.989822135	0.990041124	0.990117699
Correlation RW	0.991047061	0.993272731	0.994688241	0.992894395	0.993695834	0.994688241

Table 4: Comparison of neural network one day forecast results with random walk results as well as the correlation of the forecasts to the true values.

We also tested several other forecasting approaches which we want to discuss here without stating the results explicitly in order to keep the exposition compact. These include a version, where yields of various maturities are turned over to the network as input. However, that approach failed to perform well. An explanation for this observation is the high dimensionality of the input and output data in proportion to the number of historical dates in that case. As it is explained in Friedman, Hastie and Tibshirani (2017) a large number of input dimensions corresponds rather poorly with small data sets. The latter limits the number of training runs such that the neural network is not able to learn which data has to be considered important and which redundant in view of every single output dimension.

Besides, we implemented an alternative network based on a parameterized plug-in model. For a given training set, we extract the Nelson-Siegel yield curve parameters β_0, β_1 and β_2 and assume these to follow a certain time series model. As input we provide the three parameters for every curve and mid, bid and ask values which results in a total of 27 input parameters. After training the network, we build forecasts for the Nelson-Siegel parameters which are then used to compute future yields via the Nelson-Siegel model. Our results showed that this did not yield any improvements over the approach which is directly based on the yields.

Furthermore, we have included different macroeconomic variables such as the prime interest rate, inflation rates and GDP growth rates, in the input layer as these have been shown to improve predictions in standard time series based approaches. In our neural network this, however, produces less precise forecasts. An explanation for this might be the seldom updates of those values. While macroeconomic variables might be powerful indicators of level shifts in time series whenever those values are updated, they represent only an extra dimension in the network without any additional use between these updating time points. In contrast, our results show that including the bid, mid and ask prices really improves the results. The information which lies in the spread obviously corresponds to the evolution of the time series which is well detected by the artificial intelligence.

4. Trend Forecasting via Support Vector Machines

In this section, we investigate the relationship between the different movements of the term structures and various market variables by using Support Vector Machines (SVMs). This type of machine learning classifier can be applied to forecast the up and down trends of the levels as well as the slopes for the multiple term structures of interest rates. For a detailed treatment of the subject of SVMs we refer to Cortes and Vapnik (1995), Burges (1998) and Evgeniou et al. (2000). Furthermore, an introductory discussion of SVMs is given by Bennett and Campbell (2000) and a more rigorous study is stated by Steinwart and Christmann (2008). SVMs have been successfully applied for trend predictions in financial markets. For instance, Huang et al. (2005) analyse the accuracy of stock market movement prediction with SVMs by comparing their performance with classical forecasting classification models. Their SVM-based models outperform the other classification methods in their empirical analysis. In the following, we apply support vector machines to predict level and slope trends in term structures of interest rates.

4.1. Model Input Selection

We consider the bootstrapped discount, three-month and six-month yield curves for the period of September 2005 to May 2016. For each term structure we derived the bid, mid and ask curves up to a maturity of 10 years. The term structures are fitted to the parametric model of Nelson and Siegel (1987). Consequently, we obtain a time series of estimated parameters of level $\beta_{0,t}^{q,k}$, slope $\beta_{1,t}^{q,k}$, curvature $\beta_{2,t}^{q,k}$ and exponentially decaying rate $\lambda_t^{q,k}$ of the loadings where t represents the date, $q \in \{\text{bid, mid, ask}\}$ and $k \in \{\text{d, 3m, 6m}\}$. From the weekly changes of the estimated level and slope

parameters, we derive the categorical variables indicating the up or down movement of the yield curve shape parameters. As explanatory variables we use the fitted level and slope parameters of each curve, the changes in the level and slope of the preceding period, the ratio between the bid-ask spread to mid bond prices for the time-to-maturity of 10 years and the Euro STOXX 50. The ratio is used as measure for the long-term liquidity in the market. All the raw data is provided by Bloomberg and corresponds to a weekly frequency (last business day in the week). The prediction models can then be expressed in form of

$$\text{level}_t = F(\Lambda_{t-1}^l) \quad \text{and} \quad \text{slope}_t = G(\Gamma_{t-1}^s)$$

where

$$\text{level}_t = \begin{pmatrix} \text{level}_t^{\text{d,ask}} \\ \text{level}_t^{\text{d,mid}} \\ \text{level}_t^{\text{d,bid}} \\ \text{level}_t^{\text{3m,ask}} \\ \text{level}_t^{\text{3m,mid}} \\ \text{level}_t^{\text{3m,bid}} \\ \text{level}_t^{\text{6m,ask}} \\ \text{level}_t^{\text{6m,mid}} \\ \text{level}_t^{\text{6m,bid}} \end{pmatrix}, \quad \text{slope}_t = \begin{pmatrix} \text{slope}_t^{\text{d,ask}} \\ \text{slope}_t^{\text{d,mid}} \\ \text{slope}_t^{\text{d,bid}} \\ \text{slope}_t^{\text{3m,ask}} \\ \text{slope}_t^{\text{3m,mid}} \\ \text{slope}_t^{\text{3m,bid}} \\ \text{slope}_t^{\text{6m,ask}} \\ \text{slope}_t^{\text{6m,mid}} \\ \text{slope}_t^{\text{6m,bid}} \end{pmatrix} \quad \Lambda_{t-1}^l = \begin{pmatrix} \Lambda_{t-1}^{\text{d,ask}} \\ \Lambda_{t-1}^{\text{d,mid}} \\ \Lambda_{t-1}^{\text{d,bid}} \\ \Lambda_{t-1}^{\text{3m,ask}} \\ \Lambda_{t-1}^{\text{3m,mid}} \\ \Lambda_{t-1}^{\text{3m,bid}} \\ \Lambda_{t-1}^{\text{6m,ask}} \\ \Lambda_{t-1}^{\text{6m,mid}} \\ \Lambda_{t-1}^{\text{6m,bid}} \end{pmatrix}, \quad \Gamma_{t-1}^s = \begin{pmatrix} \Gamma_{t-1}^{\text{d,ask}} \\ \Gamma_{t-1}^{\text{d,mid}} \\ \Gamma_{t-1}^{\text{d,bid}} \\ \Gamma_{t-1}^{\text{3m,ask}} \\ \Gamma_{t-1}^{\text{3m,mid}} \\ \Gamma_{t-1}^{\text{3m,bid}} \\ \Gamma_{t-1}^{\text{6m,ask}} \\ \Gamma_{t-1}^{\text{6m,mid}} \\ \Gamma_{t-1}^{\text{6m,bid}} \end{pmatrix}$$

with

$$\Lambda_{t-1}^{k,q} = \begin{pmatrix} \beta_{0,t-1}^{k,q} \\ \beta_{0,t-1}^{k,q} - \beta_{0,t-2}^{k,q} \\ \frac{B^{\text{bid},k}(t-1,10) - B^{\text{ask},k}(t-1,10)}{B^{\text{mid},k}(t-1,10)} \\ S_{t-1}^{\text{SXX50}} \end{pmatrix}^{\top}, \quad \Gamma_{t-1}^{k,q} = \begin{pmatrix} \beta_{1,t-1}^{k,q} \\ \beta_{1,t-1}^{k,q} - \beta_{1,t-2}^{k,q} \\ \frac{B^{\text{bid},k}(t-1,10) - B^{\text{ask},k}(t-1,10)}{B^{\text{mid},k}(t-1,10)} \\ S_{t-1}^{\text{SXX50}} \end{pmatrix}^{\top}.$$

As mentioned above $\text{level}_t^{k,q}$ and $\text{slope}_t^{k,q}$ are categorical variables indicating up and down movements for curve k and type of quote q .

4.2. Model Design

The input data consists of weekly quotes from September 2005 to May 2016. As training set we choose the first 150 weekly model quantities ranging from 2nd September 2005 to 12th February

2016. The test set contains 15 weeks belonging to the period 19th February 2016 to 27th May 2016. The implemented SVM for the level predictions is based on bound-constraint formulation of the classification. It is equipped with the Gaussian Radial Basis kernel function where the kernel parameter is set to $\sigma = 0.01965$. The constant C of the regularization term in the Lagrange formulation, known as cost of constraints violation, is set to be $C = 3$. Furthermore, a 3-fold cross validation on the training data is performed to assess the quality of the model performance. This type of classifiers also supports class-probabilities output. For the forecasting of the slope movements we use a similar design of the SVM with parameters $\sigma = 0.06555$, $C = 25$ and 3-fold cross validation. The parameters are obtained by using a hyperparameter optimization procedure.

4.3. Forecasting Results

Typically the evaluation of the performance for classifiers is stated in form of accuracy measures or their comparison to other classification benchmark methods. A comprehensive overview on useful and common accuracy measures based on the confusion matrix can be found in Friedman, Hastie and Tibshirani (2017) and Lantz (2015).

Method	+class	Accuracy	Error Rate	Kappa	Sensitivity	Specificity	Precision	F-score
SVM	up	0.6593	0.3407	0.3164	0.7246	0.5909	0.6494	0.6849
LDA	up	0.6519	0.3481	0.2971	0.8551	0.4394	0.6146	0.7152
QDA	up	0.6519	0.3481	0.2999	0.7681	0.5303	0.6310	0.6928
RW	up	0.4254	0.5746	-0.1498	0.4412	0.4091	0.4348	0.4380
SVM	down	0.6593	0.3407	0.3164	0.5909	0.7246	0.6724	0.6290
LDA	down	0.6519	0.3481	0.2971	0.4394	0.8551	0.7436	0.5524
QDA	down	0.6519	0.3481	0.2999	0.5303	0.7681	0.6863	0.5983
RW	down	0.4254	0.5746	-0.1498	0.4091	0.4412	0.4154	0.4122

Table 5: Comparison of model performance with positive class *up-movement* as well a positive class *down-movement* for level trends.

Table 5 shows the performance of our SVM approach for the prediction of up- and down movements in the level of interest rates and compares these with the corresponding values for the benchmark random walk model (RW), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).³ Table 6 illustrates the corresponding results for up- and down movements in the

³Compare Huang et al. (2005) for details on these approaches.

Method	+class	Accuracy	Error Rate	Kappa	Sensitivity	Specificity	Precision	F-score
SVM	up	0.6000	0.4000	0.2059	0.7077	0.5000	0.5679	0.6301
LDA	up	0.5852	0.4148	0.1783	0.7231	0.4571	0.5529	0.6267
QDA	up	0.5630	0.4370	0.1412	0.8154	0.3286	0.5300	0.6424
RW	up	0.4104	0.5896	-0.1807	0.3846	0.4348	0.3906	0.3876
SVM	down	0.6000	0.4000	0.2059	0.5000	0.7077	0.6481	0.5645
LDA	down	0.5852	0.4148	0.1783	0.4571	0.7231	0.6400	0.5333
QDA	down	0.5630	0.4370	0.1412	0.3286	0.8154	0.6571	0.4381
RW	down	0.4104	0.5896	-0.1807	0.4348	0.3846	0.4286	0.4317

Table 6: Comparison of model performance with positive class *up-movement* as well a positive class *down-movement* for slope trends.

slope of the yield curves. The results indicate that our SVM classifier performs extremely robust across various classification benchmark methods and clearly outperforms them in most cases.

5. Conclusion

In this paper we have developed robust methods to forecast interest rates of different maturities and tenors based on neural networks. Our results show extremely accurate predictions, in particular for longer term yields. Besides we have applied SVMs to forecast trends in the yield curve shape parameters. Compared to other classification methods, our SVM approach performs very well and robust. While the linear discriminant analysis also yields good results for level trends, it performs less well for slope predictions. In contrast, the quadratic discriminant analysis works well for slope trends but less good for level trends. The suggest SVM approach, however, is very stable and provides good trend predictions across the different yield curve movements.

References

- Bennett, K.P., Campbell, C., 2000. Support vector machines: hype or hallelujah? ACM SIGKDD Explorations Newsletter 2, 1–13.
- Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13, 281–305.

- Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121–167.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine Learning* 20, 273–297.
- Evgeniou, T., Pontil, M., Poggio, T., 2000. Regularization networks and support vector machines. *Advances in Computational Mathematics* 13, 1–50.
- Friedman, J., Hastie, T., Tibshirani, R., 2017. *The elements of statistical learning. volume 2.* Springer.
- Gerhart, C., Lütkebohmert, E., 2018. *Empirical Analysis and Forecasting of Multiple Yield Curves.* Preprint, University of Freiburg .
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Computation* 9, 1735–1780.
- Huang, W., Nakamori, Y., Wang, S.Y., 2005. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research* 32, 2513–2522.
- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations* .
- Lantz, B., 2015. *Machine learning with R. 2nd ed.,* Packt Publishing Ltd.
- Nelson, C., Siegel, A., 1987. Parsimonious modelling of yield curves. *Journal of Business* 60, 473–489.
- Steinwart, I., Christmann, A., 2008. *Support vector machines.* Springer Science & Business Media.

Empirical evaluation of advanced oversampling methods for improving bankruptcy prediction

W. Alswiti¹, H. Faris¹, H. Aljawazneh², S. Safi², P.A. Castillo²,
A.M. Mora³, R. Abukhurma¹, and H. Alsawalqah¹

¹ King Abdullah II School for Information Technology
The University of Jordan, Amman, Jordan

² Department of Computer Architecture and Technology,
ETSIT and CITIC, University of Granada, Spain

³ Department of Signal Theory, Telematics and Communications,
ETSIT and CITIC, University of Granada, Spain

Abstract. The relevance of bankruptcy prediction problem is evident in today's world due to its effects on banks, businesses, and companies. There might be huge financial losses encountered due to bad judgment and analysis. Thus, in order to help improving the quality of such tasks, much efforts has been invested on building prediction models for aiding the decision makers to anticipate the events before they take place. However, developing an accurate bankruptcy prediction model is a challenging task due to the usual high imbalanced distribution in data, where the number of insolvent companies are much less than those successful ones. This makes it very difficult to create an accurate model to classify or forecast healthy companies from bankrupt ones. In this work, we try to enhance the predictive ability of bankruptcy models by tackling the imbalanced distribution problem. Specifically, we focus on preprocessing the data to reduce their imbalance ratio by applying and comparing eleven advanced resampling methods. After the data balancing has been performed, we use C4.5 classifier to generate the decision trees to predict bankruptcy. We have obtained significant enhancement regarding the considered evaluation measurements, including Type II and Type I errors. The latter is decreased sometimes up to a 65%, maintaining an accuracy close to 90%. This helps in reducing the misclassification of positive instances, as they are considered the most important risk factor. We have also studied the complexity of the resulting decision trees after applying each oversampling technique, to evaluate their utility as a decision-aid tool for a human expert. Moreover, we have analyzed the importance of the features for the top oversampling methods. SMOTE with Edited Nearest Neighbor (ENN) performs the best in both type I error and complexity.

Keywords: Bankruptcy, Oversampling, Imbalanced data, Classification, Prediction

1 Introduction

Bankruptcy is a formal insolvency proceeding involving either an individual or a company who have announced their inability to pay the outstanding debts. As a result for this legal status, the debtor's assets are liquidated in order to repay part of the distressed debts, and the remaining portion is discarded [1]. For this reason, the prediction of bankruptcy is a major concern for different financial personnel such as managers, stakeholders, creditors, investors and others who may be affected by the consequences of the financial failure [2].

A successful forecasting for this problem will give a broader perspective about the healthy situation of the business and help the decision makers to predict the events before they take place. For these reasons, there has been an intensive effort in the literature for developing statistical and artificial intelligence-based models to accurately forecast the financial status of the companies. Generally, from machine learning perspective, the prior judging for the company's status either it is bankrupt or non-bankrupt is considered as a binary classification problem.

Developing robust and accurate model for bankruptcy prediction is a complex task. There are so many challenges and difficulties that may negatively impact the model building process and badly affect the generalization performance. Among these challenges are the large number of variables that need to be studied, the missing or unavailable information, or the non-stationary nature of the bankruptcy conditions, which explains the difficulty of adapting a single time period and the necessity to study the company's conditions during multiple periods to reduce the time sensitivity of the model. Moreover, from a machine learning point of view, bankruptcy datasets are imbalanced by nature. This means that the classes are unevenly distributed where the bankrupt class has rare occurrences compared with the normal (i.e non-bankrupt) class. The essence of the problem is that standard classification algorithms deal with both classes as if they would have the same importance, where in fact, the bankrupt class should have more attention for getting more successful classification results. According to this consideration the generated models are biased towards the majority class, whereas the minority class will be neglected in most cases [3,4]. Hence the model predictive power will be declined and the obtained results may not be reliable anymore.

In literature, different approaches for handling the imbalanced class distribution in datasets have been proposed. Among them, the external approach is the most frequently used. In it, the distribution of the class labels is altered by increasing the minority class patterns (i.e Oversampling) or by decreasing the majority class ones (i.e Undersampling). For bankruptcy prediction, most of the previous works that followed oversampling approaches focused on few methods which are the simple Random Oversampling (ROS) method [5] or the common Synthetic Minority Over-sampling Technique (SMOTE) [6].

In this work, we empirically evaluate and compare 9 advanced oversampling methods in addition the aforementioned two, for the bankruptcy prediction. A real dataset collected from the Spanish market is utilized for this study. The

dataset is considered very challenging due to the highly imbalanced distribution where the bankrupt cases form only a 2% of the whole sample.

This study comes as an improvement of our previous work [7],

which applied three simple data resampling methods: Oversampling, Under-sampling and a Hybrid approach in order to improve the performance of J48, Random forest and Naïve Bayes classifiers to predict the financial states of companies using the same dataset considered here.

2 State of the art

Several researchers have considered the bankruptcy prediction as a classification problem. While the financial datasets are not constantly balanced, the researchers solved this problem using data balancing approaches at a preprocessing stage. Thus, in this research line, García et al.[8] addressed a performance and efficiency comparison between several undersampling and oversampling approaches applied before classifying several imbalanced credit datasets. Then they applied four classification methods widely used for credit risk prediction, (KNN, MLP, RBF and SVM). Five real-world credit datasets with various balancing ratios had been considered to evaluate the resampling strategies and classification. This study shows that SMOTE with the Edited Nearest Neighbor (SMOTE + ENN) algorithm together with SVM classifier obtains the best results.

Kim et al. [9] solved the problem of predicting the bankruptcy situation from an imbalanced dataset. They proposed a method named geometric mean-based boosting algorithm (GMBoost), which is a modification of AdaBoost algorithm. The authors considered a dataset collected from a Korean commercial bank. Thus, after they split the dataset into 5 subsets, they compared the obtained results with and without using SMOTE. GMBoost yielded the best performance regarding the prediction and learning capabilities in the comparison with AdaBoost and cost-sensitive boosting.

Zięba et al. [10] proposed a novel approach for bankruptcy prediction which applies Extreme Gradient Boosting for learning an ensemble of decision trees. In addition, they introduce a new concept named *synthetic feature* in order to obtain higher-order statistics in data. The dataset has been used to evaluate this approach is Polish companies financial states collected during 2007 to 2013 for bankrupt companies, and 2000 to 2012 for still operating ones. The approach proposed by the authors obtained better results with respect to the referenced methods they applied such as J48, RF, SVM and AdaBoost.

Barboza et al.[11] compared the results of applying several classification methods on a validation set which was extremely imbalanced. It contains 13300 records for solvent companies and 133 records for the bankrupt ones. The authors proved that Boosting, Bagging and RF obtained the best result with relatively little difference between their outcomes.

Le et al.[12] utilized several balancing methods such SMOTE, ADASYN and SMOTE-ENN in order to also handle the problem of classifying extremely imbalanced datasets. They applied several classifiers, such as: Random Forest, De-

cision Trees, MLP and SVM. The extremely imbalanced dataset used by the authors was collected from a Korean financial companies and consists of 307 bankrupt companies and 120048 solvent ones. The results of this study proved that oversampling techniques can improve the performance of bankruptcy prediction. SMOTE-ENN used as preprocessing stage for RF yielded the best results regarding the Area Under the Curve (AUC) measurement.

The present work stands for studying the same kind of problem, but with a different and real dataset regarding Spanish companies, which is also extremely imbalanced (as described in Section 3). This is an enhancement on our previous work [7], as we study here 11 more advanced resampling methods to perform the dataset balancing. In addition, as far as we know, some of the balancing approaches compared in this study such as ADOMS, SPIDER, SPIDER2 and AHC were not considered before by other researchers in the literature, regarding the bankruptcy prediction problem.

3 Dataset description

The bankruptcy prediction problem addressed in this paper has been focused on Spanish companies, from which we have considered several financial and non-financial features. This is faced as a classification problem, considering as class for each sample, the dependent variable, *Bankruptcy*.

The dataset has been extracted from the Infotel database⁴. This company has been devoted to gather diverse information about many different companies in Spain along several years, including their financial operations and results. Thus, we have data from 470 companies, which were gathered during six consecutive years (from 1998 to 2003). There are 2860 patterns, from which 62 correspond to bankrupt enterprises.

After removing useless variables, such as internal codes, every sample has 33 independent variables, including qualitative and quantitative information, i.e. categorical and numerical values; and being some of them financial indicators and the rest non-financial ones.

So there are 2859 records with 33 variables each, 26 of them numerical and 7 non-numerical. Each record corresponds to a company in an exercise and the dependent variable or class indicates if that company was bankrupt at the end of that year. The independent variables are:

- Numeric: Financial - *Debt Structure, Debt Cost, Debt Paying Ability, Debt Ratio, Working Capital, Warranty, Operating Income Margin, Return on Operating Assets, Return on Equity, Return on Assets, Stock Turnover, Asset Turnover, Receivable Turnover, Asset Rotation, Financial Solvency, Acid Test.*

Non-Financial - *Year, Number of employees, Age of the company, Number of partners, Number of changes of location, Historic number of Judicial incidences, Number of judicial incidences, Historic amount of money spent on*

⁴ Bought from <http://infotel.es>

judicial incidences, Amount of money spent on Judicial incidences, Historic number of Serious incidences

- Categorical: *Size, Type of company, Province code, Auditor's opinion.*
- Binaries (Yes/No): *Linked to a group, Delay, Audited.*

4 Advanced oversampling methods

In this section we describe advanced oversampling techniques that will be used for handling the problem of the imbalanced distribution in the dataset.

- ROS-I: Random oversampling is a non-heuristic technique, that used to balance an imbalanced data set by increasing the number of minority class members, by simply randomly replicating existing data points. While simple and powerful this method is very sensitive to overfitting [5].
- SMOTE: Synthetic Minority Oversampling Technique, introduces new examples in the training data to enrich the data space and counter the scattered data points in the distribution. It creates new synthetic examples along the line segments joining any or all of the k minority class nearest neighbors, in order to achieve the amount of the oversampling required it selects randomly the k nearest neighbors and then takes the difference between the feature vector (sample) for each minority class sample and its nearest neighbor, and multiply that by a random number that is between 0 and 1, then added it to the feature vector [6].
- SMOTE-TL: Synthetic Minority Oversampling Technique and Tomek's modification of Condensed Nearest Neighbor, is an advanced oversampling algorithm introduced to overcome the overfitting problem and to remove noisy samples lying on the wrong side of the decision border. It applies a data cleaning method - Tomek links -. First, minor class instances are replicated using SMOTE, then Tomek links are identified and removed for both minor and major classes instances. Tomek links can be defined as follows: a pair of examples is considered a Tomek link if both examples are from different classes and they are the closest to each other, i.e given two examples E_i and E_j that belong to different classes we define the distance between them as $d(E_i, E_j)$ while there is no other example E_k such as $d(E_i, E_k) < d(E_i, E_j)$ or $d(E_j, E_k) < d(E_i, E_j)$ then the pair (E_i, E_j) is considered a Tomek links. If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline [5].
- SMOTE-ENN: Synthetic Minority Oversampling Technique and Edited Nearest Neighbor(ENN), after applying SMOTE oversampling method, this method cleans the dataset by removing the noisy instances in order to increase the classifiers' generalization ability. the cleanup process is defined as follows: For each instance E , ENN will find its 3 nearest neighbors ,if E belongs to the majority class and 2 or more of the nearest neighbors are from minority class then E is removed, and visa versa for instances from minority class [5].
- Borderline SMOTE: During the training process most of the classification algorithms tries to learn the borderline of each class. The borderline examples

are the ones that usually misclassified, hence this method focuses on these samples. Works as follows: for each instance E in the minority class it finds its k nearest neighbors, if all of them are from the majority class then this instance is considered noise and ignored. If the number of majority class in the k nearest neighbor is less than the number of instances of the minority class then this instance is considered safe and also ignored. The rest which have more majority class neighbors are considered in danger and these are the instances synthetically replicated. [13].

- Safe-Level SMOTE: Aims to create synthetic instances in safe regions only. Which works as follows: This method assigns a safe level to each minor instance p , the safe level is defined as the number of instances from the minority class in the k nearest neighbor, and then calculates the safe level ratio which is the safe level of p divided by the safe level of the nearest neighbors n . which will result in 5 different cases:
 1. safe level ratio is ∞ and safe level of p is 0, then both p and n are considered noise and ignored.
 2. safe level ratio is ∞ and safe level of p is not zero, which means n is noise so a synthetic instance is created far from n .
 3. safe level ratio is 1, then a synthetic instance is created along the line between p and n .
 4. safe level ratio is greater than 1, then a synthetic instance is created closer to p because obviously p is safer than n .
 5. safe level ratio is less than 1, then a synthetic instance is created closer to n because n is safer than p [14].
- ADASYN-I: Adaptive Synthetic Sampling: uses weights to evaluate minority class instances which are hard to predict. For each instance in the minority class, k nearest neighbors are found. Then the density distribution is calculated by dividing the number of instances in the k nearest neighbors that belong to the majority class by k . This value is a measurement of the distribution of weights for different minority class and it is used to determine the number of needed synthetic samples. This helps to reduce the bias and shifting the classification decision boundary to difficult instances [15].
- ADOMS: Adjusting the Direction of the synthetic Minority class Samples, works by generating synthetic examples to fit well in the actual data distribution of the data set, depending solely on the Principal component analysis technique, which is focuses on the variations and patterns in the data set. Synthetic examples will be created along the first principal component axis (the linear combination of the features that have the maximum variance among all linear combinations) of local data distribution which occupies the maximal amount of total variance in the feature space. Proved to be effective to reduce the drop of the classification performance of the experimental classifier in the class imbalance situations and helps to reduce drawbacks caused by newly generated synthetic samples for the minority class [16].
- SPIDER: Selective preprocessing of imbalanced data, this approach uses the internal characteristic of examples to drive their pre-processing, it classifies examples into two types, safe and noisy. Safe examples will be correctly

classified by a constructed classifier, however any example that is classified as noisy has a very high chance to get misclassified, hence require pre-processing. The example get classified to be safe or noise by applying the NNR with the heterogeneous value distance metric (HVDM). Then, there is three techniques for pre-processing, they all involve modification of the minority class, however, the degree and scope of changes varies:

1. Weak amplification: amplifies the noisy examples from the minority class by adding as many of their copies as there are safe examples from the majority class in their 3 nearest neighbors.
2. weak amplification and relabeling: Extends on the first technique adding a labeling step: noisy examples from the majority class is located in the 3 nearest neighbor of noisy examples in the minority class are relabeled by changing their assignment from majority class to minority class.
3. strong amplification: Applies weak amplification on safe examples from the minority class. But for noisy examples each example is reclassified using an 5 nearest neighbor.

After that this technique pre-process examples according to their type [17].

- SPIDER2: Similar to SPIDER this method distinguish between safe, borderline and noisy examples, and it claims that the distribution of borderline and noisy examples causes difficulties for learning algorithms. The difference between SPIDER and SPIDER2, is that SPIDER2 method applies a two phase pre-processing for the examples from the majority and the minority classes, while SPIDER identifies the nature of the examples and then simultaneously process the minority and the majority examples, which could result in too extensive modifications in some regions of the Majority class examples [18].
- AHC: Agglomerative Hierarchical Clustering is a clustering algorithm that is used in data mining, this algorithm starts with each example as a cluster of one, and then finds min distance between two clusters and merge them. Removes the original clusters merged from the data set and add the resulted cluster, then iteratively repeat the process until reaches the required number of clusters. To overcome the imbalance issue in the data set, those clusters are used as a prototype to create synthetic examples. This techniques helps in increasing sensitivity for different classifiers [19].

5 Evaluation measurements

In order to evaluate the classifier performance together with the oversampling methods, we will refer to the 2X2 confusion matrix which is shown in Table 1.

TP and TN are correctly classified positive and negative instances, misclassified instances are represented by FP and FN [20], and are used to calculate four performance measures Type I error, Type II error, average Geometric Mean and Geometric Mean Standard Deviation (Stdev):

- Type I error: Error rate represented by the false negative rate

$$TypeIError = FN/(TP + FN) \quad (1)$$

Table 1. confusion matrix for the C4.5 classifier

	Predict positive	Predict negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- Type II error : Error rate represented by false positive rate [21]

$$TypeIIError = FP/(FP + TN) \quad (2)$$

- Geometric Mean : Popular Evaluation Measure [20]

$$GMean = TPrate \times TNrate \quad (3)$$

$$TPrate = TP/(TP + FN) \quad (4)$$

$$TNrate = TN/(TN + FP) \quad (5)$$

6 Experiments and results

In all the experiments the C4.5 Decision Tree has been used as a classification algorithm. C4.5 is very common classifier based on ID3 algorithm with enhancements on handling missing values and continuous attribute value ranges with the ability to choose an appropriate attribute selection measure [22]. In general, decision trees are preferable for this kind of applications because they produce models that are interpretable and easy to explain by the decision makers.

For training and testing we created different datasets, 5-folds cross validation is applied with stratified sampling which stands on splitting the training dataset into several equal (or almost equal) partitions, then use one of these partitions as a test set and the remaining partitions uses to train the classifier, this procedure will be repeated for each partition. In other words, each partition will be a test set at each single step of cross validation. The final step of this technique is calculating the average accuracy of test each partition. Also, Stratified sampling is used to preserve the ratios of the two classes in the training and testing partitions, and make them as close as possible to the ratios in the all dataset.

We followed this approach: First, the bankruptcy data are normalized, then, the Oversampling methods are applied to handle imbalanced class distribution. Finally, the resulting data are processed using the C4.5 Classifier.

For the oversampling methods that uses k nearest neighbors algorithm to generate new instances for the minor class, we examined different values for k , starting from $k = 3$ up to 19 with a step of 2.

Table 2 shows the parameter settings for C4.5 classifier and all oversampling methods. In the first step of the experiment, the C4.5 Classifier is evaluated without applying any resampling method. The results of this evaluation are

Table 2. Experiment parameters

Algorithm	Parameter	Value
C4.5-C	Pruned	true
	Confidence	0.25
	Instances Per Leaf	2
SMOTE-I	Type of Interpolation	standard
SMOTE TL-I	Distance Function	HVDM
SMOTE ENN-I	Distance Function	Euclidean
ADASYN-I	Type of Interpolation	standard
ADOMS-I	Type of Interpolation	standard
SPIDER-I	Preprocessing Option	WEAK
Safe Level SMOTE	Type of Interpolation	standard
SPIDER2-I	Relabel	true
Borderline SMOTE I	Type of Borderline SMOTE	1
Common	Type of Interpolation	standard
	Alpha	0.5
	Mu	0.5
	Distance Function	HVDM
	Type of SMOTE	both
	Quantity of generated examples	1

Table 3. Prediction results with original data

Classifier	Type I Error	Type II Error	G-Mean
C4.5-C	77.4%	0.46 %	0.4583

shown in Table 3. Due to the high imbalanced data distribution, the classifier shows a poor performance in terms of Type I error.

The results in terms of Accuracy, Type I, Type II and G-mean are shown in Table 4 for all the oversampling approaches. While Type 2 error increased after applying all oversampling methods, our main focus was on Type I error because it is much more relevant in a bankruptcy prediction problem [23].

Figure 1 shows the changes in the evaluation measures for each of the top 5 methods when changing the number of neighbors. For example, $k = 13$ used with SMOTE ENN gave the lowest value for Type I error and an adequate value for Type II error; while $k = 7$ used with ADASYN gave the lowest value for type I error and an acceptable value for Type II error.

Finally, for each one of the previously mentioned methods we study the effect of oversampling on the decision tree complexity.

Table 5 shows the tree characteristics after applying each sampling method. We found that the tree complexity for SMOTE ENN is the best considering having the least number of leafs and nodes. From the results of this experiment we believe that SMOTE ENN is the best oversampling technique to be used with decision tree classifiers.

Table 4. Evaluation Results of all the oversampling techniques combined with the decision tree algorithm C4.5. Only the best k approach per method is shown.

Oversampling method	Type I Error	Type II Error	G-Mean	Accuracy
SMOTE (k=5)	24.19%	6.79%	0.838	92.82%
SMOTE TL (k=15)	17.74%	11.55%	0.853	88.32%
SMOTE ENN (k=13)	12.90%	12.37%	0.874	87.61%
Borderline SMOTE (k=19)	48.39%	2.86%	0.702	96.15%
Safe Level SMOTE (k=13)	27.41%	20.02%	0.757	79.81%
ROS	54.84%	2.00%	0.651	96.85%
ADASYN (k=5)	30.65%	6.76%	0.803	94.78%
ADOMS (k=5)	38.70%	4.76%	0.754	94.51%
SPIDER (k=3)	53.23%	2.03%	0.672	96.85%
SPIDER2 (k=11)	48.39%	1.82%	0.632	97.17%
AHC	48.38%	2.40%	0.699	96.53 %

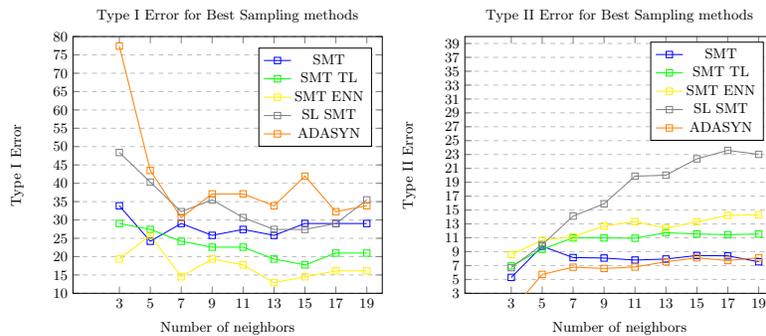


Fig. 1. Impact of the number of neighbors on the best 5 sampling methods

Table 5. decision tree complexity for C4.5

Method	Leafs	Nodes	Type I Error
C4.5 no Sampling	8.4	9.8	77.4%
SMOTE (k=5)	76	151.4	24.19%
SMOTE TL (k=15)	58.8	97.6	17.74%
SMOTE ENN (k=13)	54.2	79.8	12.90%

7 Conclusions and future work

Oversampling methods have proven their ability to handle imbalance data and enhance the performance of classifiers. These techniques have been previously applied to bankruptcy prediction problems, providing promising results. In this work, we have analyzed eleven different oversampling options on a real dataset for solving this problem using a Decision Tree-based classifier as C4.5. SMOTE ENN method obtained superior outcomes with the lowest type I error of 12.9% and G-mean with value of 0.874. The obtained decision tree has been considered as the best regarding having the least number of leafs and nodes.

Moreover, we noticed that all oversampling methods used lead to enhance results regarding type I error compared with classifying the original dataset without using any oversampling method.

The noticeable enhancements achieved by SMOTE methods might be due to the nature of this resampling technique, which replicates instances of minor classes depending on their neighbors. The number of neighbors represents an important factor in such sampling methods. Thus, for the considered dataset $K=13$ neighbors gives the best result for SMOTE ENN.

Future work can be done in this line, first, it could be used a different (and better) classification method than C4.5, which could be more benefited by the impact of the resampling techniques. Moreover, a study on the influence of feature selection algorithms on the performance of the classification methods could be conducted.

Acknowledgements

This work has been partially funded by projects EphemCH TIN2014-56494-C4-3-P, DeepBio TIN2017-85727-C4-2-P, SPIP2017-02116 and TEC2015-68752 (Spanish Ministry of Economy and Competitiveness and FEDER).

References

1. Altman, E.I., Hotchkiss, E.: Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt. Volume 289. John Wiley & Sons (2010)
2. Janer, J., Schneider, C.: Bankruptcy prediction and its advantages. Empirical Evidence from SMEs in the French Hospitality Industry (2011)
3. Alsawalqah, H., Faris, H., Aljarah, I., Alnemer, L., Alhindawi, N.: Hybrid smote-ensemble approach for software defect prediction. In: Computer Science On-line Conference, Springer (2017) 355–366
4. AlAgha, A.S., Faris, H., Hammo, B.H., Al-Zoubi, A.M.: Identifying β -thalassemia carriers using a data mining approach: The case of the gaza strip, palestine. Artificial Intelligence in Medicine **88** (2018) 70 – 83
5. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter **6**(1) (2004) 20–29

6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16** (2002) 321–357
7. Aljawazneh, H., Mora García, A.M., Castillo Valdivieso, P.A.: Predicting the financial status of companies using data balancing and classification methods. In: *Proceedings of the International Work-Conference on Time Series- ITISE*. Volume 2. (2017) 661–673
8. García, V., Marqués, A.I., Sánchez, J.S.: Improving risk predictions by preprocessing imbalanced credit data. In: *International Conference on Neural Information Processing*, Springer (2012) 68–75
9. Kim, M.J., Kang, D.K., Kim, H.B.: Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications* **42**(3) (2015) 1074–1082
10. Zięba, M., Tomczak, S.K., Tomczak, J.M.: Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications* **58** (2016) 93–101
11. Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. *Expert Systems with Applications* **83** (2017) 405–417
12. Le, T., Lee, M.Y., Park, J.R., Baik, S.W.: Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry* **10**(4) (2018) 79
13. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing* (2005) 878–887
14. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *Advances in knowledge discovery and data mining* (2009) 475–482
15. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence)*. IEEE International Joint Conference on, IEEE (2008) 1322–1328
16. Tang, S., Chen, S.P.: The generation mechanism of synthetic minority class examples. In: *Information Technology and Applications in Biomedicine, 2008. ITAB 2008. International Conference on, IEEE* (2008) 444–447
17. Stefanowski, J., Wilk, S.: Selective pre-processing of imbalanced data for improving classification performance. *Lecture Notes in Computer Science* **5182** (2008) 283–292
18. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: *Rough sets and current trends in computing*, Springer (2010) 158–167
19. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine* **37**(1) (2006) 7–18
20. Li, S., Wang, Z., Zhou, G., Lee, S.Y.M.: Semi-supervised learning for imbalanced sentiment classification. In: *IJCAI proceedings-international joint conference on artificial intelligence*. Volume 22. (2011) 1826
21. Chen, N., Vieira, A., Duarte, J.: Cost-sensitive lvq for bankruptcy prediction: An empirical study. In: *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on, IEEE* (2009) 115–119
22. Salzberg, S.L.: C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning* **16**(3) (1994) 235–240
23. Weiss, L.A., Capkun, V.: The impact of incorporating the cost of errors into bankruptcy prediction models. (2005)

The changing shape of sovereign default intensities

Yusho Kagraoka

Musashi University, Toyotama-kami 1-26-1, Nerima-ku, Tokyo 176-8534, Japan
kagraoka@cc.musashi.ac.jp

Zakaria Moussa

Université de Nantes, IAE Nantes - Économie & Management, Chemin la Censive du Tertre
BP 52231, 44322 Nantes, Cedex 3 France
Zakaria.Moussa@univ-nantes.fr

Abstract. The term structure of sovereign default intensities evolves over time along with fall/rise levels and steeping/flattening of the slope; a hump shape may exist in the default intensity curve, and the location of the hump changes. Thus, the default intensity model should have the flexibility to capture most of the variations in the term structure of the default intensities. The dynamic Nelson-Siegel (DNS) model with a time-varying decay parameter is appropriate to generate such default intensity curves. The paper studies the default intensities estimated from credit default swap (CDS) spreads by the DNS model with a time-varying decay parameter. Empirical studies were conducted on the German and U.S. CDS markets. The model parameters were successfully estimated using the Kalman filter. It is found that the magnitude of the decay parameter is positively related to the level of default intensities.

Keywords: sovereign credit default swap, sovereign default intensities, the Nelson-Siegel model, state space model, Kalman filter

1 Introduction

In the aftermath of the 2007-2008 financial crisis and the European sovereign debt crisis that started in 2008, sovereign credit default swap (CDS) spreads surged dramatically. Not only the levels of CDS spreads but also the shape of the term structure of CDS spreads changed dramatically during both these periods. The term structure of CDS spreads reflects the default probabilities of a reference entity and the timing of default. Thus, studying the shape of the term structure of CDS spreads and their evolution is important for forecasting CDS spreads and controlling credit risk.

Nelson and Siegel (1987) employed a set of simple functions to describe the shape of the term structure of interest rates. Diebold and Li (2006) applied the Nelson-Siegel (NS) model to forecast the term structure of interest rates. Diebold et al. (2006) investigated the term structure of interest rates using the NS model accompanied by macroeconomic variables and found that the macroeconomic factors improved forecasting of the term structure of interest rates. Many variations were developed to im-

prove the NS model. For instance, Koopman et al. (2010) extended the NS model by allowing the time-varying decay parameter and incorporating a GARCH structure into the error terms. Christensen et al. (2011) and Coroneo et al. (2011) constructed the arbitrage-free NS model. Levant and Ma (2017) and Zhu and Rahman (2015) extended the NS model to include Markov switching. The NS model has been applied not only to government bond yields but also to other term structures. Yu and Zivot (2011) and Krishnan et al. (2010) applied the NS model to corporate bond yields. Shaw et al. (2014) adopted a straightforward approach and applied the NS model to the European sovereign CDS spreads.

Credit default swap spreads depend on the default intensities and risk-free rates because risk-free rates are inevitable to obtain the present value of the fixed and floating legs of the CDS. The CDS spread is set to the level at which the present value of periodic payment made by a protection buyer and the present value of a loss compensation caused by the default of a reference entity are both equal. Therefore, CDS spreads are complicated functions of the risk-free rates and default intensities. For instance, Shaw et al. (2014) stated that CDS is a pure credit instrument and is isolated from the interest rate risk. Thus, the interpretation that CDS spread reflects pure credit risk is wrong. Shaw et al.'s (2014) study had another disadvantage. They applied the NS model to CDS spreads. However, it is not clear what a fundamental object is. The original NS model represents the dynamics of forward rates, and its model parameters are estimated from the term structures of spot rates. Shaw et al. (2014) applied a functional form for 'spot rates' to the CDS spreads. Thus, in Shaw et al. (2014), the counterpart to 'forward rates' is not interpreted clearly. In the original NS model, the functional form for forward rates has a clear interpretation: level, slope, and curvature of a term structure of forward rates. However, Shaw et al. (2014) did not provide any financial or economic interpretation for the underlying CDS spreads.

As explained above, CDS spreads are determined not only by the default intensities of a reference entity but also by the risk-free rates. Thus, to investigate CDS spreads, the risk-free rates are estimated. The U.S. Treasuries and German government bonds (bunds) are regarded as least risky. However, non-zero CDS spreads for the U.S. and Germany imply that these bonds are risky. Kagraoka and Moussa (2014) and Kagraoka (2018) proposed a method to estimate the risk-free rates and default intensities from government bond yields and the corresponding sovereign CDS spreads. They conducted empirical analysis and simultaneously estimated the risk-free rates in USD and EUR and default intensities for the U.S. and Germany. Their main idea was to estimate the risk-free rates by subtracting the credit spreads implied in the CDS spreads from the government bond yields.

The purpose of this study is to investigate the dynamics of sovereign default intensity by applying the DNS model with the time-varying decay parameter to the default intensities implied in the sovereign CDS spreads. The NS model parameters correspond to the level, slope, and curvature of the term structure of default intensities. To obtain the default intensities, this study follows the procedure proposed by Kagraoka (2018). An empirical study of the sovereign CDS spreads for several countries is interesting; however, this study is restricted to the CDS spreads for the U.S. and Ger-

many because these spreads are indispensable for estimating the risk-free rates in the USD and the EUR.

The rest of the paper is organised as follows. Section 2 explains how to extract the risk-free rates and default intensities from government bond yields and CDS spreads. Subsequently, the state space model represents the NS model, and the estimation method using the Kalman filter is described. Data and empirical results are given in Section 3. Section 4 concludes.

2 Methodology

2.1 Default intensity and risk-free rate

The valuation formula for CDS in the reduced form model is given in several textbooks, including Brigo and Mercurio (2006), Duffie (2003), Lando (2004), and Schönbucher (2003). This study follows Lando (1994, 1998) and Houweling and Vorst (2006). In their model the default event of a reference entity is modeled by a point process. Let τ , λ_t , and $\Pr(t, T)$ denote a default time, the default intensity at time t , and the survival probability at time t up to time T , respectively. The following relationship holds for them:

$$\Pr(t, T) = E_t \left[1_{\{\tau > T\}} \right] = E_t \left[\exp \left(- \int_t^T \lambda_s ds \right) \right]. \quad (1)$$

The value of the fixed leg of periodic payment is the expected present value of the periodic payments prior to the default time under a martingale measure. The value of the floating leg is the expected present value of compensation for a default. The CDS spread is set to the level at which the value of the fixed leg equals that of the floating leg. Apparently CDS spreads depend not only on the default probabilities of a reference entity but also on the term structure of risk-free rates.

Investors and researchers regard government bond yields as proxies for the risk-free rates. However, non-zero spreads of the sovereign CDS imply that government bonds are not risk-free assets. In fact, government bonds should be evaluated assuming that their default probability is similar to corporate bonds. Thus, government bond yields depend on the risk-free rates and the default intensities of the government.

Both CDS spreads and government bond yields are a complicated function of the risk-free rates and default intensities. It is not simple to disentangle the risk-free rates and the default intensities from the CDS spreads and government bond yields. A procedure to decompose government bond yields to the risk-free rates and credit spreads was proposed by Kagraoka and Moussa (2014) and Kagraoka (2018). The procedure is essentially subtracting the credit risk spread implied in the sovereign CDS spreads from the bond yields to extract the risk-free rates under the assumption that the risk-free rates and default intensities are independent. This study employs this method to obtain the default intensities.

2.2 State space models for default intensities

Nelson and Siegel (1987) proposed a set of parsimonious functions to express a yield curve. This model has been used to express the term structure of interest rates by many central banks worldwide as reviewed by the Bank for International Settlements (2005). Diebold and Li (2006) is the first paper that advocates the NS methodology to forecast the time evolutions of the term structures of interest rates. In the NS model, time- t forward rate maturing at $t + \tau$ is expressed as

$$f_t(\tau) = \beta_{1,t} + \beta_{2,t} e^{-\lambda^{NS} \tau} + \beta_{3,t} \lambda^{NS} e^{-\lambda^{NS} \tau}. \quad (2)$$

This parametrisation of the forward rate can reproduce various shapes of the term structure of forward rates. Changes in the parameters generate variations of the term structure of forward rates. The NS model has four parameters; the level parameter, β_1 , represents a long-term interest rate and changes in the level parameter cause parallel shifts in the term structure of forward rates; the slope parameter, β_2 , dictates a gradient of the term structure of forward rates and its movements induce steepening or flattening of the term structure of forward rates. The curvature parameter, β_3 , represents a hump shape of the term structure of forward rates and its variations produce changes in the degree of hump in the term structure of forward rates. The decay parameter, λ^{NS} , governs both the speed of the exponential decay rate of the slope factor and the location of the hump in the curvature factor. Spot rate,

$$r_t(\tau) = \frac{1}{\tau} \int_t^{t+\tau} ds f_t(s), \quad (3)$$

is an average of forward rates, and it is expressed as

$$r_t(\tau) = \beta_{1,t} + \beta_{2,t} \left(\frac{1 - e^{-\lambda^{NS} \tau}}{\lambda^{NS} \tau} \right) + \beta_{3,t} \left(\frac{1 - e^{-\lambda^{NS} \tau}}{\lambda^{NS} \tau} - e^{-\lambda^{NS} \tau} \right) \quad (4)$$

in the NS model. All previous empirical studies estimate the NS parameters from the term structure of spot rates by using equation (4).

In the context of CDS valuation, forward rates correspond to default intensities. Therefore, spot rates correspond to average default intensities,

$$\Phi_t(\tau) = \frac{1}{\tau} \int_t^{t+\tau} ds \lambda_s. \quad (5)$$

This study empirically examines the average default intensity by applying the DNS model.

Diebold and Li (2006) proposed to fix to an a priori value the decay parameter and formulated the dynamics of the term structure of interest rates by using a vector autoregressive model. They estimated the NS model parameters by applying univariate AR processes. Diebold et al. (2006) employed the state space model to describe the

dynamics of the term structure of interest rates expressed by the NS model. The evolution of the term structure of interest rates is generated by the transition equation,

$$\beta_{t+1} - \mu^\beta = T^\beta (\beta_t - \mu^\beta) + R_t^\beta \eta_t^\beta, \quad (6)$$

where β_t is a 3×1 unobservable state vector, T^β is a 3×3 matrix, the error term η_t^β is a 3×1 column vector following $\eta_t^\beta \sim N(0, \Sigma_\eta^\beta)$, and $R_t^\beta = I$. The vector of the observable variables is denoted by $y_t(\tau) = (y_t(\tau_1), y_t(\tau_2), \dots, y_t(\tau_p))'$, and the measurement equation is written as

$$y_t(\tau) = \Lambda \beta_t + \varepsilon_t^\beta \quad (7)$$

where

$$\Lambda = \begin{pmatrix} 1 & \frac{1 - e^{-\lambda^{NS} \tau_1}}{\lambda^{NS} \tau_1} & \frac{1 - e^{-\lambda^{NS} \tau_1}}{\lambda^{NS} \tau_1} - e^{-\lambda^{NS} \tau_1} \\ 1 & \frac{1 - e^{-\lambda^{NS} \tau_2}}{\lambda^{NS} \tau_2} & \frac{1 - e^{-\lambda^{NS} \tau_2}}{\lambda^{NS} \tau_2} - e^{-\lambda^{NS} \tau_2} \\ \vdots & \vdots & \vdots \\ 1 & \frac{1 - e^{-\lambda^{NS} \tau_p}}{\lambda^{NS} \tau_p} & \frac{1 - e^{-\lambda^{NS} \tau_p}}{\lambda^{NS} \tau_p} - e^{-\lambda^{NS} \tau_p} \end{pmatrix}, \quad (8)$$

and $\varepsilon_t^\beta \sim N(0, \Sigma_\varepsilon^\beta)$. It is assumed that the white noise transition disturbance, η_t^β , and the measurement disturbances, ε_t^β , are orthogonal to each other. The initial state is assumed to be orthogonal to the transition and measurement disturbances; $E[\beta_1 \eta_1^{\beta \prime}] = 0$ and $E[\beta_1 \varepsilon_1^{\beta \prime}] = 0$. The non-diagonal elements of Σ_η^β allow for the shocks to the term structure to be correlated. It is further assumed that Σ_ε^β is an orthogonal matrix.

Koopman et al. (2010) developed the DNS model with the time-varying decay parameter, λ_t^{NS} . This study follows their methodology. Logarithm of the decay parameter is included in the fourth element of the state space to ensure positivity of it,

$$\alpha_t = \begin{pmatrix} \beta_t \\ \ln \lambda_t^{NS} \end{pmatrix}. \quad (9)$$

This model is not linear with respect to $\ln \lambda_t^{NS}$, and this property makes it difficult to use a Kalman filter technique for model estimation. Koopman et al. (2010) linearised this model as follows,

$$\alpha_{t+1} - \mu^\alpha = T^\alpha (\alpha_t - \mu^\alpha) + R_t^\alpha \eta_t^\alpha, \quad (10)$$

$$y_t = Z_t \alpha_t + (\Lambda(a_{t|t-1}) - Z_t a_{t|t-1}) + d_t + \varepsilon_t^\alpha, \quad (11)$$

$$Z_t = \left(\Lambda_1(\tau) \quad \Lambda_2(\tau) \quad \Lambda_3(\tau) \quad \lambda_t^{NS} \left(\sum_{j=1}^3 \dot{\Lambda}(a_{j,t|t-1}) a_{j,t|t-1} \right) \right), \quad (12)$$

and

$$\dot{\Lambda}(\alpha_t) = \begin{pmatrix} \frac{\partial \Lambda_1(\tau)}{\partial \lambda_t^{NS}} & \frac{\partial \Lambda_2(\tau)}{\partial \lambda_t^{NS}} & \frac{\partial \Lambda_3(\tau)}{\partial \lambda_t^{NS}} \end{pmatrix} = \begin{pmatrix} 0 & \frac{\partial \Lambda_2(\tau)}{\partial \lambda_t^{NS}} & \frac{\partial \Lambda_3(\tau)}{\partial \lambda_t^{NS}} \end{pmatrix}, \quad (13)$$

where α_t is an unobservable state vector, T^α is a 4×4 matrix, and the error term η_t^α is a 4×1 column vector with the coefficient matrix $R_t^\alpha = I$. This state space model is estimated by maximising the log-likelihood function, as explained by Durbin and Koopman (2012).

3 Empirical analysis

3.1 Data

Weekly data were collected from 8 October 2008 to 27 December 2017 to estimate the NS model. Datastream provides CDS spread data for Germany in USD and for the U.S. in EUR, maturing at six months and one, two, three, four, five, seven, and ten years. The spot rates of the AAA-rated European government bonds maturing from three months to ten years, with one-month increments, are obtained from the European Central Bank. The Federal Reserve Board gives the spot rates for U.S. Treasuries from one year to ten years, with one-year increments.

The risk-free rates and default intensities are estimated following Kagraoka and Moussa (2014) and Kagraoka (2018). Kagraoka (2018) advocates employing CDS spreads in a foreign currency to estimate the default intensities. Thus, CDS spreads for Germany in USD and for the U.S. in EUR are employed to estimate default intensities. Spot rates in EUR and USD are depicted in Figure 1 where spot rates are expressed in percent. In addition, CDS spreads for Germany in USD and for the U.S. in EUR are depicted in Figure 1 where CDS spreads are expressed in percent. The estimated risk-free rates and (average) default intensities for Germany and for the U.S.

are illustrated in Figure 2 where all the values are expressed as a percentage. Summary statistics of the average default intensities for Germany and for the U.S. are given in Table 1.

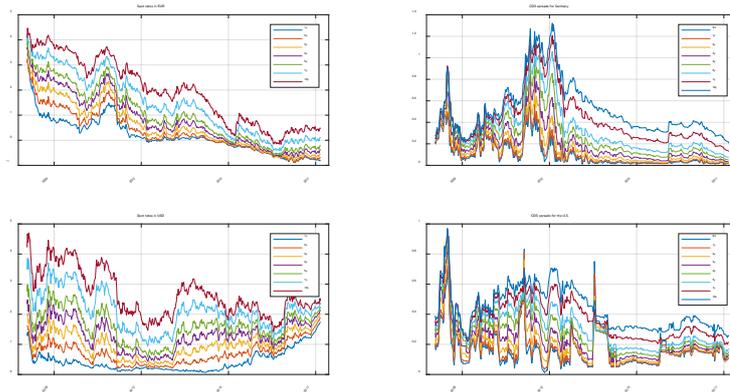


Fig. 1. Term structure of spot rates in EUR (spot rates of the AAA-rated European government bonds) (upper left) and that in USD (U.S. Treasury spot rates) (lower left). Term structure of CDS spreads for Germany in USD (upper right) and for the U.S. in EUR (lower right).

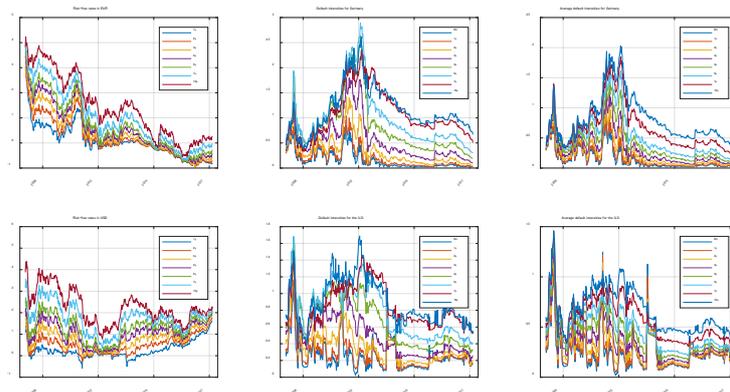


Fig. 2. Term structure of risk-free rates in EUR (upper left) and USD (lower left). Term structure of default intensities for Germany (upper middle) and for the U.S. (lower middle). Term structure of average default intensities for Germany (upper right) and for the U.S. (lower right).

Average default intensity for Germany							
	mean	sd	min	max	$\rho(1)$	$\rho(4)$	$\rho(13)$
6m	0.151908	0.173576	0.015387	0.815305	0.954321	0.835564	0.527828
1y	0.175605	0.187375	0.019888	0.897661	0.961482	0.853667	0.555690
2y	0.232932	0.221224	0.031533	1.085208	0.971972	0.881604	0.604502
3y	0.309992	0.262280	0.049142	1.275732	0.978321	0.902066	0.656700
4y	0.412099	0.311455	0.076321	1.467882	0.982889	0.919855	0.718082
5y	0.517864	0.356432	0.113359	1.635132	0.985442	0.930605	0.758047
7y	0.669876	0.375916	0.200385	1.844075	0.986121	0.936778	0.786894
10y	0.815823	0.388512	0.325306	2.027354	0.986390	0.941117	0.809478

Average default intensity for the U.S.							
	mean	sd	min	max	$\rho(1)$	$\rho(4)$	$\rho(13)$
6m	0.244607	0.173464	0.017624	1.247212	0.900185	0.619577	0.118924
1y	0.259836	0.172130	0.049238	1.215493	0.914750	0.661749	0.180961
2y	0.297056	0.176276	0.091932	1.185185	0.941540	0.747669	0.331098
3y	0.347332	0.188019	0.121615	1.277714	0.960292	0.816248	0.475995
4y	0.412049	0.202571	0.170679	1.364057	0.971382	0.861509	0.582163
5y	0.478162	0.215528	0.209673	1.431280	0.977092	0.886523	0.641468
7y	0.574024	0.213913	0.301622	1.453785	0.977731	0.890182	0.654322
10y	0.666951	0.218047	0.351210	1.460032	0.976889	0.891806	0.669006

Table 1. Summary statistics for average default intensities for Germany and for the U.S. For each maturity mean, standard deviation (sd), minimum, maximum, and three autocorrelation coefficients at 1 week ($\rho(1)$), 4 weeks ($\rho(4)$), and 13 weeks ($\rho(13)$) are reported.

3.2 Parameter estimation

The NS model parameters are estimated from the weekly data of average default intensities by maximising likelihood. The average default intensities maturing at six months and one, two, three, four, five, seven, and ten years are regarded as observable by matching the CDS maturities.

The CDS spreads and the average default densities of Germany and the U.S. were volatile between 2008 and 2013 and the shapes of their term structure of average default intensities changed drastically. For Germany, the term structure of average default intensities was very low and flat at October 2008. Later, the level of term structure of average default intensities surged abruptly and had a hump at around five years. From 2013, the term structure of average default intensities resumed to flat at low level. For the U.S., the term structure of average default intensities was very volatile from 2008 to 2014 and it took a convex or concave shape. Therefore, the NS model with time-varying decay parameter is appropriate to model the term structure of the average default intensities for both Germany and the U.S.

The estimated parameters of the NS model with time-varying decay parameters for Germany and the U.S. are reported in Table 2. The diagonal elements of Σ_e^β are expressed by taking their logarithm. Trajectories of the state variables for Germany and the U.S. are presented in Figure 3. This figure shows the time decay parameters for Germany and the U.S. vary with time.

Germany									
T					Σ_{η}				
L_1	S_1	C_1	$\ln(\lambda_t)$	L_1	S_1	C_1	$\ln(\lambda_t)$	μ	
0.997346 (0.054613)	-0.011206 (0.056234)	-0.006998 (0.017658)	0.011472 (0.018163)	0.009645 (0.001188)				0.798911 (1.786479)	
0.042320 (0.065951)	1.049558 (0.079344)	0.024198 (0.026728)	-0.025959 (0.031767)	-0.008074 (0.001305)	0.010052 (0.001565)			-0.385022 (2.025525)	
-0.211497 (0.074002)	-0.171875 (0.092735)	0.979983 (0.029477)	0.019775 (0.038131)	-0.008320 (0.001246)	0.009525 (0.001594)	0.013144 (0.001813)		-2.180955 (2.849398)	
0.153534 (0.036830)	0.143602 (0.046514)	0.034850 (0.015260)	0.960038 (0.017258)	-0.003259 (0.000755)	0.003944 (0.000842)	0.003365 (0.000833)	0.002738 (0.000435)	-3.212022 (0.151155)	
Σ_{ϵ}									
6m	1y	2y	3y	4y	5y	7y	10y		
-13.425748 (3.788603)	-7.804227 (2.000860)	-12.907577 (1.199934)	-8.018255 (0.469985)	-12.222887 (1.259355)	-7.107245 (0.404486)	-6.170986 (0.122477)	-6.888993 (0.088776)		

the U.S.									
T					Σ_{η}				
L_1	S_1	C_1	$\ln(\lambda_t)$	L_1	S_1	C_1	$\ln(\lambda_t)$	μ	
0.878359 (0.020443)	-0.064492 (0.021695)	-0.011075 (0.008482)	0.010287 (0.005668)	0.001894 (0.000303)				1.07773 (0.153635)	
0.056069 (0.051403)	0.922664 (0.028603)	0.039282 (0.016508)	-0.004535 (0.023868)	-0.001852 (0.000593)	0.009829 (0.001013)			-0.860613 (0.212719)	
0.216840 (0.073223)	0.376384 (0.063708)	0.894909 (0.018866)	-0.032526 (0.036477)	-0.000882 (0.000989)	-0.004373 (0.000882)	0.021298 (0.003137)		-1.060062 (0.412509)	
-0.032409 (0.032712)	-0.126832 (0.037530)	0.060063 (0.010157)	1.008322 (0.016439)	-0.000558 (0.000523)	0.002761 (0.000800)	-0.007518 (0.001408)	0.005847 (0.001090)	-3.953290 (0.085850)	
Σ_{ϵ}									
6m	1y	2y	3y	4y	5y	7y	10y		
-7.928865 (0.253388)	-11.144247 (2.207895)	-9.259559 (0.228988)	-8.867821 (0.739341)	-10.529601 (0.690370)	-8.089161 (0.609211)	-7.842452 (0.106278)	-8.242364 (0.197203)		

Table 2. Estimated parameters of the Nelson-Siegel model with the time-varying decay parameter for average default intensity for Germany (upper panel) and for the U.S. (lower panel).

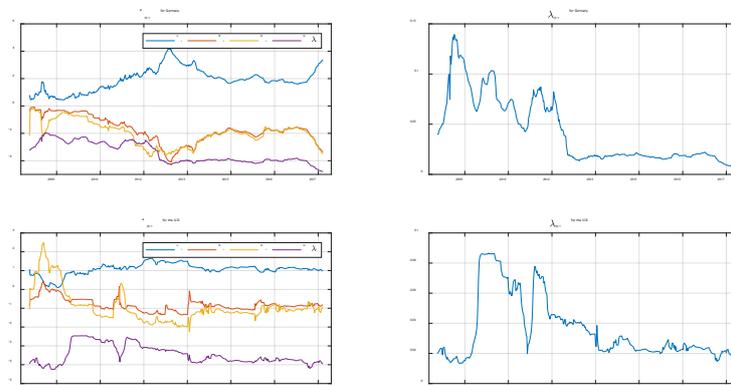


Fig. 3. Time series of the state variables for Germany (upper left) and that for the U.S. (lower left) are drawn. Time series of the time-varying decay parameter, λ_t^{NS} , for Germany (upper right) and that for the U.S. (lower right) are depicted.

First, the estimated parameters of default intensities for Germany are discussed. Level, L_t , is likely to remain the same because $T_{1,1} = 0.997346$ and $T_{1,j}$ ($j = 2, 3, 4$) are close to zero. Likewise, the slope, S_t , is likely to remain unchanged because $T_{2,2} = 1.049558$ and $T_{2,j}$ ($j = 1, 3, 4$) are close to zero. The curvature, C_t , is mainly affected by itself because $T_{3,3} = 0.979983$, and negatively affected by level and slope because $T_{3,1} = -0.211497$ and $T_{3,2} = -0.171875$. The logarithm of the time decay parameter, $\ln \lambda_t$, is mainly affected by itself because $T_{4,4} = 0.960038$, and it is affected by all the factors because $T_{4,1} = 0.153534$, $T_{4,2} = 0.143602$, and $T_{4,3} = 0.034850$. The trajectories of the level, slope, curvature and decay parameters are intriguing. The level correlates negatively to the slope and curvature. The decay parameter varies over time; it rises when the default intensities become higher and declines when the default intensities decrease. During the financial market turmoil, the decay parameter peaked to 0.1394 on 15 April 2009. It took around 0.02 after mid-2012. The correlations between the state variables are given in the left panel in Table 3. The negative correlation between the decay parameter and the level is apparent. The correlations between the decay parameter and the slope or curvature are positive. The correlation between the decay parameter and the five-year average default intensity is 0.649595. Thus, the higher the average default intensity, the greater is the decay parameter.

The estimated parameters of default intensities for the U.S. are discussed. Level is affected by itself and slope. Slope is affected by itself as well as the level and curvature. Curvature is affected by level, slope, and curvature. These parameters produce upward or downward sloping curves of default intensities for the U.S. The logarithm of the time decay parameter is mainly affected by itself because $T_{4,4} = 1.008322$, and by the slope ($T_{4,2} = -0.126832$) and curvature ($T_{4,3} = 0.060063$). The trajectories of the level, slope, curvature and time decay parameters are intriguing. The level negatively correlates to the slope and curvature as in the case of Germany. The decay parameter varies over time; it rises when the default intensities become higher and drops when the default intensities lower. In the financial market turmoil, it peaked at 0.084926 at 15 April 2009. It was below 0.03 since 2013. The correlations between the state variables are given in the right panel in Table 3. The correlation signs between the decay parameter and the level, slope or curvature are opposite to that for Germany. The correlation between the decay parameter and the five-year average default intensity is 0.404720. Thus, the higher the average default intensity, the greater the decay parameter becomes.

Germany					the U.S.				
	L_t	S_t	C_t	$\ln(\lambda_t)$		L_t	S_t	C_t	$\ln(\lambda_t)$
L_t	1.000000				L_t	1.000000			
S_t	-0.980039	1.000000			S_t	-0.887663	1.000000		
C_t	-0.937059	0.918793	1.000000		C_t	-0.875178	0.865943	1.000000	
$\ln(\lambda_t)$	-0.663970	0.766833	0.543543	1.000000	$\ln(\lambda_t)$	0.327648	-0.273805	-0.401542	1.000000

Table 3. Correlation between the state variables.

4 Conclusion

This study applies the dynamic Nelson-Siegel (DNS) model with the time-varying decay parameter to the average default intensities. The DNS factors correspond to the level, slope, and curvature of the default intensities, and the empirical results provide a clear interpretation of the estimated parameters. The DNS model with a time-varying decay parameter accommodates the time-varying features of the term structure of average default intensities and their variations. Empirical studies for the German and U.S. CDS markets are conducted and the Kalman filter is used to estimate the model parameters. The empirical results show that the decay parameter is time-varying and closely relates to the level, slope and curvature of average default intensities. The higher the average default intensity, the greater is the magnitude of the decay parameter. Future studies will investigate forecasting of the average default intensity by the DNS model.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 18K01707.

Reference

1. Bank for International Settlements (2005), Zero-coupon yield curves: technical documentation, BIS Papers No 25.
2. Brigo, Damiano and Fabio Mercurio (2006), Interest Rate Models – Theory and Practice: With Smile, Inflation and Credit, 2nd Edition, Springer.
3. Christensen, Jens H.E., Francis X. Diebold, and Glenn D. Rudebusch (2011), The affine arbitrage-free class of Nelson-Siegel term structure models, *Journal of Econometrics*, 164-1, 4-20.
4. Coroneo, Laura, Ken Nyholm and Rositsa Vidova Koleva (2011), How arbitrage-free is the Nelson-Siegel model?, *Journal of Empirical Finance*, 18-3, 393-407.
5. Diebold, Francis X. and Canlin Li (2006), Forecasting the term structure of government bond yields, *Journal of Econometrics*, 130-2, 337-364
6. Diebold, Francis X., Glenn D. Rudebusch, and S. Boragan Aruoba (2006), The macroeconomy and the yield curve: a dynamic latent factor approach, *Journal of Econometrics*, 131-1 (2006) 309-228.
7. Duffie, Darrell (2003), *Credit Risk: Pricing, Measurement, and Management*, Princeton University Press.

8. Durbin, James and Siem Jan Koopman (2012), *Time Series Analysis by State Space Methods*, Second Edition, Oxford University Press.
9. Houweling, Patrick and Ton Vorst (2005), Pricing default swaps: Empirical evidence, *Journal of International Money and Finance*, 24-8, 1200-1225.
10. Kagraoka, Yusho (2018), Dose the risk-free interest rate correlate with sovereign default intensities?, Working paper, Musashi University.
11. Kagraoka, Yusho and Zakaria Moussa (2014), Estimation of the Term Structure of CDS-Adjusted Risk-Free Interest Rates, *The Journal of Fixed Income* 24-2, 29-44.
12. Koopman, Siem Jan, Max I. P. Mallee, and Michel Van der Wel (2010), Analyzing the term structure of interest rates using the dynamic Nelson-Siegel model with time-varying parameters, *Journal of Business & Economic Statistics* 28-3, 329-343.
13. Krishnan, C.N.V., Peter H. Ritchken, and James B. (2010), Thomson, Predicting credit spreads, *Journal of Financial Intermediation*, 19-4, 529-563
14. Lando, David (1994), *Three essays on contingent claims pricing*, PhD Dissertation, Cornell University.
15. Lando, David (1998), On Cox processes and credit risky securities, *Review of Derivatives Research*, 2, 99-120.
16. Lando, David (2004), *Credit Risk Modeling: Theory and Applications*, Princeton University Press.
17. Levant, Jared and Jun Ma (2017), A dynamic Nelson-Siegel yield curve model with Markov switching, *Economic Modelling*, 67, 73-87
18. Nelson, Charles R. and Andrew F. Siegel (1987), Parsimonious modeling of yield curves, *The Journal of Business*, 60-4, 473-489.
19. Schönbucher, Philipp J. (2003), *Credit Derivatives Pricing Models: Models, Pricing and Implementation*.
20. Shaw, Frances, Finbarr Murphy, and Fergal O'Brien (2014), The forecasting efficiency of the dynamic Nelson Siegel model on credit default swaps, *Research in International Business and Finance*, 30, 348-368.
21. Yu, Wei-Choun and Eric Zivot (2011), Forecasting the term structures of Treasury and corporate yields using dynamic Nelson-Siegel models, *International Journal of Forecasting*, 27-2, 579-591.
22. Zhu, Xiaoneng and Shahidur Rahman (2015), A regime-switching Nelson-Siegel term structure model of the macroeconomy, *Journal of Macroeconomics*, 44, 1-17.

PoARX models for count time series

Jamie Halliday and Georgi N. Boshnakov

University of Manchester, United Kingdom
jamie.halliday@manchester.ac.uk

Abstract. This paper introduces multivariate Poisson autoregressive models with exogenous covariates (PoARX) for modelling multivariate time series of counts. We state conditions for a PoARX process to be stationary and ergodic before proposing a computationally efficient procedure for estimation of parameters by the method of inference functions (IFM) and stating asymptotic normality of these estimators. Lastly, we demonstrate an application to count data for the number of people entering and exiting a building, and show how the different aspects of the model combine to produce a strong predictive model. We conclude by listing directions for future work.

Keywords: Multivariate time series, Count data, Prediction, Copula

1 Introduction

The abundance of data brought about by the digital revolution has increased the availability of time series of counts. Such data appear in many areas, including statistics, econometrics, and the social and physical sciences. The most popular distribution for modelling count data is the Poisson distribution, which has attractive properties and is in some respects the count analogue of the Gaussian distribution. One restrictive property of the Poisson distribution however is that the mean and the variance are equal – this is rarely observed in applications.

For time series data, the correlation between observations provides additional challenges. The classic example of time series models is the ARMA model, which has multivariate extensions. A fruitful approach, employed in ARCH and GARCH models [6,3], uses a separate equation to model directly the dependence of the variance on the past. In order to improve the predictive accuracy, the aforementioned models have been augmented with additional exogenous covariates. However, these models do not make specific provision for the non-negativity and integer-valued nature of count data. An integer-valued analogue of the GARCH model, called INGARCH [7], uses Poisson deviates rather than normal innovations to combat these issues. Furthermore, a Poisson model for integer-valued time series has been proposed, called the Poisson autoregressive model [8], which has an autoregressive feedback mechanism for the mean. Subsequently, a class of dynamic Poisson models allowing for exogenous covariates was suggested called PARX [1]. Whilst the Poisson distribution has been widely used for univariate count models, multivariate generalisations have been relatively sparse so far. Recently, a summary of multivariate (Poisson) distributions for count data has been

made [15], including multivariate extensions of a parametric (Poisson) distribution and copula modelling using univariate (Poisson) marginal distributions.

In this article, we use a copula to extend the (univariate) PARX model [1] to multivariate count time series. This approach is flexible and tractable. Use of covariates in the Poisson model offers clear potential for better modelling and, by including the time series covariates, we allow over-dispersed data to be considered by our model. Implementation in R [24] is available in the developmental package PoARX [11].

This paper is organised as follows. Section 2 introduces the univariate and multivariate PoARX model, giving stationarity and ergodicity conditions. In Section 3 we discuss estimation of parameters by the method of inference functions (IMF) [17] and asymptotic results for the resulting estimators are stated. We consider prediction in Section 4, looking at the generating functions for future horizons. Then we demonstrate an application of the PoARX model in Section 5 by analysing a bivariate time series of count data for number of people entering and exiting a building on the University of California, Irvine (UCI) campus [14]. Exogenous covariates, such as the occurrence of a meeting or conference are included in the model to aid predictive accuracy. We summarise our findings in Section 6 and outline suggestions for future work.

2 The multivariate PoARX model

In this section we present the new class of models, introducing the necessary background material about the univariate PoARX model and copulas before generalising to higher dimensions. For the purpose of this article we focus on using Frank's copula to capture dependence between time series, but any suitable copula could be used. We use Frank's copula because in two dimensions it can account for both positive and negative dependence by allowing the dependence parameter to take values along the entire real line except zero.

2.1 The univariate PoARX model

First, a note on terminology – [1] use the abbreviation PARX for this model but we prefer PoARX since it seems to suggest more clearly “Poisson” and avoids confusion with other meanings of “P” in similar abbreviations. For example, PAR is often used to mean periodic autoregression.

Let $\{Y_t; t = 1, 2, \dots\}$ denote an observed time series of counts, so that $Y_t \in \{0, 1, 2, \dots\}$ for all $t = 1, 2, \dots$. Further, let $x_{t-1} \in \mathbb{R}^r$ denote a vector of additional covariates considered for inclusion in the model. We say that $\{Y_t\}$ is a univariate PoARX(p, q) process and write $\{Y_t\} \sim \text{PoARX}_1(p, q)$, if its dynamics can be written as follows:

$$\begin{aligned} Y_t | \mathcal{F}_{t-1} &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= \omega + \sum_{l=1}^p \alpha_l Y_{t-l} + \sum_{l=1}^q \beta_l \lambda_{t-l} + \eta \cdot x_{t-1}, \end{aligned} \tag{1}$$

where $\text{Poisson}(\lambda)$ denotes a Poisson distribution with parameter λ , \mathcal{F}_{t-1} denotes the σ -field of past knowledge, $\sigma\{Y_{1-p}, \dots, Y_{t-1}, \lambda_{1-q}, \dots, \lambda_{t-1}, x_1, \dots, x_{t-1}\}$, $\omega \geq 0$ is an intercept term, $\{\alpha_1, \dots, \alpha_p\}$ and $\{\beta_1, \dots, \beta_q\}$ are non-negative autoregressive coefficients, and η is a vector of non-negative coefficients for the exogenous covariates. Thus, the model for the intensity, λ_t , uses the past p values of the process, the past q values of the intensity and the covariates.

In order to ensure that the process is stationary and ergodic with polynomial moments of a given order, we place two further restrictions on the model [1]. Firstly, the autoregressive coefficients must obey the following condition,

$$\sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i) < 1. \quad (2)$$

Additionally, we require that each component of the exogenous covariates, denoted $x_t(k)$ to avoid confusion later, follows a Markov structure, that is,

$$x_t(k) = g(x_{t-1}(k), \dots, x_{t-m}(k); \epsilon_t), \quad k = 1, \dots, r, \quad (3)$$

for some $m > 0$ and some function $g(\mathbf{x}, \epsilon)$ with vector \mathbf{x} independent of the observed Y_t and unobserved λ_t , and with ϵ_t an i.i.d. error term.

2.2 Copulas

Copulas provide a well-defined approach to model multivariate data, with the dependence structure considered separately from the univariate margins [17]. Copula theory can be attributed to a theorem stating that any multivariate distribution can be represented as a function of its marginals [25]. Estimation of the copula is relatively straightforward as a two-stage procedure [16]. First the univariate margins are fitted to respective parameters before the copula fit to find the value of the dependence parameter. An important class of copulas are called Archimedean copulas [23], which can be constructed easily from a generator function $\varphi(\cdot)$.

Frank's copula [23] is one example of such a copula. As mentioned, the dependence parameter can take any value except zero in the two-dimensional case ($\rho \in \mathbb{R} \setminus \{0\}$). In higher dimensions, however, the dependence parameter is limited to values in $(0, \infty)$. In any case the limit as $\rho \rightarrow 0$ corresponds to independence. Using a subscript ρ to denote the case of Frank's copula, we write

$$C_\rho(u_1, \dots, u_K) = \varphi_\rho^{[-1]} \left(\sum_{i=1}^K \varphi_\rho(u_i) \right), \quad (4)$$

where the generator function is given by

$$\varphi_\rho(t) = -\log \left(\frac{\exp(-\rho t) - 1}{\exp(-\rho) - 1} \right), \quad (5)$$

and its inverse

$$\varphi_\rho^{[-1]}(t) = \varphi_\rho^{-1}(t) = -\frac{1}{\rho} \log(1 + \exp(-t)(\exp(-\rho) - 1)). \quad (6)$$

For discrete random variables the copula is no longer unique due to the presence of stepwise marginal distribution functions [18]. Despite this issue, copula models are still valid constructions for discrete distributions [9]. Additionally, evidence has been provided that suggests there are fewer identification problems when the marginal distributions are conditioned non-trivially upon covariates [27]. Joint probabilities are computed as rectangle probabilities.

2.3 The multivariate PoARX model

Let $\{Y_t = (Y_t^1, \dots, Y_t^K), t = 1, 2, \dots\}$ be a multivariate time series and let $\{x_{t-1}^j = (x_{t-1}^j(1), \dots, x_{t-1}^j(r))^\top, j = 1, 2, \dots, K\}$ be the matrix of exogenous covariates associated with Y_t . We say that $\{Y_t\}$ is a PoARX process and write $\{Y_t\} \sim \text{PoARX}_K(p, q)$, if each of the component time series is a univariate PoARX process and the joint conditional distribution is a copula Poisson. Let the intensities of PoARX processes be $\{\lambda_t^j; t = 1, 2, \dots, j = 1, \dots, K\}$ and be denoted using $\lambda_t = (\lambda_t^1, \dots, \lambda_t^K)$.

Let $\mathcal{D}(\lambda^1, \dots, \lambda^K; \rho)$ be a multivariate distribution based on Frank's copula (Equations (4)–(6)) with marginal distributions $\text{Poisson}(\lambda^1), \dots, \text{Poisson}(\lambda^K)$ and dependency parameter ρ . Before stating the entire behaviour of the multivariate model, the distribution function corresponding to $\mathcal{D}(\lambda^1, \dots, \lambda^K; \rho)$ is

$$F(y; \lambda, \rho) = C_\rho(F_1(y^1; \lambda^1), \dots, F_K(y^K; \lambda^K)), \quad (7)$$

where F_1, \dots, F_K are the distribution functions of the Poisson marginals, i.e.

$$F_j(x; \mu) = \sum_{k=0}^x e^{-\mu} \frac{\mu^k}{k!}, \quad j = 1, \dots, K.$$

The conditional distribution of Y_t is a Frank's copula distribution

$$Y_t | \mathcal{F}_{t-1} \sim \mathcal{D}(\lambda_t^1, \dots, \lambda_t^K; \rho), \quad (8a)$$

where \mathcal{F}_{t-1} denotes the σ -field defined by all previous observations and exogenous covariates, $\sigma\{Y_{1-p}, \dots, Y_{t-1}, \lambda_{1-q}, \dots, \lambda_{t-1}, x_1, \dots, x_{t-1}\}$, where each term contains information on all components of the time series. As before, the dynamics of the components of Y_t are specified by the equations:

$$Y_t^j | \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t^j), \quad j = 1, \dots, K; \quad (8b)$$

$$\lambda_t^j = \omega^j + \sum_{l=1}^p \alpha_l^j Y_{t-l}^j + \sum_{l=1}^q \beta_l^j \lambda_{t-l}^j + \eta^j \cdot x_{t-1}^j, \quad j = 1, \dots, K; \quad (8c)$$

where $\alpha_l^j, \beta_l^j \geq 0$ denote coefficients for the past values of the observations and intensities respectively, η^j denotes the vector of (non-negative) coefficients for the exogenous covariates, and $\omega^j \geq 0$ denotes an (optional) intercept term. For each univariate process, the two conditions in Equations (2) and (3) must hold.

2.4 Properties of multivariate PoARX

Properties such as stationarity and ergodicity for PoARX models can be based upon the univariate results [1] developed using τ -weak dependence. For notational ease, impose the simpler Markov structure found below,

$$x_t^j(k) = g^j \left(x_{t-1}^j(k); \epsilon_t^j \right), \quad j = 1, \dots, K, \quad k = 1, \dots, r. \quad (9)$$

However, the statements hold for the more general structure found in Equation (3). We also make three assumptions similar to those found for the univariate model [1].

Assumption 1 (Markov) The innovations ϵ_t^j and Poisson processes $N_t^j(\cdot)$ are i.i.d. for all $j = 1, 2, \dots, K$.

Assumption 2 (Exogenous stability)

$$\mathbb{E} \left\| g^j \left(x^j; \epsilon_t^j \right) - g^j \left(\tilde{x}^j; \epsilon_t^j \right) \right\|^s \leq \kappa \|x^j - \tilde{x}^j\|^s$$

for some $\kappa < 1$ and $\mathbb{E} \left\| g^j \left(0; \epsilon_t^j \right) \right\|^s < \infty$ for all $j = 1, 2, \dots, K$, for some $s \geq 1$.

Assumption 3 (PoARX stability) $\sum_{i=1}^{\max(p,q)} \alpha_i^j + \beta_i^j < 1$, for each $j = 1, 2, \dots, K$.

In the formulae below the operator vec has its usual meaning. For a matrix A , $\text{vec}(A)$ is a (column) vector obtained by stacking the columns of A on top of each other. As a shorthand, $\text{vec}(A_1, \dots, A_m)$ is equivalent to the more verbose $\text{vec}(\text{vec}(A_1), \dots, \text{vec}(A_m))$.

Theorem 1. *Under Assumptions 1 – 3 and the Markov assumption in Equation (9), there exists a weakly dependent stationary and ergodic solution, $X_t^* = \text{vec} \left((Y_t^*, \lambda_t^*, x_{t-1}^*) \right)$, to Equations (8). The solution is such that $\mathbb{E} (\|X_t^*\|^s) < \infty$, where $s \geq 1$ is found in Assumption 2, $Y_t^* = (Y_t^{*1}, \dots, Y_t^{*K})^\top$ and $\lambda_t^* = (\lambda_t^{*1}, \dots, \lambda_t^{*K})^\top$ are K -vectors, and $x_{t-1}^* = (x_{t-1}^{*1}, \dots, x_{t-1}^{*K})^\top$ is a $K \times r$ matrix.*

Proof. See the preprint [12].

A consequence of Theorem 1 is that it allows PoARX models to use the (weak) law of large numbers (LLN) for stationary and ergodic processes. To ensure the correct analysis of asymptotic behaviour, we need to be able to use the LLN for any initialisation, rather than a set of fixed initial values. Lemma 1 extends the LLN to hold for this case.

Lemma 1. *Let $X_t = \text{vec} \left((Y_t, \lambda_t, x_{t-1})^\top \right)$ be a process satisfying the equation $X_t = F(X_{t-1}; \xi_t)$ where ξ_t are i.i.d., $\mathbb{E} \|F(x; \xi_t) - F(\tilde{x}; \xi_t)\|^s \leq \kappa \|x - \tilde{x}\|^s$, and $\mathbb{E} \|F(0; \xi_t)\|^s < \infty$. For any function $h(x)$ satisfying:*

$$(i). \quad \|h(x)\|^{1+\delta} \leq M(1 + \|x\|^s) \text{ for some } M, \delta > 0,$$

(ii). for some $c > 0$ there exists $L_c > 0$ such that $\|h(x) - h(\tilde{x})\| \leq L_c \|x - \tilde{x}\|$ for $\|x - \tilde{x}\| < c$,

it holds that

$$\frac{1}{T} \sum_{t=1}^T h(X_t) \xrightarrow{P} \mathbb{E}(h(X_t^*)), \quad \text{as } T \rightarrow \infty.$$

Proof. See [21], or apply the main result from [22].

3 Estimation

We consider the model specified by Equations (8), where we denote the unknown parameters by ϑ . Then with $\alpha^j = (\alpha_1^j, \dots, \alpha_p^j)^\top$, $\beta^j = (\beta_1^j, \dots, \beta_q^j)^\top$, and $\eta^j = (\eta_1^j, \dots, \eta_r^j)^\top$,

$$\begin{aligned} \vartheta &= (\omega^1, (\alpha^1)^\top, (\beta^1)^\top, (\eta^1)^\top, \dots, \omega^K, (\alpha^K)^\top, (\beta^K)^\top, (\eta^K)^\top, \rho)^\top, \\ &= ((\theta^1)^\top, \dots, (\theta^K)^\top, \rho), \end{aligned}$$

where $\theta^j \in \Theta^j \subset [0, \infty)^{1+p+q+r}$.

The probability mass function of the copula PoARX model, derived from the cumulative mass function as rectangle probabilities is

$$\begin{aligned} &\Pr(Y_t^1 = y_t^1, \dots, Y_t^K = y_t^K) \\ &= \sum_{l_1=0}^1 \dots \sum_{l_K=0}^1 (-1)^{l_1 + \dots + l_K} C_\rho(F_1(y_t^1 - l_1; \lambda_t^1), \dots, F_K(y_t^K - l_K; \lambda_t^K)), \end{aligned}$$

with $C_\rho(\cdot)$ representing Frank's copula and $F_j(\cdot)$ the Poisson distribution function for $j = 1, \dots, K$. The conditional log-likelihood for ϑ given the multivariate observations y_1, \dots, y_n with initial values y_0 and λ_0 (denoted by the σ -field \mathcal{F}_0) is given by the following.

$$l(\vartheta) = \sum_{t=1}^n \log(\Pr((y_t^1, \dots, y_t^K)^\top | \mathcal{F}_{t-1}; \vartheta)) = \sum_{t=1}^n l_t(\vartheta).$$

With the large dimension of ϑ it is computationally more feasible to use a two-stage procedure known as the method of inference functions (IFM) [17]. We estimate the marginal parameters separately from the dependence parameter, hence reducing the dimension of the unknown parameters in each maximisation process. The marginal log-likelihood for θ^j is written as

$$l^j(\theta^j) = \sum_{t=1}^n \log(\Pr(y_t^j | \mathcal{F}_{t-1}; \theta^j)) = -\lambda_t^j + y_t^j \log(\lambda_t^j) - \log(y_t^j!), \quad (10)$$

with λ_t^j calculated using Equation (8c). Before we state the asymptotic result we impose two further conditions [1] on the parameters and the exogenous covariates.

Assumption 4 The space of possible parameters for each marginal distribution j , Θ^j , is compact for all $j = 1, \dots, K$. This means that for all $\theta^j = (\omega^j, \alpha^j, \beta^j, \eta^j) \in \Theta^j$, $\beta_i^j \leq \beta_i^{j,U}$, for each $i = 1, \dots, q$, and $\omega^j \geq \omega_L^j$ for some constants $\omega_L^j > 0$ and $\beta_i^{j,U} > 0$ with $\sum_{i=1}^q \beta_i^{j,U} < 1$.

Assumption 5 The polynomials $A^j(z) := \sum_{i=1}^p \alpha_{0,i}^j z^i$ and $B^j(z) := 1 - \sum_{i=1}^q \beta_{0,i}^j z^i$ have no common roots; and for any $a \neq 0$ and $g \neq 0$, $\sum_{i=1}^p a_i Y_{t-i}^{*j} + \sum_{i=1}^r g_i x_{i,t}^{*j}$ has a non-degenerate distribution. This should be true for each $j = 1, \dots, K$.

Theorem 2. *Suppose that Assumptions 1 – 5 hold with $s \geq 2$ and the true value of ϑ is denoted by ϑ_0 . Then ϑ is consistent and if $\vartheta \in \text{int } \Theta$,*

$$\sqrt{n}(\tilde{\vartheta} - \vartheta_0) \xrightarrow{d} \mathcal{N}(0, V), \tag{11}$$

where V is a valid covariance matrix.

Proof. See preprint [12]. Details of V can be found there.

4 Forecasting

Forecasting with PoARX models is to some extent similar to the forecasting of GARCH-X processes [13]. Predictions for the intensities can be obtained recursively using Equation (8c) and the property $E(Y_t^j | \mathcal{F}_{t-1}) = \lambda_t^j$. This procedure also gives point predictions for the process. However, there is substantial difference when predictive distributions are required.

One-step ahead forecasts at time t of the intensities $\lambda_{t+1}^j, \dots, \lambda_{t+h-1}^j$, given information \mathcal{F}_t , parameters θ^j , and covariates x_t results in a known value of $\lambda_{t+1}^j | \mathcal{F}_t$ using Equation (8c). By the specifications of the model, the one-step ahead marginal predictive distributions are Poisson with predicted intensities $\lambda_{t+1}^j | \mathcal{F}_t$. The joint predictive distribution is obtained by substituting the predicted intensities in Equation (7).

For multi-step-ahead forecasts, the procedure is not so straightforward. Firstly, the computation of the h -step-ahead forecast at time t assumes that the exogenous covariates x_t, \dots, x_{t+h-1} are known. In practice, these will often need to be replaced by their own forecasts or projections. This is not a problem when the covariates are leading indicators, see the example in Section 5. With a slight abuse of notation we use $\lambda_{t+h}^j | \mathcal{F}_t$ to represent the “intensity for horizon h conditional on \mathcal{F}_t and x_t, \dots, x_{t+h-1} ”. We let this knowledge be denoted by the σ -field \mathcal{G}_t . We show below that the predictive distributions for $h \geq 2$ are not necessarily Poisson using a characteristic function-type approach [4]. Since the Poisson distribution is discrete, we use conditional probability generating functions.

The probability generating functions can be calculated as follows, starting with $h = 2$. For a time series Y_t following a PoARX process with intensity λ_t ,

we can write $\lambda_{t+2|t} = c_{t+2} + \alpha_1 y_{t+1}$, where c_{t+2} is measurable w.r.t. \mathcal{G}_t . In the derivation below we will need the following result:

$$\begin{aligned} \mathbb{E}(\exp((-1+z)\alpha_1 y_{t+1}) | \mathcal{G}_t) &= \sum_{k=0}^{\infty} \frac{\lambda_{t+1}^k}{k!} \exp(-\lambda_{t+1}) \exp((-1+z)\alpha_1 k) \\ &= \exp\left(\lambda_{t+1}(-1 + e^{(-1+z)\alpha_1})\right). \end{aligned}$$

The 2-step ahead forecast has the following generating function ($P_2(z)$ depends also on t but we omit that to keep the notation transparent):

$$\begin{aligned} P_2(z) &= \mathbb{E}(z^{Y_{t+2}} | \mathcal{G}_t) = \mathbb{E}(\mathbb{E}(z^{Y_{t+2}} | \mathcal{G}_{t+1}) | \mathcal{G}_t) \\ &= \mathbb{E}(\exp((-1+z)\lambda_{t+2}) | \mathcal{G}_t) \\ &= \exp((-1+z)c_{t+2}) \mathbb{E}(\exp((-1+z)\alpha_1 y_{t+1}) | \mathcal{G}_t) \\ &= \begin{cases} \exp((-1+z)c_{t+2}) & \text{if } \alpha_1 = 0, \\ \exp((-1+z)c_{t+2}) \exp(\lambda_{t+1}(-1 + \exp(-1+z)\alpha_1)) & \text{if } \alpha_1 \neq 0. \end{cases} \end{aligned}$$

We can see that if $\alpha_1 \neq 0$, then $P_2(z)$ is not Poisson, by the uniqueness property of generating functions. The joint distribution can be obtained by computing analogously the joint probability generating functions.

For $h > 2$ the above calculation can be extended by repeatedly using the property of the iterated conditional expectation. It can also be expressed recursively as follows:

$$\begin{aligned} P_h(z) &= \mathbb{E}(z^{Y_{t+h}} | \mathcal{G}_t) = \mathbb{E}(\mathbb{E}(z^{Y_{t+h}} | \mathcal{G}_{t+1}) | \mathcal{G}_t) \\ &= \mathbb{E}(P_{h-1}(z) | \mathcal{G}_t). \end{aligned}$$

Clearly, for $h \geq 2$ the forecast distribution is not necessarily Poisson. Nevertheless, using iterated conditional expectations we have that

$$\begin{aligned} \mathbb{E}(Y_{t+h} | \mathcal{G}_t) &= \mathbb{E}(\mathbb{E}(Y_{t+h} | \mathcal{G}_{t+h-1}) | \mathcal{G}_t) \\ &= \mathbb{E}(\lambda_{t+h} | \mathcal{G}_t). \end{aligned}$$

Therefore, we can generate h -step ahead forecast of the intensity with the following equation,

$$\lambda_{t+h|t} = \omega + \sum_{l=1}^p \alpha_l Y_{t+h-l|t} + \sum_{l=1}^q \beta_l \lambda_{t+h-l|t} + \eta \cdot x_{t+h-1}. \quad (12)$$

where

$$Y_{t+k|t} = \begin{cases} \lambda_{t+k|t} & \text{if } k > 0, \\ y_{t+k} & \text{if } k \leq 0. \end{cases}$$

Prediction intervals can be obtained by computing the probabilities from the probability generating functions discussed above. Since these are probably feasible only for small horizons, simulation would be a more practical alternative.

To obtain a prediction interval for Y_{t+h}^j , simulate a trajectory of the PoARX time series until time $t + h$, resulting in one simulated value Y_{t+h}^j . Repeating this process B times allows access to the quantiles from which we can obtain a prediction interval for the time series. Simulating a joint predictive region is an area for further work and not discussed here.

5 Applications

We illustrate the use of PoARX models with a data set originally used for event detection [14]. The computations were done with R [24] using the implementation of the PoARX models in package PoARX [11].

5.1 Data

The data contains counts of the estimated number of people that entered and exited a building over thirty-minute intervals of a UCI campus building. Counts were recorded by an optical sensor at the front door starting from the end of 23/07/2005 until the end of 05/11/2005. The data has periodic tendencies but is also influenced by events within the building causing an influx of traffic. Originally, the data was used to build a novel event detection framework under a Bayesian scheme.

We will estimate the number of people entering ($N^I(t)$) and exiting ($N^O(t)$) the building using the Poisson distribution, as found in its original use [14]. The basis of model predictions will be the lagged values of the observations and mean value, as well as some exogenous covariates. These covariates are all indicator variables, representing the following. The first is a “weekday” indicator, that takes value 1 when the day is Monday–Friday. This corresponds to an uplift for working days. The second indicator is a “daytime” indicator, taking value 1 when the time is between 07:30 and 19:30, representing an uplift in the traffic during working hours. The third indicator is associated with the presence of an event occurring. For the flow count into the building, the variable takes the value 1 when an event will occur in the next hour. For the flow out of the building, the variable takes the value 1 in the hour after an event finished. These represent the arrival and departure of people coming to the building for the event. We will investigate whether the use of Frank’s copula improves the predictions.

5.2 Estimation and prediction

We fit four types of models to the data in an attempt to find the best predictive model. Models 1 and 2 contain no covariates whilst Models 3 and 4 contain all covariates. Additionally, Models 1 and 3 are modelled using independent PoARX distributions whilst Models 2 and 4 fit the joint distribution using Frank’s copula. To assess the quality, we used the log score [2] and training was implemented using 5-fold cross validation [26]. We also removed some observations to use as an independent test set.

For analysis, the lagged values chosen differed slightly for each time series. For the number of people entering the building ($N^I(t)$), we chose to use 4 lagged values for the observations (lags 1, 2, 48, 336) and 1 lagged value for the means (lag 1). Lagged values from the previous 2 observations represent the flow of people within the last hour, whilst the lag of 48 corresponds to the same time point on the previous day, and 336 to the same time point on the same day in the previous week. For $N^O(t)$ we used the same 4 lagged values for the observations (lags 1, 2, 48, 336) but included an extra lag for the mean values (lags 1, 48). These were chosen based on the cross-validated log scores.

In Table 1 we present the cross-validated (training) log score and the log score from the independent (test) data for each of the four models.

Table 1: Model training scores from cross-validated fit on 4000 observations

Model number	Training log score	Test log score
1	-15444	-4184
2	-15411	-4182
3	-25088	-4190
4	-16856	-4164

Looking solely at the training score, we notice that Model 2 appears to be the best model, while Model 1 is second. It seems as though the addition of the covariates weakens the fit of the model, despite the parameters of the relevant models being significantly greater than zero, statistically speaking. Furthermore, using this metric, we deduce that the use of Frank's copula improves the predictions compared to those using the independence assumption. The smallest score and therefore the worst performance is found in the results from Model 3. This model contains covariates along with the independence assumption.

However, since we are building a predictive model we also assess the predictive capabilities on data not used in the training process, to combat any overfitting. On the external data it seems as though the first three models have similar scores whilst Model 4 produces a significant improvement in predictive accuracy. This would suggest that the combination of the time series aspects, the covariates and the multivariate modelling produces the most accurate out-of-sample predictions for this kind of data. Decomposing the improvements, we start with Models 1 and 2. There is a very small increase in performance by removing the independence assumption and using Frank's copula, but perhaps this is not worth the extra complexity gained from using a copula model. From Model 1 it appears as though adding covariates does not increase either the model fit or the predictive accuracy. As mentioned earlier, one reason for this could be the violation of the assumption of independence due to the common covariates. It is only with the combination of the covariates and the copula model that we find the best performance. To the best of our knowledge, no tests for probabilistic forecast performance exist. Formally showing that the forecasts from Model 4

are best is an area of further work, with tests such as the Diebold–Mariano test not applicable for nested models [5] or probabilistic forecasts.

6 Conclusion

We introduced the multivariate PoARX model as an extension of the univariate PoARX model. Using previously established properties of the univariate PoARX model and copulas, we stated results for stability and asymptotic normality of estimates obtained via the method of inference functions [17]. Our discussion on forecasting, especially predictive distributions for horizons larger than one, seems novel even for univariate PoARX models. In particular, it is important to point out that the predictive distributions for lags greater than one are not Poisson.

In the example in Section 5 we illustrated the use of bivariate PoARX models for modelling the counts of the number of people entering and exiting a building, using lagged values and covariates. The results of adding covariates and a copula structure separately did not suggest that either addition significantly improved the model, but the combination of time series covariates, exogenous covariates, and the dependence structure produced the best predictions. However, we acknowledge that the log score was an arbitrary choice and that any strictly proper scoring rule would have sufficed (see [10] for more details). We feel that the analysis provides material for further thought and work on model evaluation for count data time series models.

There is much scope for further work in this area. The choice of Frank’s copula was made for its ability to model a negative dependence in two dimensions, but any copula could be used. Another suggestion for change would be to consider distributions other than Poisson. We are considering the possibility of using the renewal count distributions [20] implemented in the R package Rcount [19]. Combining these renewal distributions with the ideas found in this paper could lead to a fascinating new family of count time series models.

References

1. A. Agosto, G. Cavaliere, D. Kristensen, and A. Rahbek. Modelling corporate defaults: Poisson autoregression with exogenous covariates (PARX). *Journal of Empirical Finance*, 38:640 – 663, 2016.
2. J. E. Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49 – 65, 2007.
3. T. Bollerslev. Generalised autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307 – 327, 1986.
4. G. N. Boshnakov. Analytic expressions for predictive distributions in mixture autoregressive models. *Statistical & Probability Letters*, 79(15):1704–1709, 2009.
5. F. X. Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1), 2015.

6. R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987 – 1008, 1982.
7. R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH processes. *Journal of Time Series Analysis*, 27(6):923 – 942, 2006.
8. K. Fokianos, A. Rahbek, and D. Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430 – 1439, 2009.
9. C. Genest and J. G. Nešlehová. A primer on copulas for discrete data. *The ASTIN Bulletin*, 37:475 – 515, 2007.
10. T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
11. J. Halliday and G. N. Boshnakov. *PoARX: Fit PoARX models to multivariate time series*, 2018. R package version 0.3.2 (under development, to be published on CRAN).
12. J. Halliday and G. N. Boshnakov. PoARX modelling for multivariate count time series. *eprint arXiv:1806.04892*, 2018.
13. P. R. Hansen, Z. Huang, and H. H. Shek. Realised GARCH: A joint model for returns and realised measures of volatility. *Journal of Applied Econometrics*, 27:877 – 906, 2012.
14. A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying Poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*, pages 207 – 216. ACM Press, 2006.
15. D. I. Inouye, E. Yang, G. I. Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3), 2017.
16. H. Joe. *Multivariate models and dependence concepts*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, 1997.
17. H. Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401 – 419, 2005.
18. H. Joe. *Dependence Modeling with Copulas*. New York: Chapman and Hall/CRC., 2014.
19. T. Kharrat and G. N. Boshnakov. *Countr: Flexible univariate count models based on renewal processes*, 2016. R package version 3.2.8.
20. T. Kharrat, G. N. Boshnakov, I. G. McHale, and R. Baker. Flexible regression models for count data based on renewal processes: The Countr package (to appear). *Journal of Statistical Software*, 2018.
21. D. Kristensen and A. Rahbek. Quasi-likelihood estimation of multivariate GARCH models: A weak dependence approach. Working Papers, 2015.
22. A. M. Lindner and A. Szimayer. A limit theorem for copulas, 2005. urn:nbn:de:bvb:19-epub-1802-0.
23. R. B. Nelsen. *An Introduction to Copulas*. New York: Springer, Second edition, 2006.
24. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
25. A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229 – 231, 1959.
26. M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(2):111 – 147, 1974.
27. P. K. Trivedi and D. Zimmer. A note on identification of bivariate copulas for discrete count data. *Econometrics*, 5(10), 2017.

Gaussian Variational Bayes Kalman Filtering for Dynamic Sparse Bayesian Learning

Christo Kurisummoottil Thomas* and Dirk Slock

EURECOM, Sophia-Antipolis, France
{kurisumm, slock}@eurecom.fr
<http://www.eurecom.fr/cm/>

Abstract. Sparse Bayesian Learning (SBL) provides sophisticated (state) model order selection with unknown support distribution. This allows to handle problems with big state dimensions and relatively limited data. The techniques proposed in this paper allow to handle the extension of SBL to time-varying states, modeled as diagonal first-order vector autoregressive (VAR(1)) processes with unknown parameters. Adding the parameters to the state leads to an augmented state and a non-linear (at least bilinear) state-space model. The proposed approach, which applies also to more general non-linear models, uses Variational Bayes (VB) techniques to approximate the posterior distribution by a factored form, with Gaussian or exponential factors. The granularity of the factorization can take on various levels. In one extreme instance, called Gaussian Space Alternating Variational Estimation Kalman Filtering (GSAVE-KF), all state components are treated individually, leading to low complexity filtering. Simulations illustrate the performance of the proposed GVB-KF techniques, which represent an alternative to Linear MMSE (LMMSE) filtering.

Keywords: Sparse Bayesian Learning, Variational Bayes, Kalman Filtering

1 Introduction

Sparse signal reconstruction and compressed sensing (CS) has received significant attraction in the recent years. The signal model for the recovery of a time varying sparse signal can be formulated as,

$$\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t, \quad (1)$$

where \mathbf{y}_t is the observations or data at time t , \mathbf{A}_t is called the measurement or the sensing matrix which is known and is of dimension $N \times M$ with $N < M$, \mathbf{x}_t is the M -dimensional sparse signal and \mathbf{v}_t is the additive noise. \mathbf{x}_t contains only K non-zero entries, with $K \ll M$ and is modeled by a diagonal AR(1) (autoregressive) process. \mathbf{v}_t is assumed to be a white Gaussian noise, $\mathbf{v}_t \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$.

* EURECOM's research is partially supported by its industrial members: ORANGE, BMW, ST Microelectronics, Symantec, SAP, Monaco Telecom, iABG, and by the projects DUPLEX (French ANR), MASS-START and GEOLoc (French FUI).

In the time invariant case, to address this problem, there exists a variety of algorithms such as the basis pursuit method [1] and the orthogonal matching pursuit [2]. In Bayesian learning, sparse Bayesian learning (SBL) algorithm was first proposed by [3, 4]. Performance can be further improved by exploiting the temporal correlation across the sparse vectors [5]. However, most of these algorithms do offline or batch processing, whose complexity doesn't scale with the problem size. In order to render low complexity or low latency solutions, online processing algorithms (which processes small set of measurement vectors at any time) will be necessary.

In sparse adaptive estimation [6], a time varying signal \mathbf{x}_t is estimated time-recursively by exploiting the sparsity property of the signal. Conventional adaptive filtering methods such as LMS or recursive least squares (RLS) doesn't exploit the underlying sparseness in the signal \mathbf{x}_t to improve the estimation performance. Compared to the state of the art, we introduce not only sparse filter (state) but also sparse filter variations. We apply SBL now to the prediction error variances of \mathbf{x}_t , then trying to sparsify a prediction error variance actually encourage both that the actual variance gets sparse and that the variation gets sparse because a prediction error variance is small if either the quantity variance is small or its variation is small.

In the literature, there exist different KF based methods to handle the joint filtering and parameter estimation problem. One such example is the widely used EM-KF algorithm ([7, 8]) which uses the famous Expectation Maximization technique (EM), and alternating optimization technique for ML estimation. To handle general nonlinear state space models, another variation called as Extended KF (EKF) algorithm exists. In this case, the state is extended with the unknown parameters, rendering the new state update equation nonlinear. A third derivation is the truncated Second-Order EKF (SOEKF) introduced by [9, 10] in which nonlinearities are expanded up to second order, third and higher order statistics being neglected. [11] present a corrected derivation of SOEKF and show that the state of the art contains illogical approximations. In ([10, 12]), the Gaussian SOEKF is derived in which fourth-order terms in the Taylor series expansions are retained and approximated by assuming that the underlying joint probability distribution is Gaussian. In [13], Villares et al. introduced the Quadratic Extended Kalman Filter (QEKF). The authors extend the EKF to deal with quadratic signal models and exploiting the fourth order signal statistics. We proposed a space alternating variational estimation based technique for single measurement vectors in [14].

1.1 Contributions of this paper

- We propose a novel Gaussian approximation Space Alternating Variational Estimation (GSAVE) based SBL technique for LMMSE filtering called GSAVE-KF. The proposed solution is for a multiple measurement case with an AR(1) process for the temporal correlation of the sparse signal. The update and prediction stages of the proposed algorithm reveals links to the Kalman filter.

- For the static state case, numerical results presented elsewhere [14] suggest that the proposed solution has a faster convergence rate (and hence lower complexity) than (even) the existing fast SBL and performs better than the existing fast SBL algorithms in terms of reconstruction error in the presence of noise.
- For the dynamic state case considered here, simulations suggest that in spite of both significantly reduced computational complexity and the estimation of the unknown (hyper) parameters, the GSAVE-KF algorithm exhibits hardly any MSE degradation in steady-state compared to the standard Kalman filter with known parameters, but at the cost of a significantly increased transient duration.

In the following, boldface lower-case and upper-case characters denote vectors and matrices respectively. the operators $tr(\cdot)$, $(\cdot)^T$ represents trace, and transpose respectively. The operator $(\cdot)^H$ represents the conjugate transpose or conjugate for a matrix or a scalar respectively. A complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$ is distributed as $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Theta})$. $diag(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $\langle x \rangle$ or $E(\cdot)$ represents the expectation of x . $\|\cdot\|$ represents the Frobenius norm. $\Re\{\cdot\}$ represents the real part of (\cdot) . All the variables are complex here unless specified otherwise.

2 State Space Model

Sparse signal \mathbf{x}_t is modeled using an AR(1) process with correlation coefficient matrix \mathbf{F} , with \mathbf{F} diagonal. The state space model can be written as follows,

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, && \text{State Update,} \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, && \text{Observation,} \end{aligned} \tag{2}$$

where $\mathbf{x}_t = [x_{1,t}, \dots, x_{M,t}]^T$. Matrices \mathbf{F} and $\boldsymbol{\Gamma}$ are defined as,

$$\mathbf{F} = \begin{bmatrix} f_1 & 0 & \dots & 0 \\ 0 & f_2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_M \end{bmatrix}, \boldsymbol{\Gamma} = \begin{bmatrix} \frac{1}{\sqrt{\alpha_1}} & \dots & 0 \\ 0 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sqrt{\alpha_M}} \end{bmatrix}, \tag{3}$$

Here f_i represents the correlation coefficient and α_i represents the inverse variance of $x_{i,t} \sim \mathcal{CN}(0, \frac{1}{\alpha_i})$. Further, $\mathbf{w}_t \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^H))$ and $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\gamma}\mathbf{I})$. \mathbf{w}_t are the complex Gaussian mutually uncorrelated state innovation sequences. \mathbf{v}_t is independent of the \mathbf{w}_t process. Further we define, $\mathbf{A} = \boldsymbol{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^H) = \text{diag}(\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_M})$.

Although the above signal model seems simple, there are numerous applications such as 1) Bayesian adaptive filtering [15, 16], 2) Wireless channel estimation: multi-path parameter estimation as in [17, 18]. In this case, $\mathbf{x}_t =$ FIR filter response, and $\boldsymbol{\Gamma}$ represents e.g. the power delay profile.

3 VB-SBL

In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the \mathbf{x} as in [3]. The hierarchical prior is such that it encourages the sparsity property of \mathbf{x}_t or of the innovation sequences \mathbf{v}_t .

$$\begin{aligned} p(\mathbf{x}_t/\Gamma) &= \prod_{i=1}^M p(x_{i,t}/\alpha_i) = \prod_{i=1}^M \mathcal{N}(0, \alpha_i^{-1}), \\ p(\mathbf{x}_t/\mathbf{x}_{t-1}, \mathbf{F}, \Gamma) &= \prod_{i=1}^M p(x_{i,t}/x_{i,t-1}, \alpha_i, f_i) = \prod_{i=1}^M \mathcal{N}(f_i x_{i,t-1}, \frac{1}{\alpha_i}). \end{aligned} \quad (4)$$

For the convenience of analysis, we reparameterize α_i in terms of λ_i and assume a Gamma prior for Λ ,

$$p(\Lambda) = \prod_{i=1}^M p(\lambda_i/a, b) = \prod_{i=1}^M \Gamma^{-1}(a) b^a \lambda_i^{a-1} e^{-b\lambda_i}. \quad (5)$$

The inverse of noise variance γ is also assumed to have a Gamma prior,

$$p(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma^{c-1} e^{-d\gamma}. \quad (6)$$

Now the likelihood distribution can be written as,

$$p(\mathbf{y}_t/\mathbf{x}_t, \gamma) = (2\pi)^{-N} \gamma^N e^{-\frac{\gamma \|\mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t\|^2}{2}}. \quad (7)$$

To make these priors non-informative, we choose them to be small values $a = c = b = d = 10^{-5}$. For the AR(1) coefficients f_k , we don't assume any prior distribution.

3.1 Variational Bayesian Inference

The computation of the posterior distribution of the parameters is usually intractable. In order to address this issue, in variational Bayesian framework, the posterior distribution $p(\mathbf{x}_t, \Lambda, \gamma/\mathbf{y}_{1:t})$ is approximated by a variational distribution $q(\mathbf{x}_t, \Lambda, \gamma)$ that has the factorized form:

$$q(\mathbf{x}_t, \Lambda, \gamma) = q_\gamma(\gamma) \prod_{i=1}^M q_{x_{i,t}}(x_{i,t}) \prod_{i=1}^M q_{\lambda_i}(\lambda_i), \quad (8)$$

where $\mathbf{y}_{1:t}$ represents the observations till the time t ($\mathbf{y}_1, \dots, \mathbf{y}_t$), similarly we define $\mathbf{x}_{1:t}$. Variational Bayes compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{x}_t, \Lambda, \gamma/\mathbf{y}_{1:t})$ and the $q(\mathbf{x}_t, \Lambda, \gamma)$. From [19],

$$KLD_{VB} = KL(p(\mathbf{x}_t, \Lambda, \gamma/\mathbf{y}_{1:t}) || q(\mathbf{x}_t, \Lambda, \gamma)) \quad (9)$$

The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO) [19]. To elaborate on this, we can write the marginal probability of the observed data as,

$$\begin{aligned} \ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1}) &= L(q) + KLD_{VB}, \text{ where,} \\ L(q) &= \int q(\mathbf{x}_t, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1})}{q(\boldsymbol{\theta})} d\mathbf{x}_t d\boldsymbol{\theta}, \\ KLD_{VB} &= - \int q(\mathbf{x}_t, \boldsymbol{\theta}) \ln \frac{p(\mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t})}{q(\mathbf{x}_t, \boldsymbol{\theta})} d\mathbf{x}_t d\boldsymbol{\theta}, \end{aligned} \tag{10}$$

where $\boldsymbol{\theta} = \{\mathbf{A}, \gamma\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$. Since $KLD_{VB} \geq 0$, it implies that $L(q)$ is a lower bound on $\ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1})$. Moreover, $\ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1})$ is independent of $q(\mathbf{x}_t, \boldsymbol{\theta})$ and therefore maximizing $L(q)$ is equivalent to minimizing KLD_{VB} . This is called as ELBO maximization and doing this in an alternating fashion for each variable in $\mathbf{x}_t, \boldsymbol{\theta}$ leads to,

$$\begin{aligned} \ln(q_i(\theta_i)) &= \langle \ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) \rangle_{\boldsymbol{\theta}_{\bar{i}}, \mathbf{x}_t} + c_i, \\ \ln(q_i(x_{i,t})) &= \langle \ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) \rangle_{\boldsymbol{\theta}, \mathbf{x}_{\bar{i},t}} + c_i, \\ p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) &= p(\mathbf{y}_t/\mathbf{x}_t, \gamma, \mathbf{y}_{1:t-1}) p(\mathbf{x}_t/\mathbf{A}, \mathbf{y}_{1:t-1}) p(\mathbf{A}) p(\gamma). \end{aligned} \tag{11}$$

Here $\langle \rangle_{k \neq i}$ represents the expectation operator over the distributions q_k for all $k \neq i$. $\mathbf{x}_{\bar{i},t}$ represents \mathbf{x}_t without x_i and $\boldsymbol{\theta}_{\bar{i}}$ represents $\boldsymbol{\theta}$ without θ_i . In section 5, we consider another variant where the components of \mathbf{x}_t are treated jointly, where the approximate posterior becomes $q(\mathbf{x}_t, \mathbf{A}, \gamma) = q_\gamma(\gamma) q_{\mathbf{x}_t}(\mathbf{x}_t) \prod_{i=1}^M q_{\lambda_i}(\lambda_i)$.

3.2 Gaussian Posterior Minimizing the KL Divergence

In [20], for any distribution $p(\mathbf{x})$, the Gaussian distribution $q(\mathbf{x}) \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which minimizes the Kullback-Leibler divergence, $KL(p||q)$, reduces to matching the mean and covariance,

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle_{p(x)}, \quad \boldsymbol{\Sigma} = \langle \mathbf{x}\mathbf{x}^H \rangle_{p(x)} - \langle \mathbf{x} \rangle_{p(x)} \langle \mathbf{x} \rangle_{p(x)}^H. \tag{12}$$

4 SAVE Sparse Bayesian Learning and Kalman Filtering

In this section, we propose a Space Alternating Variational Estimation (SAVE) based alternating optimization between each element of \mathbf{x}_t or γ . For SAVE, no particular structure of \mathbf{A}_t is assumed, in contrast to AMP which performs poorly when \mathbf{A}_t is not i.i.d or is sub-Gaussian. The joint distribution w.r.t the observation of (2) can be written as,

$$p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = p(\mathbf{y}_t/\mathbf{x}_t, \boldsymbol{\theta}) p(\mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}). \tag{13}$$

In the following, $c_{x_{k,t}}, c'_{x_{k,t}}, c_{\alpha_k}, c_{\lambda_k}, c_{x-1}, c_{x_t}, c'_{x_t}$ and c_γ represents normalization constants for the respective pdfs.

4.1 Diagonal AR(1) (DAR(1)) Prediction Stage

In this stage, we compute the prediction about \mathbf{x}_t given the observations till time $t - 1, \hat{x}_{k,t|t-1}$. The prediction about \mathbf{x}_t can be computed from the time update equation of the standard Kalman filter,

$$x_{k,t} = \hat{f}_{k|t-1} x_{k,t-1} + \tilde{f}_{k|t-1} x_{k,t-1} + w_{k,t}. \tag{14}$$

Here we denote $\widehat{f}_{k|t-1}$ as the estimate of f_k given the observations till $t-1$ and $\widetilde{f}_{k|t-1}$ represents the error in the estimation. Similarly we can represent $x_{k,t-1} = \widehat{x}_{k,t-1|t-1} + \widetilde{x}_{k,t-1|t-1}$, $\widetilde{x}_{k,t-1|t-1}$ being the estimation error.

$$\begin{aligned} \widehat{x}_{k,t|t-1} &= \widehat{f}_{k|t-1} \widehat{x}_{k,t-1|t-1}, \quad \widetilde{x}_{k,t|t-1} = \widehat{f}_{k|t-1} \widetilde{x}_{k,t-1|t-1} + \widetilde{f}_{k|t-1} x_{k,t-1} + w_{k,t}, \\ \implies \sigma_{k,t|t-1}^2 &\stackrel{(a)}{=} |\widehat{f}_{k|t-1}|^2 \sigma_{k,t-1|t-1}^2 + \sigma_{f_k}^2 (|\widehat{x}_{k,t-1|t-1}|^2 + \sigma_{k,t-1|t-1}^2) + \frac{1}{\widehat{\lambda}_{k|t-1}}, \end{aligned} \quad (15)$$

In the variational approximation, we assume that the posterior of f_k and $x_{k,t}$ are independent. (a) in (15) follows from this argument. Further the predictive distribution $p(\mathbf{x}_t/\mathbf{y}_{1:t-1})$ can be approximated to be Gaussian distributed (refer to the discussion in section 3.2) with mean $\widehat{\mathbf{x}}_{t|t-1} = [\widehat{x}_{1,t|t-1}, \dots, \widehat{x}_{M,t|t-1}]^T$ and diagonal error covariance $\widehat{\mathbf{P}}_{t|t-1} = \text{diag}(\sigma_{1,t|t-1}^2, \dots, \sigma_{M,t|t-1}^2)$. Actually this parametric $q(\cdot)$ fitting, we only need it for the prediction stage of \mathbf{x}_t , all other q 's (filtering or smoothing of \mathbf{x}_t , all hyper parameters) come out simple due to the choice of conjugate priors. Further the joint distribution in (13) can be obtained as,

$$\begin{aligned} \ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) &= N \ln \gamma - \gamma \|\mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t\|^2 - M \ln \det(\widehat{\mathbf{P}}_{t|t-1}) - \\ &(\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t-1})^H \widehat{\mathbf{P}}_{t|t-1}^{-1} (\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t-1}) + (c-1) \ln \gamma + c \ln d - d\gamma + \text{constants}, \end{aligned} \quad (16)$$

4.2 Measurement or Update Stage

Update of $q_{x_{k,t}}(x_{k,t})$: Using (11), $\ln q_{x_{k,t}}(x_{k,t})$ turns out to be quadratic in $x_{k,t}$ and thus can be represented as a Gaussian distribution as follows,

$$\begin{aligned} \ln q_{x_{k,t}}(x_{k,t}) &= -\langle \gamma \rangle \left\{ (\mathbf{y}_t - \mathbf{A}_{t,\bar{k}} \langle \mathbf{x}_{\bar{k},t} \rangle)^H \mathbf{A}_{t,k} x_{k,t} - x_{k,t}^H \mathbf{A}_{t,k}^H \right. \\ &(\mathbf{y}_t - \mathbf{A}_{t,\bar{k}} \langle \mathbf{x}_{\bar{k},t} \rangle) + \|\mathbf{A}_{t,k}\|^2 |x_{k,t}|^2 \left. \right\} - \frac{1}{\sigma_{k,t|t-1}^2} \left(|x_{k,t}|^2 - x_{k,t}^H \widehat{x}_{k,t|t-1} - \right. \\ &x_{k,t} \widehat{x}_{k,t|t-1}^H \left. \right) + c_{x_{k,t}} = -\frac{1}{\sigma_{k,t|t}^2} |x_{k,t} - \widehat{x}_{k,t|t}|^2 + c'_{x_{k,t}}. \end{aligned} \quad (17)$$

Note that we split $\mathbf{A}_t \mathbf{x}_t$ as, $\mathbf{A}_t \mathbf{x}_t = \mathbf{A}_{t,k} x_{k,t} + \mathbf{A}_{t,\bar{k}} \mathbf{x}_{\bar{k},t}$, where $\mathbf{A}_{t,k}$ represents the k^{th} column of \mathbf{A}_t , $\mathbf{A}_{t,\bar{k}}$ represents the matrix with k^{th} column of \mathbf{A}_t removed. Clearly, the mean and the variance of the resulting Gaussian distribution becomes,

$$\begin{aligned} \sigma_{k,t|t}^{-2,(i)} &= \langle \gamma \rangle \|\mathbf{A}_{t,k}\|^2 + \sigma_{k,t|t}^{-2,(i-1)}, \\ \langle x_{k,t|t}^{(i)} \rangle &= \sigma_{k,t|t}^{2,(i)} \left(\mathbf{A}_{t,k}^H (\mathbf{y}_t - \mathbf{A}_{t,\bar{k}} \langle \mathbf{x}_{\bar{k},t}^{(i-1)} \rangle) \langle \gamma \rangle + \frac{\widehat{x}_{k,t|t-1}}{\sigma_{k,t|t-1}^2} \right), \end{aligned} \quad (18)$$

where i represents the iteration stage with $\lim_{i \rightarrow \infty} \langle x_{k,t|t}^{(i)} \rangle = \widehat{x}_{k,t|t}$ represents the point estimate of $x_{k,t}$. However, in (18) the computation of $\langle x_{k,t|t}^{(i)} \rangle$ requires the knowledge of $\langle \mathbf{x}_{\bar{k},t}^{(i)} \rangle$. So we need to perform enough iterations between the components of $\langle x_{k,t|t} \rangle$ till convergence. Moreover, we initialize $\langle x_{k,t|t}^{(0)} \rangle$

by $\hat{x}_{k,t|t-1}$ and $\sigma_{k,t}^{-2,(0)} = \sigma_{k,t|t-1}^{-2}$, which is obtained in the prediction stage. One remark is that forcing a Gaussian posterior q with diagonal covariance matrix on the original Kalman measurement equations gives the same result as SAVE. Note that the derivations in [21] for VB-KF are not correct as it does not have the correct variance expressions that vary with iteration! For the convenience of the derivations in the following sections, we define $\hat{\mathbf{P}}_{t|t} = \text{diag}(\sigma_{1,t|t}^2, \dots, \sigma_{M,t|t}^2)$, $\hat{\mathbf{x}}_{t|t} = [\hat{x}_{1,t|t}, \dots, \hat{x}_{M,t|t}]^T$.

4.3 Fixed Lag Smoothing

Kalman filtering in the EM-KF is not enough to adapt the hyper parameters, instead we need atleast a lag 1 smoothing [22]. Motivated by this result, we propose fixed lag smoothing with delay 1 for SAVE-KF. We rewrite the state space model as follows,

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}_t \mathbf{F} \mathbf{x}_{t-1} + \underbrace{\mathbf{A}_t \mathbf{w}_{t-1}}_{\tilde{\mathbf{v}}_t} + \mathbf{v}_t, \\ p(\mathbf{y}_t, \mathbf{x}_{t-1}, \boldsymbol{\theta} | \mathbf{y}_{1:t-1}) &= p(\mathbf{y}_t | \mathbf{x}_{t-1}, \boldsymbol{\theta}) p(\mathbf{x}_{t-1}, \boldsymbol{\theta} | \mathbf{y}_{1:t-1}), \end{aligned} \quad (19)$$

where $\tilde{\mathbf{v}}_t \sim \mathcal{CN}(\mathbf{0}, \tilde{\mathbf{R}}_t)$, $\tilde{\mathbf{R}}_t = \mathbf{A}_t \boldsymbol{\Lambda} \mathbf{A}_t^H + \frac{1}{\gamma} \mathbf{I}$. The posterior distribution $p(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1})$ is approximated using variational approximation as $q(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1})$ with mean and covariance as $\hat{\mathbf{x}}_{t-1|t-1}$ and $\hat{\mathbf{P}}_{t-1|t-1}$.

$$\begin{aligned} \ln p(\mathbf{y}_t, \mathbf{x}_{t-1}, \boldsymbol{\theta} | \mathbf{y}_{1:t-1}) &= \frac{-1}{2} \ln \det \tilde{\mathbf{R}}_t - \\ &(\mathbf{y}_t - \mathbf{A}_{t,k} f_k x_{k,t-1} - \mathbf{A}_{t,\bar{k}} \mathbf{F}_{\bar{k}} \mathbf{x}_{\bar{k},t-1})^H \tilde{\mathbf{R}}_t^{-1} (\mathbf{y}_t - \mathbf{A}_{t,k} f_k x_{k,t-1} - \mathbf{A}_{t,\bar{k}} \mathbf{F}_{\bar{k}} \mathbf{x}_{\bar{k},t-1}) \\ &- \frac{1}{2} \det(\hat{\mathbf{P}}_{t-1|t-1}) - (\mathbf{x}_{t-1} - \hat{\mathbf{x}}_{t-1|t-1})^H \hat{\mathbf{P}}_{t-1|t-1}^{-1} (\mathbf{x}_{t-1} - \hat{\mathbf{x}}_{t-1|t-1}) + c_{x-1}, \end{aligned} \quad (20)$$

where $\mathbf{F}_{\bar{k}}$ represents \mathbf{F} with k^{th} column and row removed.

Prediction of \mathbf{x}_{t-1} : Using (11), $\ln q_{\mathbf{x}_{t-1}}(\mathbf{x}_{t-1} | \mathbf{y}_{1:t})$ turns out to be quadratic in \mathbf{x}_{t-1} and thus can be represented as a Gaussian distribution with mean and covariance as $\hat{\mathbf{x}}_{t-1|t}$ and $\hat{\mathbf{P}}_{t-1|t}$ respectively,

$$\begin{aligned} \sigma_{k,t-1|t}^{-2,(i)} &= (\hat{f}_{k|t}^2 + \sigma_{f_k}^2) \mathbf{A}_{t,k}^H \tilde{\mathbf{R}}_t^{-1} \mathbf{A}_{t,k} + \sigma_{k,t-1|t}^{-2,(i-1)}, \\ \hat{\mathbf{P}}_{t-1|t} &= \text{diag}(\sigma_{1,t-1|t}^2, \dots, \sigma_{M,t-1|t}^2), \\ \langle x_{k,t-1|t}^{(i)} \rangle &= \sigma_{k,t-1|t}^{2,(i)} (\hat{f}_{k|t}^H \mathbf{A}_{t,k}^H \tilde{\mathbf{R}}_t^{-1} (\mathbf{y}_t - \mathbf{A}_{t,\bar{k}} \mathbf{F}_{\bar{k}} \mathbf{x}_{\bar{k},t-1}) \langle x_{\bar{k},t-1|t}^{(i-1)} \rangle) + \frac{\hat{x}_{k,t-1|t-1}}{\sigma_{k,t-1|t-1}^2}. \end{aligned} \quad (21)$$

Note that, in the algorithm implementation as shown in Algorithm 1 below, we introduce an iterative procedure (with i denoting the stage number) for the smoothing updates unlike [21] where there is no iteration for the covariance part. Note that we initialize the mean and variance in (33) from the converged values from the filtering stage.

4.4 Estimation of Hyper-Parameters

Update of $q_\gamma(\gamma)$: The Gamma distribution from the variational Bayesian approximation for the $q_\gamma(\gamma)$ can be written as,

$$\begin{aligned} \ln q_\gamma(\gamma) &= (c-1+N)\ln\gamma - \gamma(\langle \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|^2 \rangle + d) + c_\gamma, \\ q_\gamma(\gamma) &\propto \gamma^{c+N-1}e^{-\gamma(\langle \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|^2 \rangle + d)}. \end{aligned} \quad (22)$$

The mean of the Gamma distribution for γ is given by,

$$\begin{aligned} \langle \gamma \rangle &= \hat{\gamma}_t = \frac{c+\frac{N}{2}}{(\zeta_t+d)}, \quad \zeta_t = \beta\zeta_{t-1} + (1-\beta)\langle \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|^2 \rangle, \quad \text{where,} \\ \langle \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|^2 \rangle &= \|\mathbf{y}_t\|^2 - 2\Re(\mathbf{y}_t^H \mathbf{A}_t \hat{\mathbf{x}}_{t|t}) + \text{tr}\left(\mathbf{A}_t^H \mathbf{A}_t (\hat{\mathbf{x}}_{t|t} \hat{\mathbf{x}}_{t|t}^H + \hat{\mathbf{P}}_{t|t})\right), \end{aligned} \quad (23)$$

where we introduced temporal averaging also and β denotes the weighting coefficients which are less than one.

Update of $q_{f_k}(f_k)$: Using variational approximation we get a quadratic expression for $\ln q(f_k|\mathbf{y}_{1:t}) \sim \mathbb{E}_{q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{A}|\mathbf{y}_{1:t})} \ln p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{A}, \mathbf{y}_{1:t})$. Finally we write the mean and variance of the resulting Gaussian distribution as,

$$\sigma_{f_k|t}^2 = \frac{1}{\lambda_k \langle x_{k,t-1}^2 \rangle_t}, \quad \hat{f}_{k|t} = \frac{\langle x_{k,t|t} x_{k,t-1|t}^H \rangle_t}{\langle x_{k,t-1}^2 \rangle_t} \quad (24)$$

Here $\langle \cdot \rangle_t$ represents the temporal average given the observations till time t . We introduce temporal averaging here to approximate terms of the form $\langle x_{k,t|t} x_{k,t-1|t}^H \rangle$. This is done using the orthogonality property of LMMSE. So $\langle x_{k,t|t} x_{k,t-1|t}^H \rangle = \langle \hat{x}_{k,t|t} \hat{x}_{k,t-1|t}^H \rangle + \langle \tilde{x}_{k,t|t} \tilde{x}_{k,t-1|t}^H \rangle$. The Kalman filter (in linear state-space models and Gaussian noise) provides instantaneous $\hat{x}_{k,t|t}, \hat{x}_{k,t-1|t}$ and $\sigma_{k,t|t}^2, \sigma_{k,t-1|t}^2$. This explains why we do temporal averaging (sample average replacing statistical average). We define $\hat{\mathbf{P}}_{\mathbf{F}|t} = \text{diag}(\sigma_{f_1|t}^2, \dots, \sigma_{f_M|t}^2)$. Also we define the following covariance matrices, $\mathbf{R}_t^{m,n} = \langle \mathbf{x}_{t-n} \mathbf{x}_{t-m}^H \rangle_t$ and ξ_t represents the temporal weighting coefficient which is less than one [22],

$$\begin{aligned} \mathbf{R}_t^{0,0} &= (1-\xi_t)\mathbf{R}_{t-1}^{0,0} + \xi_t(\hat{\mathbf{x}}_{t|t} \hat{\mathbf{x}}_{t|t}^H + \hat{\mathbf{P}}_{t|t}), \quad \mathbf{R}_t^{1,0} = (\mathbf{R}_t^{0,1})^H = (1-\xi_t)\mathbf{R}_{t-1}^{1,0} + \\ &\xi_t \mathbf{F}(\hat{\mathbf{x}}_{t-1|t} \hat{\mathbf{x}}_{t-1|t}^H + \hat{\mathbf{P}}_{t-1|t}), \quad \mathbf{R}_t^{1,1} = (1-\xi_t)\mathbf{R}_{t-1}^{1,1} + \xi_t(\hat{\mathbf{x}}_{t-1|t} \hat{\mathbf{x}}_{t-1|t}^H + \hat{\mathbf{P}}_{t-1|t}). \end{aligned} \quad (25)$$

Further, we denote the $(i, j)^{th}$ element of $\mathbf{R}_t^{m,n}$ as $\mathbf{R}_t^{m,n}(i, j)$.

Update of $q_{\lambda_k}(\lambda_k)$: Using variational approximation $\ln q(\lambda_k|\mathbf{y}_{1:t}) \sim \mathbb{E}_{q(\mathbf{x}_t, \mathbf{x}_{t-1}, f_k|\mathbf{y}_{1:t})} \ln p(\mathbf{x}_t, \mathbf{A}, f_k | \mathbf{y}_{1:t})$, leading to

$$\begin{aligned} \ln \lambda_k - \lambda_k(\langle |x_{k,t} - f_k x_{k,t-1}|^2 \rangle + b) + (a-1)\ln \lambda_k + c_{\lambda_k}, \\ q_{\lambda_k}(\lambda_k) \propto \lambda_k^a e^{-\lambda_k(\langle |x_{k,t} - f_k x_{k,t-1}|^2 \rangle + b)}. \end{aligned} \quad (26)$$

The resulting gamma distribution is parameterized just by one quantity, the mean value, which gets used in the prediction stage and can be written as,

$$\langle \lambda_k \rangle = \frac{(a+1)}{(\langle |x_{k,t} - f_k x_{k,t-1}|^2 \rangle_t + b)}. \quad (27)$$

The temporal average $\langle |x_{k,t} - f_k x_{k,t-1}|^2 \rangle_t$ can be written as,

$$\begin{aligned} \langle |x_{k,t} - f_k x_{k,t-1}|^2 \rangle_t = \\ \mathbf{R}_t^{0,0}(k, k) - 2\Re\{\hat{f}_{k|t} \mathbf{R}_t^{1,0}(k, k)\} + (|\hat{f}_{k|t}|^2 + \sigma_{f_k|t}^2)\mathbf{R}_t^{1,1}(k, k). \end{aligned} \quad (28)$$

In Algorithm 1, we describe the GSAVE-KF algorithm in detail.

Algorithm 1 The GSAVE-KF Algorithm

Given: $\mathbf{A}_t, \mathbf{y}_t, N, M, \lambda_{k|0} = a/b \forall k, \gamma_0 = c/d, \sigma_{k,0|0}^2 = 0, \hat{\mathbf{x}}_{k,0|0} = 0 \forall k, t > 0$.

Prediction Stage

$$\sigma_{k,t|t-1}^2 = (|\hat{f}_{k|t-1}|^2 + \sigma_{f_{k|t-1}}^2) \sigma_{k,t-1|t-1}^2 + \frac{1}{\hat{\lambda}_{k|t-1}}, \quad \hat{\mathbf{x}}_{k,t|t-1} = \hat{f}_{k|t-1} \hat{\mathbf{x}}_{k,t-1|t-1},$$

Update Stage

Initialization: $\sigma_{k,t|t}^2 = \sigma_{k,t|t-1}^2, \hat{\mathbf{x}}_{t,\bar{k}|t}^{(0)} = \hat{\mathbf{x}}_{t,\bar{k}|t-1}$

for $i = 1, \dots$ until convergence

$$\sigma_{k,t|t}^{2,(i)} = \sigma_{k,t|t}^{2,(i-1)} (\sigma_{k,t|t}^{2,(i-1)} \hat{\gamma}_{t-1} \|\mathbf{A}_{t,k}\|^2 + 1)^{-1}, \quad \text{Kalman Gain: } \mathbf{K}_{k,t} = \sigma_{k,t|t}^{2,(i)} \mathbf{A}_{t,k}^H \hat{\gamma}_{t-1},$$

$$\hat{\mathbf{x}}_{k,t|t}^{(i)} = \frac{\sigma_{k,t|t}^{2,(i)}}{\sigma_{k,t|t-1}^2} \hat{\mathbf{x}}_{k,t|t-1} + \mathbf{K}_{k,t} \left(\mathbf{y}_t - \mathbf{A}_{t,\bar{k}} \hat{\mathbf{x}}_{t,\bar{k}|t}^{(i-1)} \right),$$

end for

Smoothing Stage

Initialization: $\hat{\mathbf{P}}_{t-1|t}^{(0)} = \hat{\mathbf{P}}_{t-1|t-1}, \hat{\mathbf{x}}_{t-1|t}^{(0)} = \hat{\mathbf{x}}_{t-1|t-1}$

for $i = 1, \dots$ until convergence

$$\hat{\mathbf{P}}_{t-1|t}^{-(i)} = (\hat{\mathbf{F}}_{|t}^H \mathbf{A}_t^H \hat{\mathbf{R}}_t^{-1} \mathbf{A}_t \hat{\mathbf{F}}_{|t} + \text{diag}(\mathbf{A}_t^H \hat{\mathbf{R}}_t^{-1} \mathbf{A}_t) \hat{\mathbf{P}}_{\mathbf{F}|t} + \hat{\mathbf{P}}_{t-1|t}^{-(i-1)}),$$

$$\hat{\mathbf{x}}_{t-1|t}^{(i)} = \hat{\mathbf{P}}_{t-1|t}^{(i)} (\hat{\mathbf{P}}_{t-1|t-1}^{-1} \hat{\mathbf{x}}_{t-1|t-1}^{(i-1)} + \hat{\mathbf{F}}^H \mathbf{A}_t^H \hat{\mathbf{R}}_t^{-1} \mathbf{y}_t).$$

end for

Estimation of Hyper-Parameters

Compute $\zeta_t, \mathbf{R}_t^{m,n}$ from (23), (25).

$$\sigma_{f_k|t}^2 = \frac{1}{\lambda_k \mathbf{R}_t^{1,1}(k,k)}, \quad \hat{f}_{k|t} = \frac{\mathbf{R}_t^{1,0}(k,k)}{\mathbf{R}_t^{1,1}(k,k)}.$$

$$\hat{\gamma}_t = \frac{c + \frac{N}{2}}{(\zeta_t + d)}, \quad \hat{\lambda}_{k|t} = \frac{a+1}{(\mathbf{R}_t^{0,0}(k,k) - 2\Re\{\hat{f}_{k|t}^H \mathbf{R}_t^{1,0}(k,k)\} + (|\hat{f}_{k|t}|^2 + \sigma_{f_k|t}^2) \mathbf{R}_t^{1,1}(k,k) + b)}.$$

5 VB-KF for Diagonal AR(1) (DAR(1))

In this section, we treat the components of the state \mathbf{x}_t jointly, with all the hyper-parameters λ_k, f_k, γ assumed to be independent in the q 's. So the expressions for the estimates of the hyper-parameters can be shown to be the same as in the previous section on SAVE-KF.

5.1 DAR(1) Prediction Stage

The prediction about \mathbf{x}_t can be computed from the time update equation of the standard Kalman filter, $\mathbf{x}_t = \hat{\mathbf{F}}_{|t-1} \mathbf{x}_{t-1|t-1} + \tilde{\mathbf{F}}_{|t-1} \mathbf{x}_{t-1|t-1} + \mathbf{v}_t$, $\mathbf{F} = \hat{\mathbf{F}}_{|t-1} + \tilde{\mathbf{F}}_{|t-1}$, where $\hat{\mathbf{F}}_{|t-1} = \text{diag}(\hat{f}_{1|t-1}, \dots, \hat{f}_{M|t-1})$. We also define $\hat{\Lambda}_{|t-1} = \text{diag}(\frac{1}{\hat{\lambda}_{1|t-1}}, \dots, \frac{1}{\hat{\lambda}_{M|t-1}})$. Substituting $\mathbf{x}_{t-1|t-1} = \hat{\mathbf{x}}_{t-1|t-1} + \tilde{\mathbf{x}}_{t-1|t-1}$,

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &= \hat{\mathbf{F}}_{|t-1} \hat{\mathbf{x}}_{t-1|t-1}, \quad \tilde{\mathbf{x}}_{t|t-1} = \hat{\mathbf{F}}_{|t-1} \tilde{\mathbf{x}}_{t-1|t-1} + \tilde{\mathbf{F}}_{|t-1} \mathbf{x}_{t-1|t-1} + \mathbf{w}_t, \implies \\ \hat{\mathbf{P}}_{t|t-1} &= \hat{\mathbf{F}}_{|t-1} \hat{\mathbf{P}}_{t-1|t-1} \hat{\mathbf{F}}_{|t-1}^H + \hat{\mathbf{P}}_{\mathbf{F}|t-1} \text{diag}(\hat{\mathbf{x}}_{t-1|t-1} \hat{\mathbf{x}}_{t-1|t-1}^H + \hat{\mathbf{P}}_{t-1|t-1}) + \hat{\Lambda}_{|t-1}. \end{aligned} \quad (29)$$

5.2 Measurement or Update Stage

Using (11),

$$\begin{aligned} \ln q_{\mathbf{x}_t}(\mathbf{x}_t) = -\langle \gamma \rangle & \left\{ -\mathbf{y}_t^H \mathbf{A}_t \mathbf{x}_t - \mathbf{x}_t^H \mathbf{A}_t^H \mathbf{y}_t + \mathbf{x}_t^H \mathbf{A}_t^H \mathbf{A}_t \mathbf{x}_t \right\} - \mathbf{x}_t^H \widehat{\mathbf{P}}_{t|t-1}^{-1} \mathbf{x}_t \\ & + \mathbf{x}_t^H \widehat{\mathbf{P}}_{t|t-1}^{-1} \widehat{\mathbf{x}}_{t|t-1} + \widehat{\mathbf{x}}_{t|t-1}^H \widehat{\mathbf{P}}_{t|t-1}^{-1} \mathbf{x}_t + c_{x_t} = -(\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t})^H \widehat{\mathbf{P}}_{t|t}^{-1} (\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t}) + c'_{x_t}, \end{aligned} \quad (30)$$

where the mean and variance are written as,

$$\widehat{\mathbf{P}}_{t|t}^{-1} = \langle \gamma \rangle \mathbf{A}_t^H \mathbf{A}_t + \widehat{\mathbf{P}}_{t|t-1}^{-1}, \quad \widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{P}}_{t|t} (\langle \gamma \rangle \mathbf{A}_t^H \mathbf{y}_t + \widehat{\mathbf{P}}_{t|t-1}^{-1} \widehat{\mathbf{x}}_{t|t-1}). \quad (31)$$

5.3 Fixed Lag Smoothing

The posterior distribution $p(\mathbf{x}_{t-1}/\mathbf{y}_{1:t-1})$ is approximated using variational approximation as $q(\mathbf{x}_{t-1}/\mathbf{y}_{1:t-1})$ with mean and covariance as $\widehat{\mathbf{x}}_{t-1|t-1}$ and $\widehat{\mathbf{P}}_{t-1|t-1}$.

$$\begin{aligned} \ln p(\mathbf{y}_t, \mathbf{x}_{t-1}, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = \frac{-1}{2} \ln \det \widetilde{\mathbf{R}}_t - (\mathbf{y}_t - \mathbf{A}_t \mathbf{F} \mathbf{x}_{t-1}) \widetilde{\mathbf{R}}_t^{-1} (\mathbf{y}_t - \mathbf{A}_t \mathbf{F} \mathbf{x}_{t-1}) \\ - \frac{1}{2} \det(\widehat{\mathbf{P}}_{t-1|t-1}) - (\mathbf{x}_{t-1} - \widehat{\mathbf{x}}_{t-1|t-1})^H \widehat{\mathbf{P}}_{t-1|t-1}^{-1} (\mathbf{x}_{t-1} - \widehat{\mathbf{x}}_{t-1|t-1}) + c_{x-1}, \end{aligned} \quad (32)$$

Prediction of \mathbf{x}_{t-1} : Using (11), $\ln q_{\mathbf{x}_{t-1}}(\mathbf{x}_{t-1}/\mathbf{y}_{1:t})$ turns out to be quadratic in \mathbf{x}_{t-1} and thus can be represented as a Gaussian distribution with mean and covariance as $\widehat{\mathbf{x}}_{t-1|t}$ and $\widehat{\mathbf{P}}_{t-1|t}$ respectively,

$$\begin{aligned} \widehat{\mathbf{P}}_{t-1|t}^{-(i)} &= (\widehat{\mathbf{F}}_{t|t}^H \mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1} \mathbf{A}_t \widehat{\mathbf{F}}_{t|t} + \text{diag}(\mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1} \mathbf{A}_t) \widehat{\mathbf{P}}_{\mathbf{F}|t} + \widehat{\mathbf{P}}_{t-1|t}^{-(i-1)})^{-1}, \\ \widehat{\mathbf{x}}_{t-1|t}^{(i)} &= \widehat{\mathbf{P}}_{t-1|t}^{(i)} (\widehat{\mathbf{P}}_{t-1|t-1}^{-1} \widehat{\mathbf{x}}_{t-1|t-1}^{(i-1)} + \widehat{\mathbf{F}}_{t|t}^H \mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1} \mathbf{y}_t). \end{aligned} \quad (33)$$

6 Simulation Results

For the observation model, \mathbf{y}_t is of dimension 100×1 and \mathbf{x}_t is of size 200×1 with 30 non-zero elements. All signals are considered to be real in the simulation. All the elements of \mathbf{A}_t (time varying) are generated i.i.d. from a Gaussian distribution with mean 0 and variance 1. The rows of \mathbf{A}_t are scaled by $\sqrt{30}$ so that the signal part of any scalar observation has unit variance. Taking the SNR to be 20dB, the variance of each element of \mathbf{v}_t (Gaussian with mean 0) is computed as 0.01.

Consider the state update, $\mathbf{x}_t = \mathbf{F} \mathbf{x}_{t-1} + \mathbf{w}_t$. To generate \mathbf{x}_0 , the first 30 elements are chosen as Gaussian (mean 0 and variance 1) and then the remaining elements of the vector \mathbf{x}_0 are put to zero. Then the elements of \mathbf{x}_0 are randomly permuted to distribute the 30 non-zero elements across the whole vector. The diagonal elements of \mathbf{F} are chosen uniformly in $[0.9, 1)$. Then the covariance of \mathbf{w}_t can be computed as $\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{F}\mathbf{F}^H)$. Note that $\boldsymbol{\Lambda}$ contains the variances of the elements of \mathbf{x}_t (including $t = 0$), where for the non-zero elements of \mathbf{x}_0 the variance is 1 and for the zero elements it is 0. In Fig. 1, the blue curve corresponds to the case of a standard Kalman Filter with known state-space model parameters. The red curve corresponds to GSAVE-KF with again all these hyper-parameters known. The green curve corresponds to the case of GSAVE-KF with all the hyper parameters also estimated with lag-1 smoothing. Further, we show that filtering for AR(1) coefficients (black curve) doesn't converge to the basic KF. NMSE is the normalized mean squared error at time t computed as $\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2$,

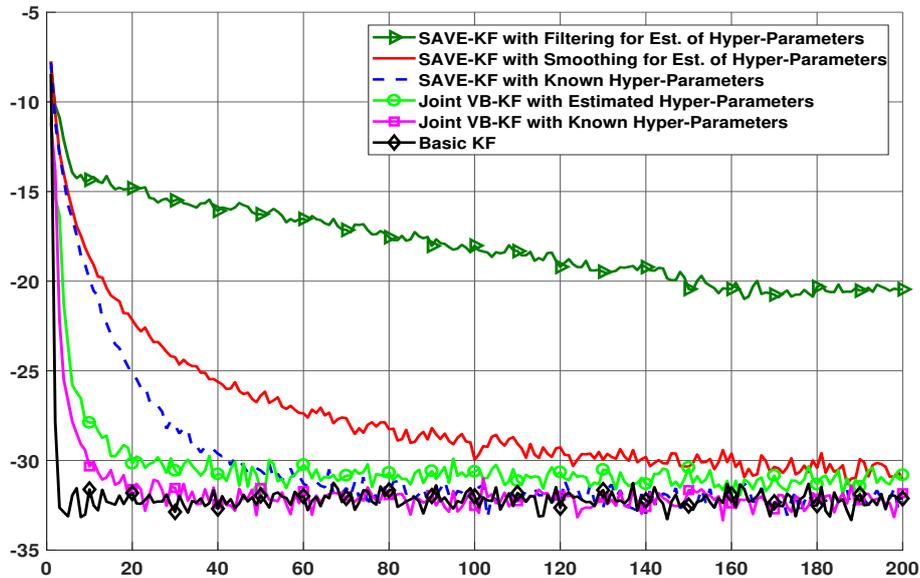


Fig. 1. NMSE as a function of time (i.e. number of measurements or iteration index).

averaged over 100 different realizations of \mathbf{A}_t , \mathbf{F} , and of course the noise realizations. The simulations show that in the scenario considered, GSAVE-KF exhibits hardly any MSE degradation over the more complex standard Kalman Filter in steady-state, but takes time to reach steady-state. Adding the estimation of the parameters leads to further slight degradations in steady-state and transient.

7 Conclusions

We presented a fast SBL algorithm called GSAVE-KF, which uses the variational inference techniques to approximate the posteriors of the data and parameters and track a time varying sparse signal. GSAVE-KF helps to circumvent the matrix inversion operation required in conventional SBL using the EM algorithm. We showed that in spite of the significantly reduced computational complexity, the proposed algorithm with estimation of the unknown model parameters has similar steady-state performance compared to the standard Kalman filter, at the price of a significantly increased transient.

References

1. S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
2. J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

3. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
4. D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. on Sig. Process.*, vol. 52, no. 8, pp. 2153–2164, August 2004.
5. Z. Zhang and B. D. Rao, "Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning," *IEEE J. of Sel. Topics in Sig. Process.*, vol. 5, no. 5, pp. 912 – 926, September 2011.
6. D. Angelosan, J. A. Bazerq, and G. B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the l1-norm," *IEEE Trans. on Sig. Process.*, vol. 58, no. 7, Jul. 2010.
7. C. Couvreur and Y. Bresler, "Decomposition of a mixture of Gaussian AR processes," *IEEE Intl. Conf. on Acous., Speech, and Sig. Process.*, vol. 3, pp. 1605–1608, 1995.
8. M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *IEEE Transactions on Acous., Speech and Sig. Process.*, vol. 36, no. 4, pp. 477–489, Apr 1988.
9. R. D. Bass, V. D. Norum, and L. Swartz, "Optimal multichannel nonlinear filtering," *J. Mufh. Anal. Appl.*, vol. 16, pp. 152 – 164, 1966.
10. A. H. Jazwinski, *Stochastic processes and filtering theory*, 1970.
11. R. Henriksen, "The truncated second-order nonlinear filter revisited," *IEEE Transactions on Automatic Control*, vol. 27, no. 1, pp. 247 – 251, feb 1982.
12. M. Athans, R. Wishner, and A. Bertolini, "Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements," *IEEE Transactions on Automatic Control*, vol. 13, no. 5, pp. 504 – 514, oct 1968.
13. J. Villares and G. Vazquez, "The quadratic extended Kalman filter," in *Sens. Arr. and Multichnl. Sig. Process. Wkshp. (SAM), 2004*, july 2004, pp. 480 – 484.
14. C. K. Thomas and D. Slock, "SAVE - space alternating variational estimation for sparse Bayesian learning," in *Data Science Workshop*, 2018.
15. T. Sadiki and D. T. Slock, "Bayesian adaptive filtering: principles and practical approaches," in *EUSIPCO*, 2004.
16. J. B. S. Ciochina, C. Paleologu, "A family of optimized LMS-based algorithms for system identification," in *Proc. EUSIPCO*, 2016, pp. 1803–1807.
17. C. K. Thomas and D. Slock, "Variational Bayesian Learning for Channel Estimation and Transceiver Determination," in *Info. Theo. and Appl. Wkshp*, San Diego, USA, February 2018.
18. B. H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, "Channel Parameter Estimation in Mobile Radio Environments Using the SAGE Algorithm," *IEEE J. on Sel. Areas in Commun.*, vol. 17, no. 3, pp. 434–450, March 1999.
19. D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Sig. Process. Mag.*, vol. 29, no. 6, pp. 131–146, November 2008.
20. R. Herbrich, "Minimising the Kullback-Leibler Divergence," in *Microsoft Research*, August 2015.
21. B. Ait-El-Fquih and I. Hoteit, "Fast Kalman-like filtering for large-dimensional linear and Gaussian state-space models," *IEEE Trans. on Sig. Process.*, vol. 63, no. 21, Nov. 2015.
22. S. Bensaid and D. Slock, "Comparison of Various Approaches for Joint Wiener/Kalman Filtering and Parameter Estimation with Application to BASS," in *IEEE 45th Asilomar Conference on Sig., Sys. and Comp.*, Pacific Grove, CA, USA, 2011.

A geometric proxy of economic uncertainty based on the disagreement in survey expectations

Oscar Claveria¹, Enric Monte² and Salvador Torra³

¹ AQR-IREA, University of Barcelona, Barcelona 08034, Spain

² Department of Signal Theory and Communications, Polytechnic University of Catalunya, 08034 Barcelona, Spain

³ Riskcenter-IREA, Department of Econometrics and Statistics, University of Barcelona, 08034 Barcelona, Spain
oclaveria@ub.edu

Abstract. In this study we present a geometric approach to proxy economic uncertainty. We design a positional indicator of disagreement among survey-based agents' expectations about the state of the economy. Previous dispersion-based uncertainty measures derived from business and consumer surveys exclusively make use of the two extreme pieces of information: the percentage of respondents expecting a variable to rise and to fall. With the aim of also incorporating the information coming from the share of respondents expecting a variable to remain constant, we propose a geometrical framework and use a barycentric coordinate system to generate a metric of disagreement, referred to as a discrepancy indicator. We assess its performance, both empirically and experimentally, by comparing it to the standard deviation of the share of positive and negative responses, which has been used by Bachman et al. (2013) as a proxy for economic uncertainty. When applied in sixteen European countries, we find that both time-varying metrics co-evolve in most countries for expectations about the country's overall economic situation in the present, but not in the future. Additionally, we obtain their simulated sampling distributions and we find that the proposed indicator gravitates uniformly towards the three vertices of the simplex representing the three answering categories, as opposed to the standard deviation, which tends to overestimate the level of uncertainty as a result of ignoring the no-change responses. Consequently, we find evidence that the information coming from agents expecting a variable to remain constant has an effect on the measurement of disagreement.

Keywords: Economic uncertainty, Expectations, Disagreement, Geometry.

Acknowledgements. This research was supported by the projects ECO2016-75805-R and TEC2015-69266-P from the Spanish Ministry of Economy and Competitiveness. We also wish to thank Anna Stangl, Johanna Garnitz and Klaus Wohlrabe at the Ifo Institute for Economic Research in Munich for providing us the data used in the study.

1 Introduction

The arrival of the 2008 financial crisis has triggered a body of research dedicated to analyse the impact of uncertainty on the economy (Ajmi et al., 2015; Arslan et al., 2015; Atalla et al., 2016; Azqueta-Gavaldón, 2017; Balcilar et al., 2017; Binder, 2017; Binding and Dibiasi, 2017; Bloom, 2014; Caggiano et al., 2014; Chua et al., 2011; Dovern, 2015; Fernández-Villaverde et al., 2015; Hartmann et al., 2017; Henzel and Rengel, 2017; Perić and Sorić, 2017; Sorić and Lolić, 2017). Since economic uncertainty is not directly observable, several strategies have been proposed to measure it.

A first approach consists on tracking the magnitude of forecast errors of macroeconomic variables (Glass and Fritsche, 2014; Jurado et al., 2015). This approach is based on the assumption that in times of high uncertainty forecast errors are expected to rise, but its ex-post nature has led researchers to develop alternative approaches to measure economic uncertainty.

A second approach is based on the assumption that notions about the future evolution of the economy are likely to be more dispersed in times of high uncertainty. This premise allows to develop dispersion-based indicators. These measures can either be based on stock market volatility (Basu and Bundick, 2012; Bekaert et al., 2013; Bloom, 2009), or on agents' economic expectations (Glass and Hartmann, 2016; Lahiri and Sheng, 2010; Mankiw et al., 2004; Mokinski et al., 2015).

Direct measures of expectations can only be derived from surveys (Claveria et al., 2017). Tendency surveys ask respondents whether they expect a variable to rise, fall or remain unchanged. By using agents' expectations coming from economic tendency surveys, Bachman et al. (2013) proposed a set of uncertainty indicators based on the dispersion of respondents' expectations about the future in Germany and the United States (US). Girardi and Reuter (2017) have recently presented three new dispersion-based uncertainty indicators derived from business and consumer surveys for the Euro Area (EA).

All these dispersion-based indicators of disagreement among respondents elicit the information exclusively from the respondents expecting a variable to rise and to fall, leaving out the responses from agents expecting no-change. This omission has led us to devise an approach that allows to derive a time-varying disagreement metric that incorporates the information coming from all three answering categories.

With this aim, we present a geometric setup to construct a positional indicator of disagreement that can be interpreted as the percentage of discrepancy among responses. We focus on agents' expectations about the country's situation regarding the overall economy both at present and by the end of the next six months. We compare the performance of the proposed measure of displacement to the standard deviation of the share of positive and negative responses, which has been used by Bachman et al. (2013) as a proxy for economic uncertainty. The analysis is carried out in sixteen European countries, focusing on the period prior to the start of the 2008 financial crisis, which provides a natural backdrop for the experiment.

The structure of the paper is as follows. The next section presents the methodological approach. Empirical results are provided in Section 3. Finally, concluding remarks and future lines of research are drawn in Section 4.

2 Methodology

In this section we present a geometric approach to derive a dispersion-based measure of positional disagreement. The proposed framework allows to capture the proportion of discrepancy among survey respondents in any given period by means of spatial vectors. Tendency surveys are addressed to economic agents in order to elicit subjective measures of their expectations about the state of the economy. Respondents are asked about the expected direction of change of a wide range of variables (inflation, consumption, etc.). In this study we focus on the expectations about the country's situation in terms of its overall economy, both at present and by end of the next six months.

Survey results are available about one quarter ahead of the publication of quantitative official data and are usually presented as balances, B_t , which consist on the subtraction between the weighted percentage of respondents expecting a variable to go up (R_t) and to go down (F_t). Nevertheless, survey results can be aggregated in a three dimensional vector denoted as V_t :

$$V_t = (R_t, E_t, F_t) \quad (1)$$

where E_t refers to the proportion of respondents expecting the variable to remain constant. The variance of the balance could be defined as:

$$D_t = R_t + F_t - B_t^2 \quad (1)$$

Theil (1955) defined expression (2) as the disconformity coefficient, due to the fact that the value of D_t would reach the minimum value zero when all the responses are concentrated in either one of the two categories. The maximum disconformity, corresponding to a value of one, would take place, if and only if, R_t and F_t each accumulates half of the responses. Expression (2) implicitly neglects the variate E_t . As a result, the 'no-change' proportion is not directly incorporated into the disagreement metric. Claveria (2010) proposed a nonlinear variation of the balance statistic that accounted for this percentage of respondents.

Bachmann et al. (2013) used an economic uncertainty proxy denoted as $DISP_t$ that can be defined as the square root of D at time t :

$$DISP_t = \sqrt{R_t + F_t - B_t^2} \quad (3)$$

The authors applied this measure to the forward-looking survey question related to the expectations of domestic production activities in Germany at the micro level. Girardi et al. (2017) developed an aggregate variation of expression (3) in order to compute the cross-sectional standard deviation of the share of positive and negative responses for all forward-looking survey questions, and then standardised the question-specific measures and rescaled the average dispersion.

With the aim of incorporating the information coming from the respondents expecting no-change in the variable, we develop a methodological framework that allows to

construct a measure of disagreement that conveys a geometrical interpretation. The proposed metric presents two inherent advantages. On the one hand, it allows to capture the trajectories of the three states. On the other hand, it has a self-explanatory interpretation, as it provides the proportion of disagreement among respondents.

In order to explicitly incorporate the three components of the surveys (R_t, E_t, F_t) , we assume that no-change responses can proxy either one of the extreme options. Note that the fraction of answers falling into the ‘no-change’ category is conveying the information about the confidence on the other two categories. Kahneman (2011) noted that when faced with a difficult question, respondents often choose an easier one instead.

As the sum of the proportions adds to a constant, a natural representation of the answers will be as a point on a simplex (Coxeter, 1969). A simplex could be defined as the smallest convex set containing the given vertices. We will use a two-dimensional simplex, which corresponds to a triangle. The interior of this simplex encompasses all possible combinations of proportions between the three answering categories.

The equilateral triangle S can be defined by its three vertices $\{x, y, z\}$ (see left panel of Fig. 2). A simplex in \mathbb{R}^3 can be defined as $a_1x + a_2y + a_3z$, such that $a_1 + a_2 + a_3 = 1$ and $a_1, a_2, a_3 \geq 0$, where a_1, a_2 and a_3 stand for the three proportions defined in (1). These proportions can be regarded as the barycentric coordinates of a point with respect to S . Therefore, each point inside S has a unique convex combination of the three vertices determined by the set of aggregated survey results.

The barycentric coordinate system allows us to compute the vertical distance of a point in the simplex to the nearest edge, as it can be seen in the right panel of Fig. 1. As there are two degrees of freedom, any set of barycentric coordinates and their corresponding basis vectors can be used to define the location of any point within S .

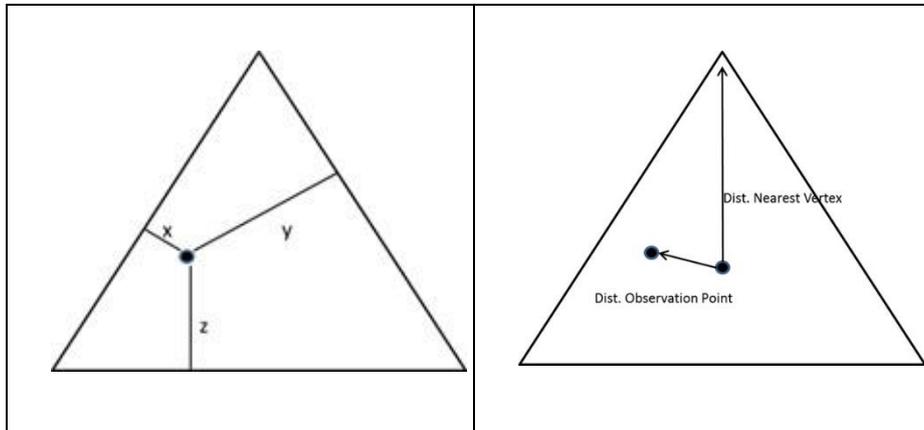


Fig. 1. Simplex S – Barycentric coordinates.

Once we have defined the location of the point within the simplex, we formalise a measure of consensus along the lines of the one proposed by Claveria (2018) for five reply options. We define a measure that summarises the notion that the centre of the

simplex corresponds to the point of maximum discrepancy among respondents. Conversely, the fact that the coordinates on the simplex are near a vertex is indicative that there is a high level of consensus (or concentration):

$$\text{Concentration} = \frac{\text{Distance of the observation point to the barycentre}}{\text{Distance of a vertex to the barycentre}} \quad (4)$$

Given that all vertices are at the same distance to the barycentre, this ratio gives the relative weight of the distance of each point in time to the centre of the triangle. We can then formalise concentration for period t as C_t as:

$$C_t = \frac{\sqrt{(R_t-1/3)^2 + (E_t-1/3)^2 + (F_t-1/3)^2}}{\sqrt{2/3}} \quad (5)$$

Consequently, the proposed geometry-based disagreement measure, which will be referred to as a discrepancy indicator, can be defined as the inverse of consensus:

$$G_t = 1 - C_t \quad (6)$$

3 Empirical results

Uncertainty is unobservable. Economic uncertainty can be defined as the situation in which economic agents are not able to anticipate future events or estimate the likelihood of their occurrence (Knight, 1921). Since the advent of the 2008 financial crisis, there has been a renewed interest in the measurement of economic uncertainty. Baker et al. (2016) designed the economic policy uncertainty (EPU) index, which uses the responses from the Surveys of Professional Forecasters.

While the development of machine learning techniques increasingly facilitates the generation of ad-hoc media indexes of frequencies of keyword combinations related to uncertainty that avoid the pre-labelling of the data (Azqueta-Gavaldón, 2017), this approach still entails a non-negligible degree of subjectivity (Girardi and Reuter, 2017). As a result, based on the assumption that the dispersion of expectations increases during periods of high uncertainty, one of the most common approaches to proxy economic uncertainty is to use measures of disagreement among survey expectations (Giordani and Söderlind 2003; Glass and Hartmann, 2016; Lahiri and Sheng, 2010; Mokinski et al., 2015; Rich and Tracy 2010; Zarnowitz and Lambros 1987).

Economic expectations are not directly observable, and therefore are elicited through survey data. Recent research has shown that the data provided by business and consumer tendency surveys is particularly useful in order to derive uncertainty measures based on the dispersion of expectations (Bachmann et al., 2013). Disagreement measures are based on the responses that fall into the two extreme answering categories, that is, the respondents expecting a variable to increase and the ones expecting it to decrease. In this study, we want to evaluate the effect of incorporating the information coming from the respondents expecting a variable to remain constant.

With this aim we use raw data from the World Economic Survey (WES) carried out quarterly by the Ifo Institute for Economic Research. The WES assesses worldwide economic trends by polling professionals and experts on current economic developments in their respective countries. We focus on the question about the country's situation in terms of its overall economy, both present and future. We use the shares of respondents expecting a variable to go up, to go down or to remain unchanged during the period ranging from 2005:Q2 to 2008:Q4 in sixteen European countries (Austria, Belgium, Finland, France, Germany, Greece, Hungary, Italy, Latvia, the Netherlands, Poland, Portugal, Romania, Spain, Sweden and the United Kingdom).

First, we project survey answers in the simplex for each period of the sample (2005:Q2-2008:Q4). Second, by means of the barycentric coordinates of each point we compute G_t . To assess the performance of this metric of positional discrepancy we compare it to the uncertainty proxy proposed by Bachmann et al. (2013), $DISP_t$, defined in (3). Both indicators are bounded between zero and one. A one value indicates maximum disagreement, while zero maximum consensus. In Table 1 we present the obtained correlations between both dispersion-based disagreement measures. We can see that G_t and $DISP_t$ co-evolve for the present, but not so much for the future.

Table 1. Correlations between both measures of disagreement.

Country	Present	Future	Country	Present	Future
Austria	0.338	0.438	Latvia	0.860**	0.528*
Belgium	0.571*	0.404	Netherlands	0.538*	0.294
Finland	0.948**	0.261	Poland	0.645**	0.175
France	0.375	0.095	Portugal	0.966**	0.337
Germany	0.044	-0.004	Romania	0.112	0.225
Greece	0.849**	0.572*	Spain	0.840**	0.462
Hungary	0.509	0.388	Sweden	0.711**	-0.262
Italy	0.569*	-0.264	UK	0.844**	0.285

Notes: Both indicators are bounded between zero and one. A one value indicates maximum disagreement; while zero, maximum consensus. * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

With the aim of further assessing the performance of both indicators, we sample the simplex defined in section 2. We generate a uniform set of points in the unit cube, and then normalise each point such that the sum of the coordinates is equal to one. This procedure is equivalent to projecting the distribution onto a plane in order to sample the simplex of both metrics of disagreement among respondents.

In Fig. 2 we depict the overlapped non-normalised histograms of both statistics. While both distributions are similar and negatively skewed, the positional discrepancy indicator proposed in this study shows a fatter tail, suggesting a higher level of granularity. In Table 2 we present the summary statistics of both simulated distributions.

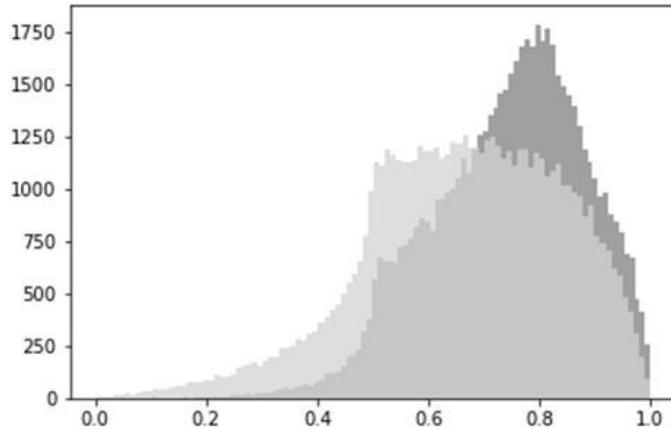


Fig. 2. Histogram of simulated distributions – $DISP_t$ vs. G_t . The lighter histogram represents the distribution of the proposed positional indicator of discrepancy; while the darker histogram at the back represents the distribution of $DISP_t$.

Table 2. Summary statistics of simulated distribution of disagreement measures.

Metric of disagreement	Mean	Std. Dev.	Min.	Max.	Range	IQR
$DISP_t$	0.742	0.137	0.054	1.000	0.946	0.195
G_t	0.662	0.176	0.004	0.998	0.994	0.255

Note: IQR stands for the interquartile range, which is a measure of dispersion obtained as the difference between upper and lower quartiles, $Q3-Q1$.

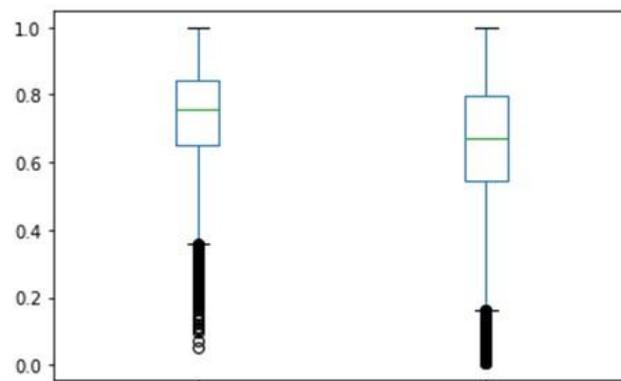


Fig. 3. Boxplot of simulated distributions – $DISP_t$ vs. G_t . The boxplot to the left represents the distribution of the disagreement measure proposed by Bachmann et al. (2013), while the one to the right that of the proposed positional indicator of discrepancy. A one value indicates maximum disagreement; while zero, maximum consensus.

The IQR in Table 2 differs between both distributions, being significantly larger for G_t . This result is indicative of a higher level of granularity for the median values of the distribution of the discrepancy indicator in comparison to Bachmann et al.'s (2013) disagreement indicator.

In Fig. 3 we graph the boxplots, which represent the distribution of each indicator through the quartiles without making any assumptions of the underlying statistical distribution. It can be seen that the distribution of the discrepancy indicator encompasses a much wider range of the scale, and its distribution of scores is more uniform.

Finally, in Fig. 4 we project the barycentric coordinates of the simulated points in the simplex for both indicators. The higher granularity of the indicator proposed in this article is manifested by the fact that the areas for each level of scores is more uniform. We can see that G_t behaves uniformly in all three directions, while the disconformity indicator shows a wider area in which gives a maximum value of disagreement. This result is caused by not taking into account the share of no-change responses.

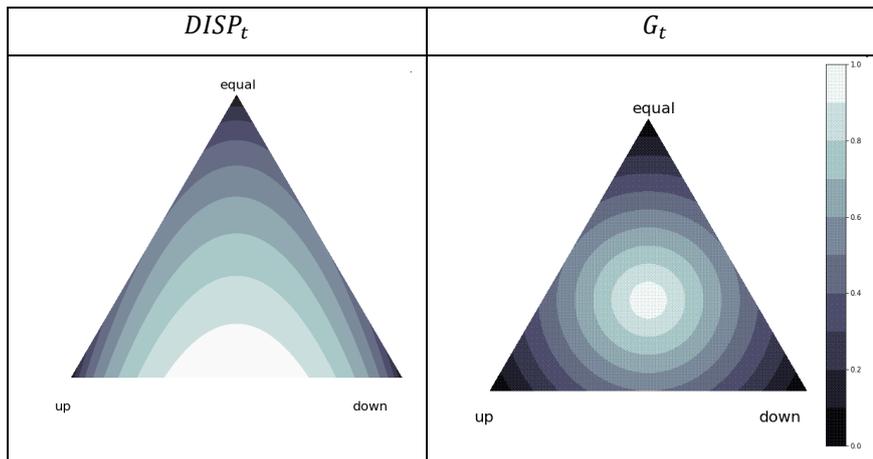


Fig. 4. Projection of barycentric coordinates of simulated points onto the simplex. The simplex to the left represents the distribution of the disagreement measure proposed by Bachmann et al. (2013), while the one to the right that of the proposed positional indicator of discrepancy. A one value indicates maximum disagreement; while zero, maximum consensus.

4 Concluding remarks

This paper presents a geometrical framework to proxy economic uncertainty by means of a survey-based measure of disagreement among respondents. The fact that tendency surveys ask agents whether they expect a particular variable to increase, decrease or remain unchanged, has led us to design an indicator that takes into account all three magnitudes. Previous dispersion-based uncertainty indicators derived from business and consumer surveys exclusively make use of the two extreme pieces of information, that is, the responses expecting a variable to rise and to fall.

Our main aim was to incorporate the share of respondents expecting a variable to remain constant. With this objective, we project survey responses onto a simplex that takes the form of an equilateral triangle, and by means of spatial vectors we derive a measure of displacement that incorporates all three pieces of information.

To assess the performance of the proposed measure of positional discrepancy we compare it, both empirically and experimentally, to the standard deviation of the share of positive and negative responses, which has been used by Bachman et al. (2013) as a measure of disagreement. First, we compute both measures for sixteen European countries, finding that they co-evolve during the sample period in most countries, especially for the expectations about the country's current economic situation.

Second, we generate the simulated sampling distributions of both the proposed geometric indicator of discrepancy and the disagreement measure used as a benchmark. In spite of the fact that both distributions are negatively skewed and similar, we find that the distribution of the proposed positional indicator of discrepancy shows a fatter tail, suggesting a higher level of granularity for the intermediate values, which is confirmed by a higher value of the interquartile range.

By projecting the barycentric coordinates of the simulated points onto the simplex, we observe that the proposed discrepancy indicator gravitates uniformly towards the three vertices of the triangle, defined by the three answering categories. Conversely, the disagreement measure used as a benchmark tends to overestimate the level of uncertainty as a result of ignoring the no-change share of responses. Arguably, it seems that the information coming from agents expecting a variable to remain constant has an effect on the measurement of disagreement among survey respondents.

In spite of the novelty of the approach, the metric presented in the paper is not without limitations. The proposed geometrically-based discrepancy indicator is a measure of disagreement among survey respondents, and as such has to be considered a proxy of uncertainty, which is a latent variable. As noted by Girardi et al. (2017), the evolution of survey-based disagreement indicators does not only reflect changes in underlying uncertainty levels, but also in heterogeneity among agents' expectations. An issue left for further research is extending the construction of the indicator on the basis of responses to additional variables. Another line of future research is the analysis of the impact of the proposed uncertainty metric on economic activity.

References

1. Ajmi, A. N., Aye, G. C., Balcilar, M., El Montasser, G.: Casualty between US economic policy and equity market uncertainties: Evidence from linear and nonlinear tests. *Journal of Applied Economics*, 18(2), 225–246 (2015).
2. Arslan, Y., Atabek, A., Hulagu, T., Şahinöz, S.: Expectation errors, uncertainty, and economic activity. *Oxford Economic Papers*, 67(3), 634–660 (2015).
3. Atalla, T., Joutz, F., Pierru, A.: Does disagreement among oil price forecasters reflect volatility? Evidence from the ECB surveys. *International Journal of Forecasting*, 32(4), 1178–1192 (2016).
4. Azqueta-Gavaldón, A.: Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters*, 158, 47–50 (2017).

5. Bachmann, R., Elstner, S., Sims, E. R.: Uncertainty and economic activity: Evidence from business survey data. *American Economic Journal: Macroeconomics*, 5(2), 217–249 (2013).
6. Baker, S. R., Bloom, N., Davis, S. J.: Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636 (2016).
7. Balcilar, M., Bekiros, S., Gupta, R.: The role of news-based uncertainty indices in predicting oil markets: a hybrid nonparametric quantile causality method. *Empirical Economics*, 53(3), 879–889 (2017).
8. Basu, S, Bundick, B.: Uncertainty shocks in a model of effective demand. *Econometrica*, 85(3), 937–958 (2017).
9. Bekaert, G., Hoerova, M., Lo Duca, M.: Risk, uncertainty and monetary policy. *Journal of Monetary Economics*, 60(7), 771–788 (2013).
10. Binder, C.: Measuring uncertainty based on rounding: New method and application to inflation expectations. *Journal of Monetary Economics*, 90, 1–12 (2017).
11. Binding, G., Dibiasi, A.: Exchange rate uncertainty and firm investment plans evidence from Swiss survey data. *Journal of Macroeconomics*, 51, 1–27 (2017).
12. Bloom, N.: The impact of uncertainty shocks. *Econometrica*, 77(3), 623–685 (2009).
13. Bloom, N.: Fluctuations in uncertainty. *Journal of Economic Perspectives*, 28(2), 153–176 (2014).
14. Caggiano, G., Castelnuovo, E., Groshenny, N.: Uncertainty shocks and unemployment dynamics in U.S. recessions. *Journal of Monetary Economics*, 67(C), 78–92 (2014).
15. Claveria, O. A new consensus-based unemployment indicator. *Applied Economics Letters*. In Press (2018).
16. Claveria, O. Qualitative survey data on expectations. Is there an alternative to the balance statistic? In Molnar. A. T. (ed.), *Economic Forecasting* (pp. 181–190). Nova Science Publishers, Hauppauge, NY (2010).
17. Claveria, O., Monte, E., Torra, S.: A new approach for the quantification of qualitative measures of economic expectations. *Quality & Quantity*, 51(6), 2685–2706 (2017).
18. Coxeter, H. S. M.: *Introduction to Geometry* (2nd Edition). John Wiley & Sons, London (1969).
19. Chua, C. L., Kim, D., Suardi, S.: Are empirical measures of macroeconomic uncertainty alike?. *Journal of Economic Surveys*, 25(4), 801–827 (2011).
20. Dovern, J.: A multivariate analysis of forecast disagreement: Confronting models of disagreement with survey data. *European Economic Review*, 80, 1–12 (2015).
21. Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., Rubio-Ramírez, J.: Fiscal volatility shocks and economic activity. *American Economic Review*, 105(11), 3352–3384 (2015).
22. Giordani, P., Söderlind, P.: Inflation forecast uncertainty. *European Economic Review*, 47(6), 1037–1059 (2003).
23. Girardi, A., Reuter, A.: New uncertainty measures for the euro area using survey data. *Oxford Economic Papers*, 69(1), 278–300 (2017).
24. Glas, A., Hartmann, M.: Inflation uncertainty, disagreement and monetary policy: Evidence from the ECB Survey of Professional Forecasters. *Journal of Empirical Finance*, 39(Part B), 215–228 (2016).
25. Glass, K., Fritsche, U.: Real-time information content of macroeconomic data and uncertainty: An application to the Euro area. DEP (Socioeconomics) Discussion Papers, Macroeconomics and Finance Series 6/2014, University of Hamburg (2014).
26. Hartmann, M., Herwartz, H., Ulm, M.: A comparative assessment of alternative ex ante measures of inflation uncertainty. *International Journal of Forecasting*, 33(1), 76–89 (2017).

27. Henzel, S., Rengel, M.: Dimensions of macroeconomic uncertainty: A common factor analysis. *Economic Inquiry*, 55(2), 843–877 (2017).
28. Jurado, K., Ludvigson, S., Ng, S. Measuring uncertainty. *American Economic Review*, 105(3), 1177–216 (2015).
29. Kahneman, D. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY (2011).
30. Karnizova, L., Khan, H.: The stock market and the consumer confidence channel: Evidence from Canada. *Empirical Economics*, 49(2): 551–573 (2015).
31. Knight, F. H. (1921). *Risk, uncertainty, and profit*. Hart, Schaffner & Marx, Houghton Mifflin Company, Boston, MA (1921).
32. Lahiri, K., Sheng, X.: Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics*, 25(4), 514–38 (2010).
33. Mankiw, N. G., Reis, R., Wolfers, J.: Disagreement about inflation expectations. In M. Gertler and K. Rogoff (Eds.), *NBER Macroeconomics Annual 2003* (pp. 209–248). MIT Press, Cambridge, MA (2004).
34. Mokinski, F., Sheng, X., Yang, J.: Measuring disagreement in qualitative expectations. *Journal of Forecasting*, 34(5), 405–426 (2015).
35. Perić, B. Š., Sorić, P. A note on the “Economic Policy Uncertainty Index”. *Social Indicators Research*, In Press (2017).
36. Rich, R., Tracy, J.: The relationships among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts. *Review of Economics and Statistics*, 92(1), 200–207 (2010).
37. Saari, D. G.: Complexity and the geometry of voting. *Mathematical and Computer Modelling*, 48(9–10): 551–573 (2008).
38. Sorić, P., Lolić, I.: Economic uncertainty and its impact on the Croatian economy. *Public Sector Economics*, 41(4), 443–477 (2017).
39. Theil, H.: Recent experiences with the Munich business test: An expository article. *Econometrica*, 23(2), 184–192 (1955).
40. Zarnowitz, V., Lambros, L. A.: Consensus and uncertainty in economic prediction. *Journal of Political Economy*, 95(3), 591–621 (1987).

Prediction of crime from time series data-driven model ^{*}

Grzegorz Borowik¹, Zbigniew M. Wawrzyniak², Paweł Cichosz²,
Radosław Pytlak², Eliza Szczechla³, Paweł Michalak³,
Dobiesław Ircha³, and Wojciech Olszewski³

¹ Police Academy in Szczytno, Szczytno, Poland
g.borowik@wspol.edu.pl

² Warsaw University of Technology, Warsaw, Poland
z.wawrzyniak@ise.pw.edu.pl, p.cichosz@elka.pw.edu.pl,
r.pytlak@mini.pw.edu.pl

³ Scott Tiger S.A., Warsaw, Poland
eliza.szczechla@tiger.com.pl, pawel.michalak@tiger.com.pl,
dobieslaw.ircha@tiger.com.pl, wojciech.olszewski@tiger.com.pl

Abstract. The paper presents a short-term prognosis of the crime rate for subsequent two years in a selected region of Poland. The estimate of the future crime rate has been developed as data-driven model on the basis of past crime time series covering the period of three years. In our research we have focused on certain types of crime. The vast majority of crime categories experienced a downward trend in recent years, therefore the majority of forecasts assume further decline. The methods used for prognoses combine time series regressive methods with generalized additive models.

Keywords: crime prediction, time series, generalized additive model

1 Introduction

The effective operation of public order institutions, especially the Police, is based on the rapid collection, processing, analysis and deployment of events, information and characteristics from the real world as well as attributes of PESTLE analysis [1,2,3].

Analyzing crime as a complex phenomenon requires looking at the environment from different perspectives: legal, economic, social and technological [2]. The prediction of crime rate from time series is an important element of predictive policing, and resource management optimization for the Police and other law enforcement agencies [4].

Efficient IT tools and other technical means guarantee the efficiency and effectiveness of operations. At the same time, modern information systems that

^{*} Supported by the Polish National Center for Research and Development under grant DOB-BIO7/05/02/2015.

allow obtaining information, introduce new standards, both in the area of reaction to events and the subsequent analysis. The quality of information plays a very important role in this scope, in particular such criteria as: effectiveness, efficiency, confidentiality, integrity, availability, compliance, reliability.

Reliable and high quality forecasts to crime activity are based on variety of time series and criminal analysts with expertise in time series modeling [5,6,7]. A practical approach to forecasting needs to combine configurable models with experimental and analyst knowledge. By fusing a modular regression model with interpretable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series, the most effective tool as data-driven model can be constructed.

To formulate the forecasts, methods based on extrapolation were used in which the basic assumption is to continue trends – the future events will happen because they had happened in the past. Unlike structural models, extrapolation omits structural changes in the studied phenomenon and does not predict changes in the trend, which is its weakness. Therefore, these methods are usually applied to short-term forecasts.

The object of this paper is the short-term prognosis for crimes events for the years 2017–2018. The forecast of future crimes has been developed on the time series covering the period 2013–2016. The data used is a representative subset of police registers in Poland. In research, we focus on police interventions, burglary, hooliganism, and traffic offenses.

For subsequent manual review and adjustment of the results, one must use a tool to evaluate forecasting procedures and automatic forecasts as well as analyze the performance of the results. In our study, we used non-linear regression combined with generalized additive model in the form of “prophet” package [8,9,10] implemented in Python [11]. This analysis yields trends and seasonality of events.

2 Methodology

To perform a short-term point forecast, one can use exponential smoothing methods and ARIMA class models [12,13].

In the method of single exponential smoothing (Brown’s method [14]), the forecast is calculated as a weighted average, where the weights decrease exponentially for observations from earlier periods. This means that the smallest weights are assigned to the oldest observations [15]:

$$\hat{y}_{t+1|t} = \alpha y_t + (1 - \alpha)\hat{y}_t = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha^2)y_{t-2} + \dots, \quad (1)$$

where $0 \leq \alpha \leq 1$ is a smoothing parameter. This method works best for data without the observed trend or seasonality.

Proposed by Ch.C. Holt Double exponential smoothing [16], is an extended method of exponential smoothing and allows forecasting of data characterized

by the trend. Then,

$$\begin{aligned} \hat{y}_{t+h|t} &= l_t + hb_t, \\ l_t &= \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}), \\ b_t &= \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}, \end{aligned} \tag{2}$$

where l_t denotes estimation of the level of the series at time t ; b_t is an estimate of the trend coefficient at time t ; α is a smoothing parameter for the level $0 \leq \alpha \leq 1$, while β is a smoothing parameter for the trend $0 \leq \beta \leq 1$.

There exists modification to this method, such as double exponential smoothing with a fading trend. It is suitable for a time series with a trend in which it is believed that the continuation of the growth dynamics observed at the end of historical data is unlikely outside the short-term forecast [4].

Autoregression model with moving average – ARIMA (AutoRegressive Integrated Moving Average) – proposed by G.E.P. Box and G.M. Jenkins [12] – is a combination of two processes: autoregressive AR and moving average MA for integrated time series. ARIMA class models (p, d, q) without seasonality can be represented by the following equation:

$$\phi(L)(1 - L)^d y_t = \theta(L)\epsilon_t, \tag{3}$$

where: $\phi(L) = (1 - \sum_{i=1}^p \phi_i L^i)$ is the autoregressive part of the order p for AR(p) model; L is a delay operator; $(1 - L)^d$ denotes a differential operator of order d ; $\theta(L) = (1 + \sum_{i=1}^q \theta_i L^i)$ is the average moving process of the order q for the model MA(q); ϵ_t denotes random error.

2.1 Proposed approach

In our study, we have used non-linear regression implemented in the Python packages. This analysis yields trends and seasonality of events. In the research we applied a decomposition of a time series model that was proposed in [13]. The model characterizes with three main components, i.e., trend, seasonality, and holidays. As was shown in [13] these components can be combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t. \tag{4}$$

Function $g(t)$ represents the trend that models non-periodic changes, function $s(t)$ represents periodic changes (e.g., weekly and yearly seasonality), and function $h(t)$ represents the effects of holidays. The error ϵ_t represents any idiosyncratic changes which are not accommodated by the model.

The piecewise logistic growth model equals:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T \delta)(t - (m + a(t)^T \gamma)))}, \tag{5}$$

where $C(t)$ is time-varying capacity, k the growth rate, m an offset parameter, δ is a vector of rate adjustments, $a(t) \in \{0, 1\}^S$.

A crucial experiment was to perform seasonality analysis of the events. The domain of frequency is an alternative field of signal description and analysis, closely related to the domain of time. Frequency analysis methods in the frequency domain are called frequency methods or spectral methods. They are more effective tools for studying signals than time domain methods. The mathematical formalism of Fourier's integral transformations is a basement for frequency analysis methods of continuous signals. They determine mutual relations between time and frequency domain. Fourier's integral transformation known for over two hundred years remains the fundamental and most widespread signal analysis tool.

We rely on the Fourier series to provide a flexible model of periodic effects. Let P be the regular period we expect the time series to have (e.g. $P = 365.25$ for yearly data or $P = 7$ for weekly data, when we scale our time variable in days). We can approximate arbitrary smooth seasonal effects with:

$$s(t) = \sum_{n=1}^N \left(a_n \cos \left(\frac{2\pi nt}{P} \right) + b_n \sin \left(\frac{2\pi nt}{P} \right) \right). \quad (6)$$

2.2 Data

The analyzed data originate from the police data collection system. The basis of the system operation is gathering the information from open sources, citizens [18] and policemen about crimes and security threats received by the police units on duty, as well as findings made by specific police services in relation to these reports. An important information is also the data of planned and implemented activities of the police services, as well as the results obtained. The purpose of the system is to streamline the organization process and planning the activities of the preventive services [17].

Implementation of the objectives of the system requires the processing of:

1. details of the notification:
 - the identifier of the notification;
 - priority assignment;
 - status assignment (ongoing, unreliable, dangerous, etc.);
 - place, time and description of the event;
 - event classification;
 - how was the event reported (by phone, in person, using electronic mail, etc.);
 - personal data identifying the reporting person;
2. operational data:
 - time and type of actions taken;
 - location of the action taken;

- basis of the action taken;
- 3. messages transmitted between users of the system;
- 4. information about queries performed by users of the system.

The system provides the following features:

- 24/7 access to information resources for eligible organizational units;
- continuous electronic logging of notifications and events;
- mobile access to resources;
- quick location of events and patrols on the digital map;
- generating reports and statistics;
- safety of information.

3 Experiments

In this section, we show elements of seasonality for time series as a way of parameterization of models containing the main information characterizing the processes of development and propagation of criminal activities (hotspots of crime). In this examination, we present the results for four types of events, i.e., police interventions, burglary, hooliganism, and traffic offenses. To conceal the actual crime rate, the data presented in charts is a representative subset of the total data from the data collection system.

The frequencies presented in Figs. 1, 2, 3, 4 show initial evidence of seasonalities, i.e., a day of the week, weekly seasonality, and yearly seasonality. For example, most police interventions (Fig. 1) have been carried out on a Saturday at 11 pm (heat map on Fig. 1 (bottom right)).

The rate of hooliganism is very much the same as the number of police interventions. As shown in Fig. 2, the highest intensity of hooliganism occurs on Saturdays around 11 pm; hooliganism is characterized by weekly seasonality as demonstrated by the frequency analysis. An increased number of hooliganism occurs in the summer months.

For the burglary category (Fig. 3), one may notice the events mostly at the beginning of the week (Monday). Most of the events happened around 10 am and 3 pm since the perpetrators exploit the knowledge that the owners are absent from home or apartment.

Traffic offenses (Fig. 4) happen mainly during working days between 7 am and 6 pm. We can also observe an increase in the number of traffic offenses around 3 pm due to the traffic jams as well the people commuting from work to shops or home. The seasonality of traffic offenses between 3 pm and 6 pm on Fridays is correlated to the beginning of the weekend and the commuting of people between cities. The yearly seasonality of traffic offenses increases in the warmer months as well as during Christmas and the All Saints' Day what is visible in the Fig. 4 (middle right).

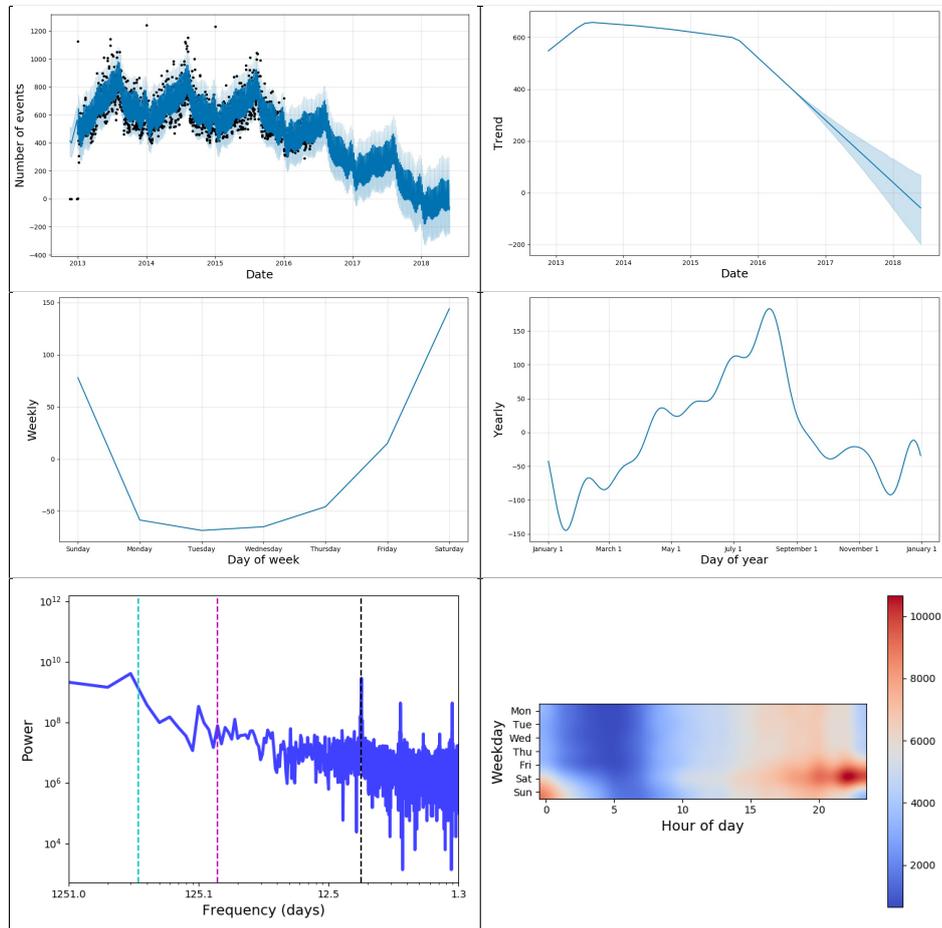


Fig. 1. Time series of police interventions – INT (upper left), trend of INT for subsequent years (upper right), weekly seasonality (middle left), yearly seasonality (middle right), frequency characteristics of INT (bottom left), and distribution of the relative difference of INT for weekday and hours profiles (bottom right)

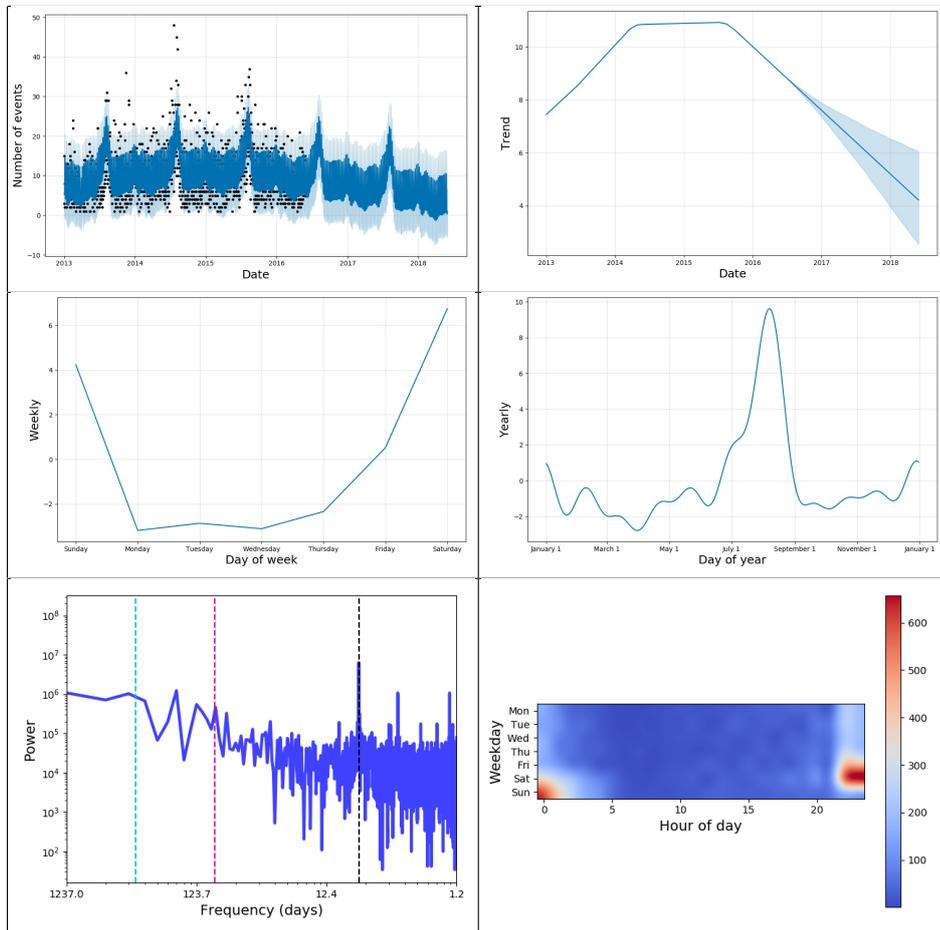


Fig. 2. Time series of hooliganism – HOO (upper left), trend of HOO for subsequent years (upper right), weekly seasonality (middle left), yearly seasonality (middle right), frequency characteristics of HOO (bottom left), and distribution of the relative difference of HOO for weekday and hours profiles (bottom right)

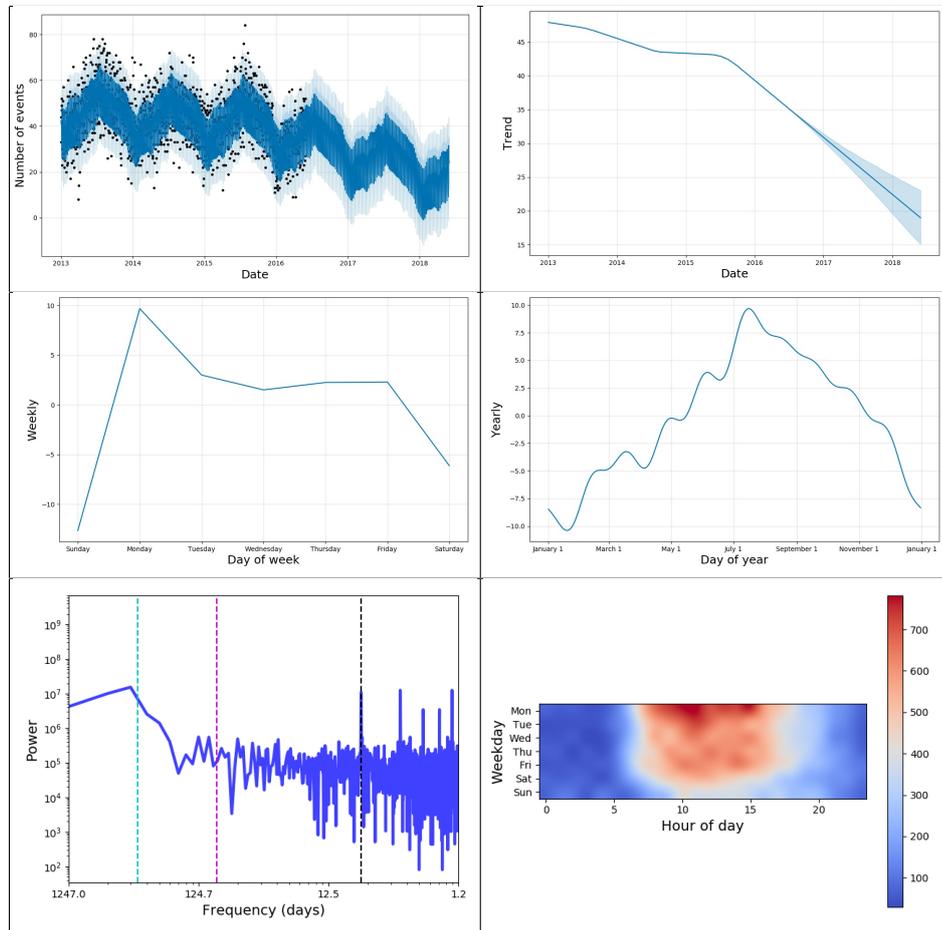


Fig. 3. Time series of burglary – BUR (upper left), trend of BUR for subsequent years (upper right), weekly seasonality (middle left), yearly seasonality (middle right), frequency characteristics of BUR (bottom left), and distribution of the relative difference of BUR for weekday and hours profiles (bottom right)

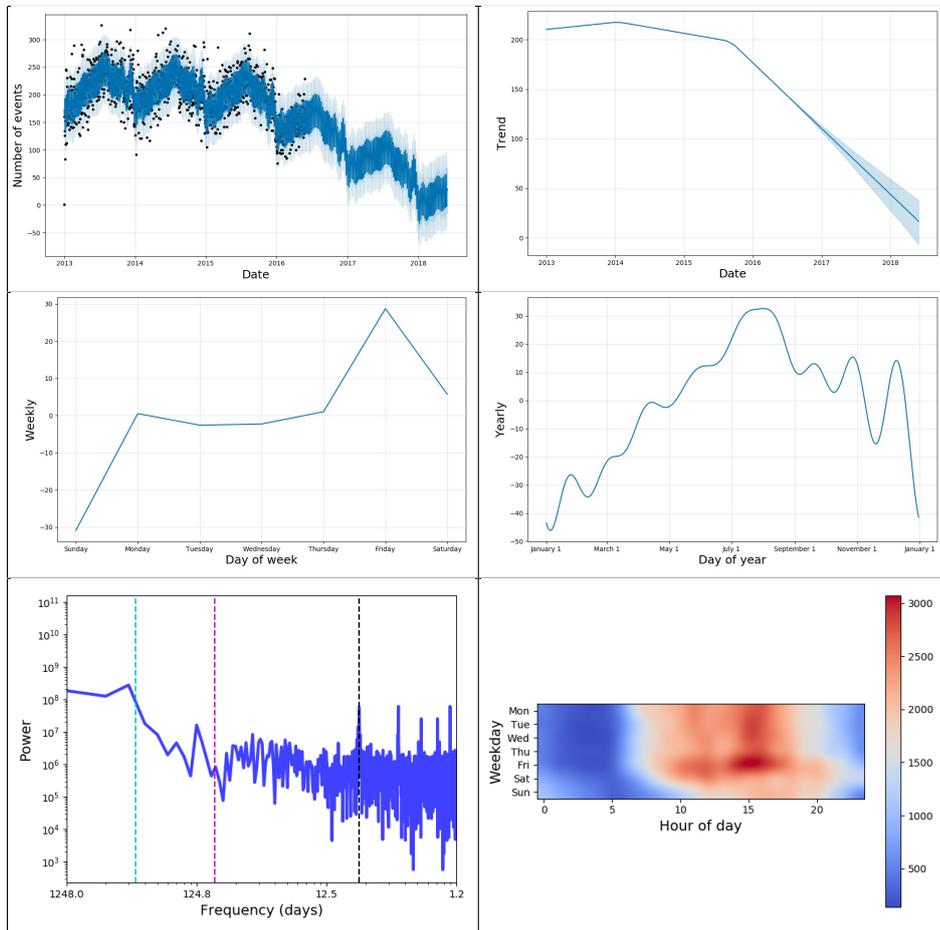


Fig. 4. Time series of traffic offenses – TRA (upper left), trend of TRA for subsequent years (upper right), weekly seasonality (middle left), yearly seasonality (middle right), frequency characteristics of TRA (bottom left), and distribution of the relative difference of TRA for weekday and hours profiles (bottom right)

4 Conclusions

The applied forecasting method leverages information about past crime events to predict the future crime rate. The vast majority of crime categories saw a downward trend in recent years, therefore the majority of forecasts assume further decline. This does not apply to corruption (not included in the paper) for which the historical growing trend is continued. It should be noted that these forecasts are potential direction indicators but cannot be treated as an approximation of specific values but rather as a data-driven model. This applies especially to the last years of the forecast, which are quite distant from the last observation when it comes to extrapolation of time series. The time series methods work best in the short term forecast, therefore the forecast for the first year is much more likely than for the last year of the forecast. Moreover, when constructing a model-based forecast that implements a seasonal-trend decomposition in time series and uses analytical processes from attributes of PESTLE analysis, the model needs to be parameterized and calibrated on the basis of high-quality real data originated from Police and open sources.

References

1. Buonanno, P., Montolio, D.: Identifying the socio-economic and demographic determinants of crime across spanish provinces. *International Review of Law and Economics* **28**(2) (2008) 89–97
2. Blanco, J., Cohen, J.: Macro-environmental factors driving organised crime. In Larsen, H., Blanco, J., Pastor, P., Yager, R., eds.: *Using Open Data to Detect Organized Crime Threats*
3. Kruse, M., Svendsen, A.: Foresight and the future of crime: Advancing environmental scanning approaches. In Larsen, H., Blanco, J., Pastor, P., Yager, R., eds.: *Using Open Data to Detect Organized Crime Threats*
4. Ahmad, F., Syal, S., Tinna, M.S.: Criminal policing using rossmo’s equation by applying local crime sentiment. In Satapathy, S.C., Bhateja, V., Raju, K.S., Janakiramaiah, B., eds.: *Data Engineering and Intelligent Computing*, Singapore, Springer (2018) 627–637
5. Nakaya, T., Yano, K.: Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* **14**(3) (2010) 223–239
6. Caplan, J.M., Kennedy, L.W., Miller, J.: Risk terrain modeling: brokering criminological theory and gis methods for crime forecasting. *Justice Quarterly* **28**(2) (2011) 360–381
7. Kang, H.W., Kang, H.B.: Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* **12**(4) (2017) e0176244
8. Taylor, S.J., Letham, B.: Forecasting at scale. *The American Statistician* **72**(1) (2018) 37–45
9. Homepage: <https://facebook.github.io/prophet/>.
10. Prophet R package: <https://cran.r-project.org/package=prophet>.
11. Prophet Python package: <https://pypi.python.org/pypi/fbprophet/>.

12. Box, G.E.P., Jenkins, G.: *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., San Francisco, CA, USA (1990)
13. Harvey, A., Peters, S.: Estimation procedures for structural time series models. *Journal of Forecasting* **9** (1990) 89–108
14. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S.: A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18**(3) (2002) 439 – 454
15. Gardner, E.S., McKenzie, E.: Forecasting trends in time series. *Manage. Sci.* **31**(10) (1985) 1237–1246
16. Gelper, S., Fried, R., Croux, C.: Robust forecasting with exponential and holt-winters smoothing. *Journal of Forecasting* **29**(3) (2010) 285–300
17. Hołyst, B.: *Prognozowanie kryminologiczne w wymiarze społecznym – Metodologia, Analiza, Tendencje rozwojowe. Tom 1* (in Polish). Wydawnictwo Naukowe PWN S.A. (2017)
18. Larsen, H.L., Blanco, J.M., Pastor, R.P., Yager, R.R.: *Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime*. Springer (2017)
19. Blanco, J.M., Cohen, J.: Macro-environmental factors driving organised crime. In: *Using Open Data to Detect Organized Crime Threats*. Springer (2017) 137–166
20. Schroeder, J., Xu, J., Chen, H., Chau, M.: Automated criminal link analysis based on domain knowledge. *Journal of the Association for Information Science and Technology* **58**(6) (2007) 842–855
21. Iriberry, A., Navarrete, C.J.: Internet crime reporting: Evaluation of a crime reporting and investigative interview system by comparison with a non-interactive reporting alternative. In: *System Sciences (HICSS), 2010 43rd Hawaii International Conference on, IEEE* (2010) 1–9
22. Sakpere, A.B., Kayem, A.V., Ndlovu, T.: A usable and secure crime reporting system for technology resource constrained context. In: *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on, IEEE* (2015) 424–429
23. Cesario, E., Catlett, C., Talia, D.: Forecasting crimes using autoregressive models. In: *Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C, IEEE* (2016) 795–802

Measurement and Modelling of Business Cycles using Linear and Nonlinear methods

Nomeda Bratčikoviene

Vilnius Gediminas Technical University, Vilnius Lithuania
nomeda.bratcikoviene@vgtu.lt

Abstract. This paper focuses on business cycle component analysis and modelling in small economy countries. Lithuanian economy is one of such example with specific properties. The list of major leading indicators business cycles was studied in this paper in the case of Lithuania. The results suggest that economical methods aren't suitable for modelling of main indicators business cycles in such cases. Standard univariate methods are preferable. However various time series business cycles may be nonlinear, therefore linear ARIMA and nonlinear SETAR methods were used and availability to capture features of the real time series business cycle with complex nature was tested.

Keywords: ARIMA, SETAR, Business Cycle.

1 Introduction

The aim of structural time series models is estimation of unobserved times series components, such as trend, cycle, seasonal, irregular. These components have a natural interpretation: trend forms general long-term tendency, seasonal component is defined as repetitive and predictable movement during year, irregular component provides random nature of the time series.

The empirical definition of business cycle was developed during Great Depression (see Schumpeter [1], Haberler [2], Burns and Mitchel [3]). Business cycle was defined as a type of fluctuation in the aggregate economic activity that consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle [3]. Business cycles occur at a certain frequency, they are not periodic or regular and their size or length is not standard or similar across countries or in time.

Cycle forecasting is the process of estimating predictions about the future of the country economy. A good comprehension of business cycle facts, correct predictions of leading economic indicators business cycle turning points are essential for macroeconomic decision making, especially in monetary and fiscal policies and this leads to better economic country development.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Often, economic forecasting involves the use of complex mathematical models that consider many different economic indicators. "Leading indicators" are variables that tend to change before the economy as whole, which are commonly used in business cycle forecasting. Examples of leading indicators include the federal funds rate, interest rates, hours worked, unemployment claims, building permits for new homes and crude oil prices, etc.

Depending on the objectives and aim of the scientific research a variety of methods can be applied for business cycle evaluation, analysis and forecasting: from simple univariate to general equilibrium models. The main groups of business cycle investigation methods can be separated into statistical (linear, nonlinear or multivariate time series methods) and economical (economic indicators, composite indicators) methods.

Economical methods have some limitations. Firstly, these methods have insufficient theoretical foundation as statistical methods. They act mostly on economic logic. Secondly, the data of many indicators that can be used in economic or composite indicators method are available too late. Sometimes time series of such indicators can be too short for sufficient forecasts. And finally, indicators developed in some countries are not necessarily act in the exact same way in other countries and economies [4]. However, economic logic couldn't be denied and we will follow it with some adjustments.

Multivariate statistical methods, for example structural vector autoregressive (SVAR) models are useful when the aim of the research is to analyse the dynamics of a model by subjecting it to an unexpected shock [5], quantifying impulse-responses of shocks or measuring the contributions of shocks to fluctuations of economic variables.

In this paper we will focus on cyclical component analysis, its modelling and forecasting that can be used for policy simulations. For such an aim standard univariate methods works better when multivariate. Linear and nonlinear modelling methods will be applied for comparison.

Social, economic, political and other changes that occur leave business cycle asymmetries or can change dynamic of economic time series. Such features cannot be captured by conventional linear models with constant parameters. Linear models, which allow infrequent structural changes in the parameters (see [6–9] or non-linear models (see [10])) can help to solve these problems. In this paper we will compare classical linear integrated autoregressive moving average (ARIMA) versus self-exciting threshold autoregressive (SETAR) business cycle modelling results.

The paper is organized as follows. Business cycle definition and its most commonly used modelling methods are described in the Introduction. In Section 2 we introduce linear ARIMA and nonlinear SETAR methods. Section 3 provides details of empirical modelling. And finally, Section 4 contains some concluding remarks and prepositions.

2 Brief Introduction of Modelling Methods

2.1 Integrated autoregressive moving average models

Auto-Regressive Integrated Moving Average (ARIMA) models are the powerful class of modelling and forecasting techniques which can be applied to many real time series.

If L is time lag operator defined as

$$L^k X_t = X_{t-k} \quad (1)$$

and difference operator

$$\Delta X_t = X_t - X_{t-1} = (1 - L)X_t \quad (2),$$

then an AutoRegressive Integrated Moving Average with autoregressive order p , integration order d , and moving average order q (ARIMA (p, d, q)) model can be written as discrete time linear equation:

$$\left(1 - \sum_{i=1}^p \alpha_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \beta_i L^i\right) \varepsilon_t \quad (3)$$

with noise ε_t .

The Box-Jenkins methodology can be applied for identification and estimation of time series models within the class of ARIMA models. The set of procedures uses an iterative approach and contains of four stages [11]:

- Step 1. Identification. This step involves determination of the ARIMA model order (p, d, q), mostly by using graphics of time series, AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF), etc.
- Step 2. Estimation and selection. It is responsible for estimation of the parameters of the models that have been indentified in step 1 and proceeds to a first selection of models by using certain information criteria.
- Step 3. Diagnostic checking includes determination of specified and estimated model is adequacy. Residual diagnostics is most often performed followed by the inference and final model selection.
- Step 4. Model's use for analysis and forecasting.

2.2 Self-Exciting Threshold AutoRegressive model

One of the popular families of modelling methods used in non-linear time series modelling is multi regime forecasting models that allows smooth transition from one linear regime to the other. These models were proposed by Bacon and Watts [12]. Threshold Auto Regressive (TAR) model were introduced by Tong [13] and extensively discussed in [14–16]. A basic feature of these models is that they allow for an

application of regime-switching and can describe the different dynamic behaviour of analysed time series in this way.

Self-Exciting Autoregressive (SETAR) model is a special case of TAR models. The class of SETAR models has been widely used for various economic time series and especially for financial time series and analysis and modelling. Basically SETAR models are linear autoregressive models in which the linear relationships varies over regimes depending on the threshold values. The regime is determined by the past values of the time series in SETAR models.

SETAR process X_t can be defined by equation:

$$X_t = (1 - I(L^d X_t > r)) \left(1 - \sum_{i=1}^{p_1} \alpha_{1i} L^i \right) X_t + \varepsilon_{1t} + I(L^d X_t > r) \left(1 - \sum_{i=1}^{p_2} \alpha_{2i} L^i \right) X_t + \varepsilon_{2t} \quad (4)$$

where $I(L^d X_t > r) = 1$ if $L^d X_t > r$ and zero otherwise. ε_{1t} and ε_{2t} are sequences of independent and identically distributed random variables. Positive integer d is delay parameter – transition variable that governs changes in regime. r is the threshold value. For a given threshold r and the position of $L^d X_t$ with respect to this threshold r , the time series X_t follows AR(p_1) model or an AR(p_2) model. The model parameters are α_{1i} , for $i = 1; 2$, and $j = 1, \dots, p_k$, where $k = 1; 2$ and the delay d and threshold r . The set of allowable threshold values r should be such that each regime must contain enough quantity of observations for the evaluation of reliable estimates of autoregressive parameters. A popular choice of r is to require that each regime must contain at least a fraction π of the observations, that is $r \in \{r | X_{[\pi(N-d)]} \leq r \leq X_{[(1-\pi)(N-d)]}\}$, where $X_{(0)}, X_{(1)}, \dots, X_{(N-d)}$ denote the order statistics of the threshold variable $X_{(N-d)}$, $X_{(0)} \leq X_{(1)} \leq \dots \leq X_{(N-d)}$ and $[.]$ denotes integer part. A safe choice for this fraction appears to be 0.15 [17].

Threshold value r and delay parameter d can be identified and SETAR model selected by using Akaike information criterion (AIC). AIC for SETAR models are defined as the sum of AIC's for the AR models in the different regimes [15]:

$$AIC(p_1, p_2) = n_1 \ln \hat{\sigma}_1^2 + n_2 \ln \hat{\sigma}_2^2 + 2(p_1 + 1) + 2(p_2 + 1) \quad (5)$$

there $\hat{\sigma}_j^2$, $j = 1; 2$, is the variance of the residuals in the j th regime. AIC must achieve its minimum value for selected threshold parameter r and delay d .

3 Empirical results

3.1 Description of real time series used in practical analysis

The importance of business cycle studies of a global economy does not lose its importance for business cycle research in separate countries. Business cycle analysis is especially relevant in Lithuania, because we have some special features of our economy in Lithuania. Lithuania has relatively short history as a free-market economy – independence of Lithuania was restored only in 1991 and free-market economy started to act a bit later. That is the reason why our time series of economic indicators aren't long. In addition, our economy survived two economic crises – Russian crisis in 1998-1999 and the global financial crisis in 2008-2011 within a relatively short period of time. It is therefore essential to determine the best methods for understanding and predicting of movements of business cycle in Lithuania. These questions are not fully answered yet.

Lithuanian business cycle may have specific properties that are different from the behaviour of the developed economies countries business cycles. In this paper we will try to examine the list of major leading indicators [18] of developed economy countries in the case of Lithuania business cycle. These results will be compared with the results of linear and nonlinear models.

The business cycles of Final consumption expenditure, Investment in tangible fixed assets at current prices, Unemployment, Number of hours worked per employee in Lithuania, Harmonised indexes of consumer prices, Number of new residential buildings for which building permits were granted, Money supply (M2) and Interest rates on new business loans to non-financial corporations and households were compared with, the business cycle of gross domestic product at current prices in Lithuania, to check hypothesis for developed economy countries leading indicators acting in Lithuanian. Business cycles of all indicators were estimated by using Hodrick-Prescott filter [19]. The graphs of these indicators are presented in the Figure 1.

Analysis of business cycles in Lithuania showed that most of analysed indicators business cycles are coincident or depend on time – in some moments of time it can be leading, lagging or coincident. Business cycles of number of hours worked and unemployment rate (except last two years) are lagging. These results show that standard developed economy methods for business cycles analysis and prediction are not valuable in Lithuania, because of described reasons above. An alternative way of business cycle investigation is forecasting. Reliable models for prediction of business cycles must be found.

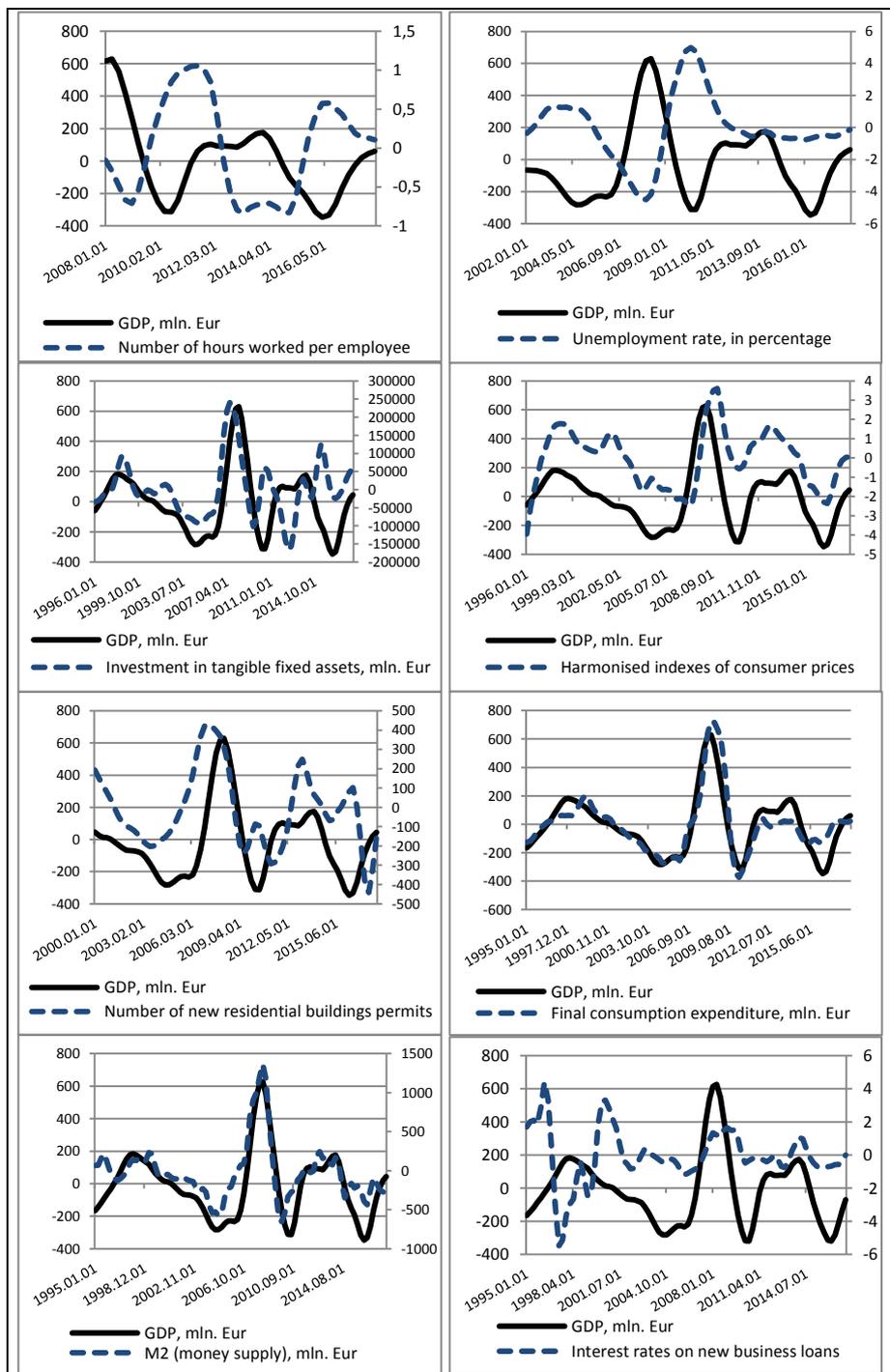


Fig. 1. Business cycles of main indicators in Lithuania

3.2 Empirical results of linear and nonlinear models of whole sample

The results of linear ARIMA and nonlinear SETAR modelling will be presented in this subsection. For comparison of models mean absolute percentage error (MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| * 100 \quad (6)$$

will be used.

ARIMA model identification was made depending on autocorrelation function (ACF) and partial autocorrelation function (PACF). Models were selected by using Akaike information criterion. AIC was chosen for comparability with nonlinear SETAR model and because the sample of modelled indicators is quite small. Under unstable conditions such as small sample and large noise levels Akaike information criterion outperforms Bayesian information criteria (BIC) [20]. Residuals of estimated models were tested for normality and independence.

SETAR models and parameters were selected depending on Akaike information criterion (5) for SETAR models. SETAR and ARIMA models were estimated using R package.

On purpose not to extend the scope of this article, here we show estimated models only for main indicator – Gross Domestic Product (GDP).

Lithuanian GDP business cycle time series is not stationary – it was stacionarized by using second order differencing. According to Box-Jenkins procedure and AIC, the best linear ARIMA model of Lithuanian GDP business cycle - ARIMA(2, 2, 0):

$$(1 - 1,2027B + 0,5489B^2)(1 - B)^2 X_t = \varepsilon_t \quad (7)$$

here B – backward shift operator ($BX_t = X_{t-1}$). $MAPE = 13.1$.

According to previously described procedure, nonlinear SETAR model with two regimes of Lithuanian GDP business cycle was estimated:

$$X_t = \begin{cases} 103.967 + 1.264 X_{t-1}, & \text{if } X_{t-2} \leq -160.911 \\ -17.141 + 1.021 X_{t-1}, & \text{if } X_{t-2} > -160.911 \end{cases} \quad (8)$$

$MAPE = 22.35$.

Analysis of SETAR and ARIMA models errors showed that ARIMA model is relatively more accurate than the SETAR model for GDP business cycle time series. Range of errors of ARIMA model is significantly smaller than range of SETAR model errors. Graphs of model estimates are presented in the Figure 2:

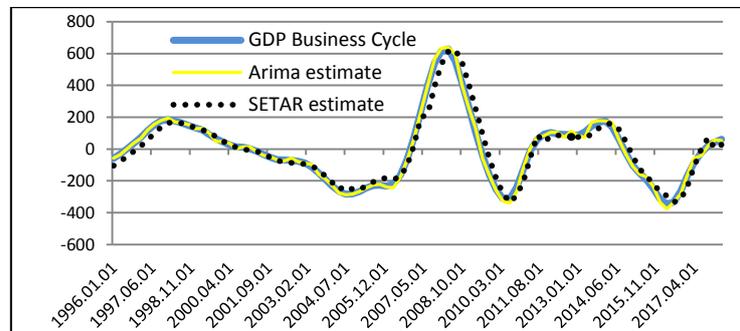


Fig. 2. Modelling results of Lithuanian GDP Business Cycle

4. Concluding remarks

Most of Lithuanian time series are complex in their nature, therefore business cycle of economic indicators in Lithuania have specific properties. Lists of major leading indicators that are used in prediction of developed economies countries business cycles were tested in Lithuania case. The results showed that most of analysed leading indicators business cycles are coincident or depend on time, therefore standard developed economy methods for business cycles analysis and prediction are not valuable in Lithuania. An alternative way of business cycle investigation is forecasting.

Linear ARIMA and nonlinear SETAR methods were used for an estimation of main indicators business cycles in Lithuania. Analysis of errors showed that ARIMA model is relatively more accurate and reliable than the SETAR model for GDP business cycle time series.

References

1. Schumpeter, J.A.: *Business Cycle*, New York: McGraw-Hill Book Co (1939),.
2. Haberler, G. (ed.): *Readings in Business Cycle Theory*, Philadelphia: The Blakiston Company (1944).
3. Burns, A. F., Mitchell, W. C.: *Measuring business cycles*, NBER Studies in Business Cycles, no 2, New York (1946)
4. Batchelor, R.: Confidence indexes and the probability of recession: a Markov switching model. In Dua, P., editor, *Business cycles and economic growth: an analysis using leading indicators*, pages 207_225. New Delhi: Oxford University Press (2004)
5. Gottschalk, J.: An Introduction into the SVAR Methodology: Identification, Interpretation and Limitations of SVAR models. In: *Kiel Working Papers*. (2001).
6. Bai, J., Perron, P.: Estimating and testing linear models with multiple structural changes, *Econometrica*, 66, pp. 47–78, (1998).
7. Culver, S.E., Papell, D.H.: Is there a unit root in the inflation rate? Evidence from sequential break and panel data models, *J. Appl. Econom.*, 12, pp. 435–444, (1997).
8. Emerson J., Chihwa K.: Testing for structural change in panel data: GDP growth, consumption growth, and productivity growth, *Economics Bulletin*, 3(14), pp. 1–12 (2006)
9. Perron, P., Zhongjun, Q.: Estimating and testing structural changes in multivariate regressions, *Econometrica*, 75(2), pp. 459–502 (2007).

10. Chatfield, Ch.: *The Analysis of Time Series: An Introduction*, CRC Press, Boca Raton, FL (2004).
11. Kaiser, R., Maravall, A.: *Notes on Time Series Analysis, ARIMA Models and Signal Extraction*, Banco de Espana, No 0012 (2000).
12. Bacon, D. W., Watts, D. G.: Estimating the transition between two intersecting lines, *Biometrika*, 58, pp. 525–534 (1971).
13. Tong, H.: On a threshold model, in: C.H. Chen (Ed.), *Pattern Recognition and Signal Processing*, pp. 101–141 Amsterdam: Sijthoff and Noordhoff, (1978).
14. Tong, H.: *Threshold Models in Non-Linear Time Series Analysis*, New York: Springer-Verlag (1983).
15. Tong, H.: *Non-Linear Time Series: A Dynamical Systems Approach*, Oxford: Oxford Univ. Press (1990).
16. Tong H., A personal overview of non-linear time series analysis from a chaos perspective, *Scand. J. Stat.*, 22, pp. 299–445 (1995).
17. Franses, P. H., Dijk, D.: *Non-Linear Time Series Models in Empirical Finance*, Cambridge: Cambridge Univ. Press (2000).
18. Klein, P. A.: *The Leading Indicators in Historical Perspective*, *Business Cycle Indicators Handbook*, pp. 23-29, New York: Conference Board, (2001).
19. Kaiser, R., Maravall, A.: *Measuring business cycles in economic time series*, *Springer Science & Business Media* (2001).
20. de-Graft Acquah, H.: Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship, *Journal of Development and Agricultural Economics* Vol. 2(1) pp. 001-006, (2010).

Examination of forecasting in education field

W. TEROUZI, F. MAHJOUBI, A.OUSSAMA

1 Laboratory of Spectro- Chemometrics and environment, Faculty of Science and Technology of Beni Mellal, University of Sultan Moulay slimane, 21000- Beni Mellal, Morocco.

E-mail : w.terouzi@usms.ma

ABSTRACT

1. INTRODUCTION

Certainly, each manager prefers to know the exact nature of future events in order to take measures or plan his actions when enough time is available to implement the plan. The effectiveness of his plan depends on the level of precision with which future events are known to him. But each manager plans for the future regardless of whether future events are known or not. This implies that he is trying to forecast the future from his best abilities, judgment and experience. Generally, forecasting starts with certain assumptions based on the management's experience [1]. Additionally, in the developed world, forecasting is utilized in several tasks related to education field. In this context, the goal of this work is to describe the main steps, technical tools used in forecasting approaches for different applications in education area and determine the most effective ones for our current research data.

2. SOME APPLICATIONS IN EDUCATION AREA

- ✓ Forecasting enrollment in higher education [2], [3];
- ✓ Predicting Student's Performance [4] ;
- ✓ Forecasting models for evidence based policy of education system [5];
- ✓ Forecasting Education and Training Needs in Transition Economies....

3. FORECASTING MODELS

Forecasting techniques can be classified into two types:

a)-**Qualitative Forecasting Methods** : Are based on judgments, opinions, intuition, emotions, or personal experiences and are subjective in nature. They do not rely on any rigorous mathematical computations.

b)- **Quantitative Forecasting Methods** : Are based on mathematical (quantitative) models, and are objective in nature. They rely heavily on mathematical computations.

In present study, we try to forecast the total number of students in the higher school of technology during enrollment / re-enrollment, and the specialties / courses to be created during the next 5 years. In this case we looked at different tools to find suitable methods.

4. CONCLUSION

A good education institute must plan its activities and areas of concentration for the coming years according to the expected demand for its various activities. The institute can, then, come out with a forecast on the future needs in the labor market, for its students who graduate. This may require a reorientation of the formation and faculty, the development of appropriate teaching materials, the recruitment of new teachers with sectoral backgrounds, experience and specific pedagogical skills.

REFERENCES

- [1] D. W., & A. S. (1994). *Forecasting: The Key to Managerial Decision Making. Management Decision*, 41-49.
- [2] Rabby Q. Lavilles and Mary Jane B (2012). *Arcilla. Enrollment Forecasting for School Management System. International Journal of Modeling and Optimization*, Vol. 2, No. 5, 563-566.
- [3] Norhaidah Abu Haris, Munaisyah Abdullah, Abu Talib Othman and Fauziah Abdul Rahman (2014). *Application of Forecasting Technique for Students Enrollment. Knowledge Management International Conference (KMICe)*, 12 – 15.
- [4] Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid (2015). *The Third Information Systems International Conference - A Review on Predicting Student's Performance using Data Mining Techniques. Procedia Computer Science* 72, 414 – 422
- [5] Gintautas DZEMYDA, Vydūnas ŠALTENIS, Vytautas TIEŠIS (2003). *Forecasting Models in the State Education System. Informatics in Education*, Vol. 2, No. 1, 3–14

Time Series Versus Causal Forecasting: An Application of Artificial Neural Networks

Prithviraj Lakkakula

Department of Agribusiness and Applied Economics, North Dakota State University
811 2nd Ave N., Fargo, ND 58102, Tel.: +1 701-231-6642,
Email: prithviraj.lakkakula@ndsu.edu

Abstract. This paper uses artificial neural networks to evaluate the performance of soybean and high fructose corn syrup (HFCS) price forecasting using both causal and time series techniques. Mean absolute deviation, and mean square error (both in-sample and out-of-sample) are used to evaluate the predictive accuracy of causal and time series neural networks. Based on the out-of-sample forecast performance, causal neural networks performed better in predicting both soybean and HFCS prices. To check the robustness of the results, a turning point test and sensitivity analysis are conducted. A turning point test is performed to evaluate the technique that captures the most turning points. Turning point test results indicate that the time series forecasting approach captures the most turning points for both soybean and HFCS prices. Finally, a sensitivity analysis is performed to analyze the relative importance of explanatory variables on soybean and HFCS prices. Results of this article help to guide forecasting approach selection based on research objectives.

Key words: Neural Networks; price forecasting; Out-of-sample performance; turning points

1 Introduction

Forecasting is an important area of research in various fields including economics, agriculture, computer science and other allied fields. Better forecasts lead to better decision making. For example, a prevalent strategy in any industry is to plan future inventory levels to maximize expected returns. However, obtaining better forecasts is a difficult task. Forecasting is even more difficult in agriculture as it is subjected to unpredictable weather, which increases volatility in yield and price.

Expectations about commodity prices, which are driven by several factors including expected demand, and supply, inform producers about crop production decisions. Commodity price forecasts are also important globally as they influence farm income. As commodity prices are frequently volatile, governments tend to support farmers to ease income volatility. For example, data compiled by the United States Department of Agriculture (USDA) indicate that net farm income projections decreased by 8.7% for the fourth consecutive year after a peak in 2013 [1]. Considering the significance of commodity prices and their impact on the macroeconomic activity of a country, their accurate prediction is important.

This article examines the performance of both causal and time series approaches of forecasting to determine which approach is better: time series or

causal? To address this research question, neural networks are used for training monthly Illinois soybean freight on board (FOB) prices between 1986:2 and 2013:1, and quarterly HFCS-42 prices between 1995:1 and 2013:3. Later, the respective trained models are evaluated with monthly soybean prices between 2013:2 and 2016:1, and quarterly HFCS prices between 2013:4 and 2015:3.

Time series forecasting and causal forecasting are two main approaches of quantitative forecasting. To examine the relative popularity of the two techniques, a Google Scholar search was done using the two terms: “Time Series Forecasting” and “Causal Forecasting”. For example, during 2010-2016, causal forecasting and time series forecasting counts are 300 and 16,300, respectively. The search results suggest that time series forecasting has been more popular among the research community compared to causal forecasting.

Time series forecasting uses the historical data of a variable to understand the autocorrelation and partial autocorrelation structures of the time series. This structural relationship between present and past data series is later used to develop and estimate the model. Causal forecasting, which is not widely used compared to time series forecasting, consists of two sets of variables: dependent and independent. Independent variables are chosen such that they explain a significant variation in the dependent variable. However, few methods such as vector autoregression seems to fit the definition of both the techniques.

1.1 Time Series and Causal Neural Network Studies

[2] were the first to model time series using neural networks. Subsequently, many studies have used neural networks for forecasting. [3] applied both autoregressive integrated moving average (ARIMA) and feed-forward neural networks to study monthly live cattle and wheat prices for the 1950–1990 period. For both price series, they found that neural networks performed better with lower mean square error as well as capturing maximum turning points compared to the ARIMA model.

[4] applied ARIMA, neural networks, and a hybrid model (using both the ARIMA and neural networks) to three different data sets, wherein the hybrid model performed well compared to the other two models based on the lowest mean square error and mean absolute deviation. Zhang’s hybrid model consists of two steps [4]. First, estimate the ARIMA model to collect residuals of the model. Second, model these residuals in neural networks by incorporating lagged residuals as input variables.

[5] claims that data preprocessing is unnecessary when applying neural networks because they are considered as the universal approximators. However, subsequent studies claim that performing data preprocessing steps such as deseasonalizing and detrending the data yield better forecast accuracy [6],[7]. Specifically, [7] found that deseasonalizing and detrending the data before incorporating into the neural networks yield better results compared to raw data. Hence, in this article, three different data preprocessing steps including deseasonalization, detrending, and Z-score normalization are performed.

[8] compared several machine learning techniques using monthly M3 time series competition data to evaluate a time series forecast. They found that data preprocessing has played an important role in the performance of the model.

Most previous studies have focused relatively more on time series forecasting compared to causal forecasting with neural networks. But, little research has focused on estimation and comparison of model performance, especially in the context of causal versus time series forecasting approaches and hence this study contributes to the literature. This comparison between causal and time series approaches is important to understand the type of approach to be used with specific types of data to make better and informed decisions. In this article, causal and time series approaches are used to predict both soybean prices and HFCS prices. Later, the performance of both causal and time series neural networks are compared through various forecast accuracy measures. Finally, turning point test results are discussed comparing causal and time series forecasting approaches.

2 Model Specification

2.1 Artificial Neural Networks

The artificial neural network contains several interconnected neurons (nodes) that exchange and process information through so-called weights. Weights are analogous to parameters, which are obtained during the process of training the network. A typical neural network consists of an input layer, hidden layer, and output layer. Each layer consists of nodes (equivalent to neurons). For example, the number of nodes in the input layer represent the number of input variables and the number of nodes in the output layer represent the number of output variables.

The hidden layer processes the information in the form of a signal received from the input layer. Signals are first processed by an *integration function* responsible for combining all incoming signals and second by an *activation function* responsible for transforming the output of the neuron. The weights or parameters in the network are assigned using a learning algorithm that minimizes an error function such as the mean square error. The inputs i into hidden neuron j are linearly combined to obtain the integration function

$$Z_j = \beta_{0j} + \sum_{i=1}^k \beta_{ij} X_i \quad (1)$$

where i , j , X_i , β_{0j} , and β_{ij} are the number of nodes in the input layer, number of nodes in the hidden layer, input variable i , the weight connecting the bias unit and hidden neuron j , and the weight connecting input neuron i and hidden neuron j . In the hidden layer, Z_j is then modified using a non-linear function (activation function) such as the sigmoid function $f(Z) = \frac{1}{1+\exp(-Z)}$ to serve as an input for the next layer. Training the network is analogous to moving down an error surface, which takes place by adapting weights as per the learning algorithm rule [3]. The training process is an iterative process, which involves the following steps [9]: 1) compute the predicted or forecast value with given inputs and estimated weights, 2) calculate the error sum of the squares, and 3)

all weights are updated given the rule of the learning algorithm. The iterative process stops when the pre-specified criterion of error sum of squares is fulfilled if all absolute partial derivatives of the error function w.r.t weights are less than the threshold (i.e., $\frac{\partial E}{\partial \beta} < threshold$). The threshold levels are assigned based on how fast the model convergence occurs. In this estimation algorithm, 0.001 is the threshold level used for predicting HFCS prices and 0.01 is the threshold level used for predicting soybean prices. A resilient back propagation algorithm is used for training the network. All algorithms minimize the error function by adding a learning rate to the weights going in the opposite direction of the gradient.

The most important characteristic of the neural networks is the initialization of the weights. Unlike most traditional models, the initialization of the weights should be non-zero values. If the initial values of the weights are zero, then the final values of all the weights will be the same for all input variables. To prevent this issue, random values drawn from standard normal distribution are assigned as initial weights other than zero. In the output layer, there will be a choice of collecting information linearly or non-linearly. Usually, in the regression context, the most prevalent option is linear. In this article, the output layer consists of one node/variable, which is the explained variable—HFCS-42 spot price or soybean FOB price depending on the dataset used.

This article designed the following steps in the methodology: 1) data-splitting, 2) data preprocessing including deseasonalizing, detrending, normalization, 3) training the NN for both the causal and time series approaches, 4) predict HFCS and soybean prices using estimated model and respective test data, and 5) evaluate the forecasts using various forecast accuracy measures.

Two important decisions need to be made before training the neural networks: 1) the selection of the autoregressive terms of the variable that explains a significant variation in that variable, and 2) the number of hidden nodes to be specified in the hidden layer. [10] claim that the number of hidden nodes should be $\sqrt{m * n}$, where m is the number of input nodes and n is the number of output nodes. However, there is no “hard-and-fast rule” in deciding the number of hidden nodes, which are usually selected arbitrarily depending on the data. The model which gives the lowest out-of-sample MSE or out-of-sample MAD is selected as the best model. A precaution must be taken to not to overfit the model. Overfitting is a scenario where the out-of-sample MSE is very high and in-sample MSE is very low.

2.2 Neural Networks

Causal networks are a representation of a dependent variable as a function of independent variables. This article designed two causal networks: soybean price causal networks and HFCS price causal networks. In a time series model, a variable of interest is a dependent variable, which is specified as a function of its own lagged historical observations.

Soybean Price Network: Soybean is a highly traded agricultural commodity in the United States. Approximately 45% of the total US soybean production is exported (<https://www.mda.state.mn.us/food/business/~media/Files/food/business/economics/exports-soybeans.ashx>). US soybean prices are affected

by many factors including US soybean farm price, crude oil price, US Dollar exchange rate (captures the soybean trade impact). This article used the crude oil price, US dollar index (USDI) value, and one lagged US soybean farm price as the major factors explaining the variation in the soybean FOB price over the 1986–2016 period. During the process of model building, when Soybean exports were used as a proxy for production data to include in the model estimation, the results did not improve compared to the results when it was not included. The correlation results also indicate that the USDI captured most correlation patterns in Soybean FOB prices compared to the Soybean exports. Hence, the soybean model is represented as follows: $P_t = \alpha_0 + \alpha_1 * CrudeOilP_t + \alpha_2 * Soybean1LFP_t + \alpha_3 * USDI_t$ where P_t is the Illinois soybean FOB price (US\$/bushel) at time t , $CrudeOilP_t$ is the crude oil price (US \$/barrel) at time t , $Soybean1LFP_t$ is one lagged US soybean farm price (US \$/bushel) at time t , and $USDI$ is the US dollar index value at time t .

HFCS Price Network

HFCS is one of the major caloric sweeteners used in the United States. Currently, HFCS accounts for about 25–30% of the total caloric sweetener consumption in the United States (<http://www.ers.usda.gov/topics/crops/sugar-sweeteners/data.aspx>). Corn is the major raw material used in the production of HFCS. The quarterly corn price, total quantity of HFCS production, and total fructose exports are used as the major contributing factors that explain the significant variation in the HFCS-42 price. Causal neural networks of the HFCS price model has three nodes in the input layer, four nodes in the hidden layer, and one node in the output layer. Typically, in a regression context, there is only one output variable, except in classification models and/or binary response models. The presence of a hidden layer makes the neural network non-linear. The HFCS model is represented as follows:

$$PH_t = \beta_0 + \beta_1 * CornP_t + \beta_2 * QtyHFCS P_t + \beta_3 * TFructoseExp_t$$

where PH_t is the HFCS-42 spot price (US cents/pound) at time t , $CornP_t$ is the corn price (US \$/bushel), $QtyHFCS P_t$ is total quantity of HFCS produced in the United States at time t (short tons), and $TFructoseExp_t$ is the total US fructose exports at time t (metric tons). Neural networks can be written as [7]

$$y_t = \alpha_0 + \sum_{j=1}^n \alpha_j f \left(\sum_{i=1}^m \beta_{ij} x_{it} + \beta_{0j} \right) + \epsilon_t \quad (2)$$

where m , and n are the number of input nodes, and hidden nodes, respectively. While f is a sigmoid transfer function such as the logistic, $f(x) = \frac{1}{1+exp(-x)}$ or a $tanh$ function. $(\alpha_j, j = 0, 1, \dots, n)$ is a vector of weights or parameters connecting the hidden and output nodes and $(\beta_{ij}, i = 0, 1, \dots, m; j = 1, 2, \dots, n)$ are the weights or parameters connecting the input and hidden nodes. Finally, α_0, β_{0j} are called bias terms, which are equivalent to the intercept (the bias terms always take a value of 1). The time series model is specified as: $Y_t = \sum_{i=1}^p \alpha_i y_{t-i} + \epsilon_t$ where Y_t is the dependent variable (also the interested variable for forecast), y_{t-i} is the lagged y_t variables, ϵ_t is the residual at time t , and $i = 1, \dots, p$ is the number of lags.

Two features need to be determined in the time series model before incorporating into the neural networks: 1) the stationarity of the raw data, 2) the appropriate number of lags in the model. There are four different criteria to determine the appropriate number of lags to be used for model estimation. They are based on 1) the statistical significance test, 2) the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), 3) adding lags of the explained variable as the independent variable until the elimination of the serial correlation in the residuals, and 4) partial autocorrelation functions [11]. This article primarily used the third and fourth criteria to determine the appropriate number of lags to be included in the neural networks. Specifically, a residual analysis is performed to check for any significant information left out in the residuals.

One of the concerns of neural networks is that it is susceptible to a local minimum trap that can be mitigated by training the model multiple times while simultaneously experimenting with different hidden nodes with different starting values. The best model was selected based on the lowest test MSE or the lowest MAD.

This article employs the following forecast accuracy measures in evaluating both models: 1) In-sample mean square error: $IMSE = \frac{1}{T} \sum_{t=1}^T (Y_t - Y_t^*)^2$, 2) out-of-sample mean square error: $OMSE = \frac{1}{s} \sum_{t=1}^s (Y_t - Y_t^*)^2$, 3) In-sample mean absolute deviation: $IMAD = \frac{1}{T} \sum_{t=1}^T |Y_t - Y_t^*|$, and 4) Out-of-sample mean absolute deviation: $OMAD = \frac{1}{s} \sum_{t=1}^s |Y_t - Y_t^*|$. where Y_t is the actual value at time t , Y_t^* is the forecast value at time t , and T in IMSE and IMAD represents a total of 75 in-sample observations and 324 in-sample observations in the HFCS price model and the soybean price model, respectively. Also, s in OMSE and OMAB represents a total of 8 out-of-sample observations and 36 out-of-sample observations in the HFCS price model and the soybean price model, respectively.

2.3 Pesaran and Timmermann's (2009) Test for Turning Point Evaluation

Several studies have been proposed to study the directional accuracy of forecasts [12],[13],[14]. However, all the above tests are based on the assumption that null distributions are independently and identically distributed (*i.i.d*). A test proposed by [15](henceforth, PT Test) is used because it addresses the serial correlation between forecasts and actual binary sequences. The PT test evaluates the turning points comparing both the actual and predicted binary sequences of both the soybean price model and HFCS price model. Specifically, the turning points are evaluated separately for the training data, test data and whole (all sample) data.

The PT test tests the dependence of two binary sequences. The binary sequences are obtained after performing a forecast analysis. Specifically, X_t is defined as a binary sequence for the forecast variable that takes a value of 1 if ΔX_t is greater than 1, and zero otherwise. That is, if the change in the value of the forecast in time t and in time $t - 1$ is greater than one, then X_t takes a value of 1, and zero otherwise. Similarly, Y_t is the binary sequence of the actual

series that takes a value of 1 if ΔY_t is greater than 1, and zero otherwise. This means that for a time period, if X_t , and Y_t take the same value—either zero or one—then the forecast and actual time series moves in the same direction. The null hypothesis is that the direction of change in a forecast and the actual binary series is independent [15]. [15] proposed two versions of tests: the trace canonical correlation test and the maximum canonical correlation test. However, in the case of a two-dimensional two-way table (such as in this case), these two tests are similar. Due to space considerations the detailed PT test is not presented here. For more details on PT test see [15], [16], and [17].

3 Data

Data for the quarterly US HFCS–42 spot price (cents/pound), US total quantity of HFCS production ('000 short tons, dry weight), total US fructose exports (metric tons, dry weight basis), and US yellow corn price (\$/bushel) are collected from the United States Department of Agriculture (USDA) sugar and sweetener yearbook tables and USDA feed grain tables for the 1995–2015 period (<http://www.ers.usda.gov/topics/crops/sugar-sweeteners/data.aspx>). The entire data are split into two groups: Training set (1995:Q1 to 2013:Q3) and Test set (2015:Q1 to 2015:Q3). Therefore, the training set consists of 75 observations while the test set consists of 8 observations accounting for nearly 10% of the sample size.

Similarly, data are collected for the monthly Illinois soybean FOB price (\$/bushel), crude oil price (US\$/barrel), a lagged US soybean farm price (\$/bushel), and US dollar index value (no units) from USDA oilseeds crop outlook database, and World Bank for the 1986 (February)– 2016 (January) period. The soybean price model consists of 324 (1986:2 to 2013:1) observations in the training set and 36 (2013:2 to 2016:1) observations in the test set, which also accounts for about 10% of the sample size (<http://www.ers.usda.gov/data-products/oil-crops-yearbook.aspx>). The US dollar index data are collected from <http://www.investing.com/quotes/us-dollar-index-historical-data>.

It is important to note that the two datasets chosen for the analysis are distinct from each other in the sense that the HFCS data are not volatile and there is no continuous movement in the HFCS prices for the period considered. In contrast to HFCS prices, the soybean price data are more volatile and there is a continuous movement in the prices for the period considered. Representing two different types of data for the analysis ensure the generalization of the results to other datasets based on their characteristics.

Data preprocessing was performed to create more uniform data and to avoid computation problems. These data preprocessing steps included deseasonalization, detrending, and Z-score normalization. A versatile and robust method of decomposing time series for deseasonalization and detrending, called as “Seasonal and Trend decomposition using Loess” (STL) was used in the article [18].

Finally, Z-score normalization is performed to both the training and test data sets of the deseasonalized and detrended data by using the following formulas: For training data: $X_{trainnew} = \frac{X_{trainold} - \text{mean}_{train}}{\text{std}_{train}}$, and for test data: $X_{testnew} = \frac{X_{testold} - \text{mean}_{train}}{\text{std}_{train}}$. Where $X_{trainnew}$, $X_{testnew}$, $X_{trainold}$, $X_{testold}$,

$mean_{train}$, and std_{train} are new training data, new test data, old training data, old test data, mean of training data, and standard deviation of training data, respectively, for both soybean prices and HFCS prices.

After convergence of the neural network to the given threshold value, the predicted values (also called in-sample forecasts) are computed. The model parameters and the out-of-sample independent test data are then used to predict the out-of-sample soybean prices and HFCS prices. Both the in-sample and out-of-sample predictions are then rescaled by seasonalizing and trending with the respective time periods. These rescaled predictions and the original prices of both the soybean prices and HFCS prices are finally used to compute their respective predictive accuracy measures such as the MSE and the MAD.

4 Results and Discussion

This section discusses the performance of both the causal and time series models and their ability to forecast soybean and HFCS prices. First, a method is chosen to decide the appropriate number of lags to be used in the model estimation of time series neural networks. Second, the selection of an appropriate number of hidden neurons employed in the neural networks is discussed. Third, forecast accuracy and turning point evaluation results are presented and discussed. Additionally, diagnostic tests are performed on the residuals of both the time series and causal neural networks of HFCS and soybean prices. Finally, the results are discussed in the broader context.

For time series neural networks, the unit root test was performed to ensure the stationarity of the data. This resulted in one differenced soybean prices and two differenced HFCS prices being stationary. Based on the partial autocorrelation functions (PACF), it was determined that five autoregressive terms (1st, 2nd, 3rd, 9th, and 23rd lags) of soybean prices and three autoregressive terms (1st, 2nd, and 3rd lags) of HFCS prices have explanatory power to predict Illinois soybean FOB prices and HFCS-42 spot prices, respectively. Therefore, five lagged explanatory variables were used for predicting soybean prices and three lagged explanatory variables were used for predicting HFCS prices in the respective time series neural networks.

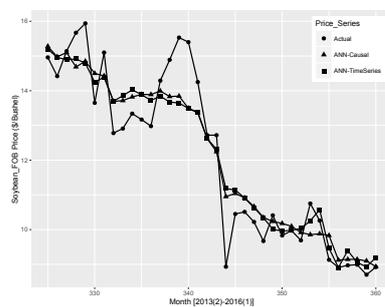
For selecting the number of hidden neurons, the article explored several feed-forward networks with varying hidden neurons and later the best network with the lowest out-of-sample forecast accuracy measures was retained. More specifically, the article used hidden neurons ranging from one to ten consecutively and later chose the appropriate number of hidden neurons based on the best performed network. For soybean neural networks, the $3 \times 2 \times 1$ network was the best for causal while the $5 \times 7 \times 1$ network was the best for time series. Similarly, for HFCS neural networks, the $3 \times 1 \times 1$ network was the best for causal while the $3 \times 7 \times 1$ network was the best for time series. In the network, the first number indicates the number of input (explanatory) variables, second number indicates the number of hidden neurons, and the final number indicates the output (explained) variable.

Table 1 shows the forecast accuracy results of both the causal and time series networks in terms of mean square error (MSE), and mean absolute deviation

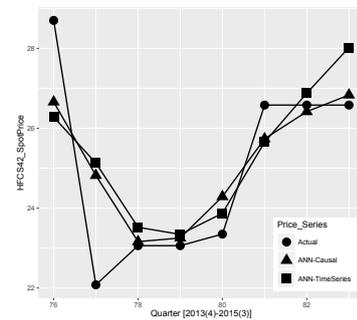
(MAD) for both in-sample and out-of-sample. Results of out-of-sample accuracy measures in table 1 show that causal forecasting performed better compared to time series forecasting in both soybean and HFCS price networks. This is a very important result given the fact that the time series forecasting technique has been more popular compared to causal forecasting. Also, the time series forecasting approach for both soybean and HFCS price networks has consistently performed well in the in-sample, but failed to generalize to the out-of-sample observations. Figure 1 show the out-of-sample fit for soybean and HFCS prices using both causal and time series networks. The out-of-sample fit in Figure 1 shows that the causal approach performed better (in predicting both prices) because the causal predictions are much closer to the actual data points minimizing the error in prediction. For example, in Figure 1 (b), the “ANN-Causal” prediction points are closer to “Actual” data points when compared to “ANN-TimeSeries” predictions.

Table 1: Forecast accuracy results of neural networks

Model Type	In-sample MSE	In-sample MAD	Out-of-sample MSE	Out-of-sample MAD
<u>Soybean Model</u>				
ANN-Causal	0.29	0.36	0.63	0.63
ANN-Time Series	0.23	0.33	0.66	0.64
<u>HFCS Model</u>				
ANN-Causal	0.26	0.38	1.67	0.91
ANN-Time Series	0.02	0.10	2.34	1.18



(a) Soybean price



(b) HFCS price

Fig. 1: Actual vs predicted soybean prices of both causal and time series neural networks (out-of-sample)

Table 2 shows the PT test results for the turning point evaluation [15]. Turning point evaluation results show alternative findings to the forecast accuracy results. The null hypothesis is rejected in all cases at least at the 1% significance level except in the case of the in-sample causal model of soybean prices, and the out-of-sample causal and time series models of HFCS prices. A rejection of the null hypothesis of the PT test implies that a forecast is a useful predictor of the actual change in the variable. Specifically, in the case of the out-of-sample

soybean model, even though both the causal and time series approaches of forecasting are strong predictors of the change in direction of the actual variable, the time series forecasting approach predicted more turning points compared to the causal approach. The turning point evaluation results of HFCS prices show that in-sample and all sample results strongly reject the null hypothesis at the 1% level, while the time series approach captured the change in the direction of the actual variable more accurately compared to the causal approach. As a whole, the turning point evaluation results indicate that the time series approach is better suited when the objective is to maximize capturing turning points in the interested time series.

Table 2: Test results of turning point evaluation

Pesaran and Timmerman (2009) Test Statistic			
Model Type	In-sample	Out-of-sample	All sample
<u>Soybean Model</u>			
ANN-Causal	2.29	6.21***	5.41***
ANN-Time Series	10.74***	7.07***	16.22***
<u>HFCS Model</u>			
ANN-Causal	9.67***	0.44	8.36***
ANN-Time Series	16.30***	1.33	14.74***

Notes: ***p<0.01.

Additionally, two diagnostic tests were performed in order to validate the specification of both the causal and the time series neural networks. The Jarque-Bera (JB) test is used to test for the normality of the residuals, and the Augmented Dickey-Fuller (ADF) test is used to test for the stationarity of the residuals [19],[20]. The null hypothesis of the Jarque-Bera test is the normality of the residuals and it is rejected if the JB test statistic is larger than the critical value of the (χ^2) distribution. For both the soybean and HFCS models, the JB test statistic is less than the 5% critical value so the null hypothesis is not rejected. The null hypothesis of the ADF test is that the residuals are non-stationary meaning that there is a significant amount of information that is left in the residuals. All tests reject the null hypothesis and conclude that the residuals of all the models are stationary at the 5% level. Based on the above diagnostic tests, it is reasonably concluded that both the causal and time series networks of soybean and HFCS prices are correctly specified.

In summary, results of the article indicate that the use forecasting approach depends on the objectives of the research. If the objective is forecast accuracy, then based on the results, it is recommended to employ causal forecasting technique. On the other hand, if the objective is to capture maximum number of turning points, then time series forecasting technique is appropriate.

Finally, a sensitivity analysis was performed in this study. The primary goal of a sensitivity analysis is to evaluate the relative importance of explanatory variables based on their effect on the output variable. Two methods were chosen for sensitivity analysis including Garson's algorithm [21], and Lek's profile method [22],[23]. Garson's algorithm uses the connection weights between each of the in-

put variables to quantify the effect of inputs on the output. A higher connection weight value between an input variable and the output variable indicates that a specific input variable has a higher impact on the output variable. On the other hand, Lek's profile method analyzes the effect of a change in an input variable on an output variable maintaining all other input variables at a certain value. Specifically, it constructs a fictitious matrix pertaining to the range of all input variables in which only one input variable is varying at a time and all other input variables are fixed at certain values such as the minimum, first quartile, median, third quartile, and maximum. Therefore, the effect of each input variable on the output variable generates a plot of five profiles. For additional details of both Garson's algorithm, and Lek's Profile method, refer to [23].

Garson's algorithm results show that the most important variable for determining the soybean price is US Dollar Index, while the least important (of three input variables) is crude oil price. Lek's profile results indicate that the soybean price increases with increase in the crude oil price when all other input variables are constant at the respective quartiles. An increase in the one lagged soybean farm price has positive or increased expectation on the soybean fob price except for minimum. Finally, a higher US Dollar Index has a negative effect on the soybean price for all profile groups except for minimum.

5 Conclusions

This paper demonstrates the forecast performance of two approaches: causal and time series using artificial neural networks, which was neglected in the literature. Both neural networks are used for predicting Illinois soybean FOB prices and HFCS-42 spot prices. Forecast accuracy measures such as mean square error and mean absolute deviation in the out-of-sample reveal that the causal approach performed better in predicting both soybean and HFCS prices compared to time series networks.

Turning point evaluation results suggest that the time series approach captured most of the turning points in both the in-sample, and out-of-sample soybean and HFCS prices. This implies that time series approaches are superior to causal approaches when considering turning point analysis. Even though time series approach seem to be more popular, results of this article indicate that the use of forecasting approach depends on the objective of the researcher. For example, if the researcher is interested in capturing maximum turning points, then the results of the article suggest that the time series approach is more appropriate. On the other hand, if the researcher is interested in the accuracy of the forecast, then the researcher needs to employ the causal forecasting approach.

Sensitivity results of soybean causal model suggest that US Dollar Index is the most important variable that affects the soybean FOB price, while results of HFCS causal model suggest that corn price is significant variable in determining the HFCS-42 price.

References

1. Highlights From the February 2017 Farm Income Forecast, <https://www.ers.usda.gov/topics/farm-economy/farm-sector-income-finances/highlights-from-the-farm-income-forecast/>.

2. Lapedes, Alan, and Robert Farber. Nonlinear signal processing using neural networks: Prediction and system modelling. No. LA-UR-87-2662; CONF-8706130-4. 1987.
3. Kohzadi, Nowrouz, Milton S. Boyd, Bahman Kermanshahi, and Iebeling Kaastra. A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing* 10, no. 2 (1996): 169-181.
4. Zhang, G. Peter. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50 (2003): 159-175.
5. Gorr, Wilpen L. Research prospective on neural network forecasting. *International Journal of Forecasting* 10, no. 1 (1994): 1-4.
6. Nelson, Michael, Tim Hill, William Remus, and Marcus O'Connor. Time series forecasting using neural networks: Should the data be deseasonalized first?. *Journal of forecasting* 18, no. 5 (1999): 359-367.
7. Zhang, G. Peter, and Min Qi. Neural network forecasting for seasonal and trend time series. *European journal of operational research* 160, no. 2 (2005): 501-514.
8. Ahmed, Nesreen K., Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews* 29, no. 5-6 (2010): 594-621.
9. Gnther, Frauke, and Stefan Fritsch. neuralnet: Training of neural networks. *The R journal* 2, no. 1 (2010): 30-38.
10. Kaastra, Iebeling, and Milton Boyd. Designing a neural network for forecasting financial and economic time series. *Neurocomputing* 10, no. 3 (1996): 215-236.
11. Hill, R. Carter, William E. Griffiths, and Guay C. Lim. *Principles of econometrics*. Vol. 5. Hoboken, NJ: Wiley, 2008.
12. Henriksson, Roy D., and Robert C. Merton. On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills. *Journal of business* (1981): 513-533.
13. Cumby, Robert E., and David M. Modest. Testing for market timing ability: a framework for forecast evaluation. *Journal of Financial Economics* 19, no. 1 (1987): 169-189.
14. Pesaran, M. Hashem, and Allan Timmermann. A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics* 10, no. 4 (1992): 461-465.
15. Pesaran, M. H. and Timmermann, A: Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*, 104(485), 325-337 (2009)
16. Chou, Cheng, and Chia-Shang J. Chu. Market timing: Recent development and a new test. *Economics Letters* 111, no. 2 (2011): 105-109.
17. Tsuchiya, Yoichi. Are government and IMF forecasts useful? An application of a new market-timing test. *Economics Letters* 118, no. 1 (2013): 118-120.
18. Cleveland, Robert B., William S. Cleveland, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6, no. 1 (1990): 3.
19. Jarque, Carlos M., and Anil K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters* 6, no. 3 (1980): 255-259.
20. Said, Said E., and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71, no. 3 (1984): 599-607.
21. Garson, David G. Interpreting neural network connection weights. (1991): 47-51.

22. Lek, S., A. Belaud, I. Dimopoulos, J. Lauga, and J. Moreau. Improved estimation, using neural networks, of the food consumption of fish populations. *Oceanographic Literature Review* 9, no. 43 (1996): 929.
23. Gevrey, Muriel, Ioannis Dimopoulos, and Sovan Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling* 160, no. 3 (2003): 249-264.

A value-based evaluation methodology for renewable energy supply prediction

Robert Ulbricht², Bijay Neupane³, Martin Hahmann¹, and Wolfgang Lehner¹

¹ Technische Universität Dresden, Database Technology Group, Germany
martin.hahmann@tu-dresden.de, wolfgang.lehner@tu-dresden.de

² Robotron Datenbank-Software GmbH, Dresden, Germany

robert.ulbricht@robotron.de

³ Aalborg University, Denmark

bn21@cs.aau.dk

Abstract. With the ongoing expansion of renewable energy supply, developing and comparing precise forecasting methods becomes important. In this paper, an evaluation metric is investigated which allows the integration of multiple accuracy criteria into one consistent performance ranking and returns information about the economic impact of a forecast. The practical applicability of the approach is demonstrated using solar energy time series observed in different real-world scenarios.

Keywords: Renewable energy forecasting, performance evaluation, ranking score, business context

1 Introduction

The capacity of renewable energy sources like solar panels and wind mills is constantly increasing world-wide. Simultaneously, many countries aim at establishing liberalized electricity markets in order to create competition between the former monopolistic organizations. As a consequence, the maintenance of the electric balance between power demand and supply is challenged (1) technically by the fluctuating character of the renewable sources and (2) economically by the need for a seamless integration of all market participants. Accordingly, a lot of research was dedicated in the past to the development of precise time-series forecasting models for renewable energy supply. However, due to the lack of a industry-wide accepted and standardized evaluation protocol for forecast quality, decisions are primarily based on context-unaware statistical error measures. As such domain-neutral evaluation criteria do not consider the varying economic impact of over- and underestimations at a certain moment, this can result in misleading decisions. Furthermore, by the use of more than one error criterion the obtained ranking for the competing methods can be inconsistent.

The purpose of this work is the introduction of a value-based performance evaluation methodology for renewable energy forecasting methods. Firstly, a time-dependent context component is introduced by the use of electricity spot

market prices to determine the economic benefit obtained from the forecast results in a modeled market environment. Secondly, we propose different forms of an evaluation criterion which combines multiple uni-dimensional accuracy measures in order to solve possible ranking inconsistencies. Finally, by bringing them together we create an integrated context-aware and multi-dimensional approach with the abilities of reflecting the impact of a decision and flexible adaptation to the underlying scenario. The paper contains 5 sections, with the first being this introduction. In Section 2 we discuss state-of-the-art forecasting performance measures used in the energy domain. Subsequently, we provide a basic description of the relevant business environment defined by the core electricity market rules in Section 3 before we introduce our novel approach of context-dependent forecast benefit determination in Section 4. After that, in Section 5 the practical applicability is demonstrated on three real-world use cases and finally, we conclude and outline future research on this topic in Section 6.

2 Accuracy evaluation in energy forecasting

Despite of a wider range of available forecast evaluation criteria like a model's robustness or technical performance, reducing quality determination to the accuracy dimension of a forecast is a common practice in the forecasting community. However, also the selection of appropriate statistical metrics to measure the forecast accuracy is a topic frequently addressed in literature (compare e.g. [2], [3] or [4]). For the renewable energy domain, the foundations of a standardized performance evaluation protocol were defined by Madsen et al. [6] more than a decade ago. As a minimum set of measures, they propose the use of normalized *Mean Bias Error* (MBE), *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE) and the usage of improvement factors for accuracy comparison between concurring methods and against naïve predictors.

Error Term	Definition
Mean Absolute Error	$MAE = \frac{1}{n} \sum_{t=1}^n y_t - y'_t $
Mean Bias Error	$MBE = \frac{1}{n} \sum_{t=1}^n (y_t - y'_t)$
Mean Square Error	$MSE = \frac{1}{n} \sum_{t=1}^n (y_t - y'_t)^2$
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - y'_t)^2}$
Mean Absolute Percentage Error	$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left \frac{y_t - y'_t}{y_t} \right $
Symmetric Mean Absolute Percentage Error	$SMAPE = \frac{100\%}{n} \sum_{t=1}^n \frac{ y_t - y'_t }{ y'_t + y_t }$

Table 1. Common statistical error criteria used for measuring forecast accuracy.

When deciding on a specific error criterion, its characteristics have to be taken into account: The simple MAE indicates the magnitude of the average error,

but focuses on the mean which leads to an underrating of high, but infrequent errors. This is corrected by the *Mean Square Error* (MSE), as squaring the error before the mean is calculated puts a higher penalty on large errors. The same information is provided by the RMSE with the exception that the square root brings the result's unit back to the original value. In contrast, the MBE describes the direction of the error bias. It's value is related to the magnitude of the value under investigation. According to the definition in Table 1, a negative MBE occurs when predictions are higher than observations, indicating a systematic over-prediction by the model. As none of these criteria provide information on the relative size of the error, forecasts of different time series can not be directly compared. This is addressed by the *Mean Absolute Percentage Error* (MAPE), which is one of the most popular measures and returns the MAE in percentage terms. However, one of the known shortcomings of MAPE is that it is biased as it will systematically select a method whose forecasts are too low. The reason is that for under-predictions the MAPE cannot exceed 100%, but there is no upper limit for over-predictions. Furthermore, it is not defined for observed zero values. To deal with such limitations the *Symmetrical Mean Absolute Percentage Error* (SMAPE) was proposed [7]. The SMAPE has a lower and an upper bound and can handle zero observations as long as predictions are not zero. Surprisingly, our former research has shown that the SMAPE is rarely found in renewable energy forecasting literature [10].

3 Energy market models

In this section we give a brief introduction to common energy market rules and describe the possible interactions for trading and balancing. We provide details for the three exemplary markets whose most relevant market characteristics are compared in Table 2.

Since the beginning of the process in the 1980s [9], many of the industrialized countries have liberalized their energy markets, thus breaking up with the traditional market roles. Customers can now freely choose their favorite energy supplier and electricity is traded between the market participants in the newly created markets. Similar to other commodities, transactions with a short delivery time are done at the short term spot market and trades that have to be fulfilled further in the future in the long term future market. While a spot market trade is meant to satisfy an urgent need, the motivation of future contracts often is to protect a trading party against the economic risks of unexpected and drastic price movements. In the short-term electricity markets, the purchased power is paid either for one trading day before it is delivered (day-ahead) or, depending on country-specific market rules, for up to x minutes before the delivery at a certain hour (intraday). Prices are fixed through auctions or through continuous trading, although literature [1] suggests that they do not vary that much between day-ahead and intraday contracts.

Once intraday trading is closed, any further deviation in the portfolio is balanced by trading in the regulating power market. The regulating power market is activated shortly before the time of the actual delivery and when the market is anticipated to have any imbalance in supply or demand. The regulating power could be activated for any duration of time. Regulating power can be either up or down as a consequence of the following situations: If the supply is less than the demand, the supplier's associated balancing responsible party (BRP) has to buy up-regulating power - at up-regulating power price - in order to maintain the energy balance in the market. The required amount of up-regulating power is fulfilled by other energy suppliers or by decreasing the demand by an amount equivalent to the difference. On the other hand, if the supply is greater than the demand, the BRP has to sell down-regulating power - at down-regulating power price - to maintain the energy balance in the market. The down-regulating power is sold in the reserve energy market, or the demand is increased by an amount equivalent to the difference. Furthermore, negative wholesale prices can be permitted or penalties can be applied to those who cause such deviations.

	Australia	Denmark	Germany
Intraday trading	No	Continuous, closure 60min before t_0	Continuous, closure 5min before t_0
Pricing restrictions	$-1,000 \leq P_{SPOT} \leq 330 \text{ AU\$/MWh}$	$-500 \leq P_{SPOT} \leq 3,000 \text{ €/MWh}$	$-500 \leq P_{SPOT} \leq 3,000 \text{ €/MWh}$
Deviation penalties	Cost of regulation power	Cost of regulation power	None for P_{FIT} ; Cost of regulation power for P_T option

Table 2. Comparison of selected characteristics for the electricity spot markets of Australia, Denmark and Germany.

(Western) Australia. The Wholesale Electricity Market (WEM) for the South West Interconnected System of Western Australia operates independently from the Australian National Energy Market. Most of the energy trading in the WEM is done directly via bilateral contracts, so in the day-ahead spot market (STEM) only the positions not already covered by such contracts are traded. Positions can be traded until 9:50 AM of each trading day and prices are settled through auctions. After the STEM trades are settled, all deviations from contract positions are exposed to the regulation price. The regulation price is determined according to the minimum and maximum STEM prices for the trading period. Balancing costs are funded by the customers based on their monthly demand [8].

Denmark. The Danish energy market is an integral part of the Nordic energy market, and trading takes place through Nord Pool. The spot market closes at 12 AM, where the market participants submit their bid for power that will be delivered or purchased on each trading interval of the following trading day. Each trading interval represents an hourly period. Once Nord Pool calculates and in-

forms the market prices to the participants, the trade is settled. Thereafter, any deviation in the commitment to the spot market is handled by bilateral trading or participating in the intraday market which opens at 2 PM and closes 60 min before delivery start. Up- and down regulation power prices are distinct and the cost of regulation is assigned to those participants who are responsible for the imbalance.

Germany. Trading in the German energy market takes place at the European Power Exchange (EEX), where bids for the spot market have to be submitted until 12 PM. Intraday trading starts at 3 PM for the following day, closing 5 min before delivery. Renewable generators can choose between fixed feed-in tariffs P_{FIT} or directly selling their energy to the market and receiving a premium tariff P_T on top of the market price. In case of the latter they are charged for imbalances, otherwise the consumer has to pay the balancing services. In the regulation energy market, only one price is determined for both up- and down-regulating power.

The chosen examples show that for every market design, there are different conditions that can affect the way forecasts are created and evaluated. This also applies to the specific requirements in the forecasting process and to the motivation of the market participants for providing accurate results.

4 Value-based forecasting performance

In this section, first we describe our approach of value-based forecasting performance measurement before we define a metric used for selecting forecasting methods based on ranking scores.

4.1 Forecast Benefit Determination

Something all the error criteria discussed in Section 2 have in common is that none of them is context-dependent. To provide information about the domain-specific economic impact of forecast accuracy when deciding for a specific method, a tailor-made criterion has to be used which allows for modeling the relevant business environment as shown by [5]. In the case of renewable energy suppliers, the economic benefit of a forecast is determined by applying the corresponding electricity market-rules and -prices to the numerical results. This returns a scalar product of two time series which is time-dependent, so whenever there are differences between spot- and regulation-prices the over- and underestimations will be fined differently thus leading to a higher diffusion of the original forecast accuracy. Subsequently we propose two criteria to measure the benefit for a forecast.

Forecast Value. The *Forecast Value* $FCV(F_n)$ aims at giving information about the absolute monetary return from the day-ahead spot market for a chosen forecasting model F_n (compare Equation 1): The predicted amount of energy y' is always sold at the corresponding electricity spot market price P_{SPOT} . When the actual delivered amount of energy y is higher than anticipated, the surplus energy is bought at 100% by the TSO at the down-regulation price P_{R_DW} . Accordingly, for underproduction the TSO will sell the missing amount to the energy supplier at the up-regulation price P_{R_UP} . In both cases, the expected trading profit can be further increased when premium tariffs P_T are paid on top; or reduced by the deviation penalty D_P .

$$FCV(F_n) = \begin{cases} y' * P_{SPOT} + (y - y') * (P_{R_DW} + P_T) - D_P & \text{if } y > y' \\ y' * P_{SPOT} + (y' - y) * (P_{R_UP} + P_T) - D_P & \text{others} \end{cases} \quad (1)$$

Energy producers participating in the spot market will be interested in maximizing their benefit in terms of FCV . This means that depending on their bidding strategy, the most accurate forecast will not necessarily result in the highest FCV . For example when $P_{SPOT} < P_{R_DW}$, intentional underestimating the scheduled output and selling the surplus energy at the regulation market is more attractive as long as the penalty costs caused by the deviation are lower than the net trading result. However, the FCV can be negative if that bidding strategy fails. An exception are producers that receive fixed feed-in tariffs P_{FIT} and do not have to pay deviation costs, they do not rely on forecast quality because the benefit model for their FCV simplifies to:

$$FCV(F_n) = y * P_{FIT} \quad (2)$$

Forecast Loss. Similar to the FCV introduced above, the *Forecast Loss* $FCL(F_n)$ includes market information but determines the monetary loss for a forecasting model compared to a perfectly fitted result. The FCL is defined as the scalar product of the absolute energy deviation and the difference between spot- and regulation energy prices according to Equation 3:

$$FCL(F_n) = \begin{cases} |(y - y')| * |(P_{SPOT} + P_T - P_{R_UP})| + D_P & \text{if } y > y' \\ |(y - y')| * |(P_{SPOT} + P_T - P_{R_DW})| + D_P & \text{others} \end{cases} \quad (3)$$

Hence, the optimal FCL is 0 which means that there is either no error in the forecast or no price difference between spot- and regulation market at the moment of energy delivery. Unlike the FCV , the FCL is insensitive to the bidding strategy of the market participant, although the significance of the time component for forecast accuracy is reflected here as well.

4.2 Multi-dimensional Ranking Scores

Using more than one evaluation criterion when comparing the accuracy of competing forecasting methods is common practice in the energy forecasting domain

[10] and in such cases a distinct ranking can be obtained for each criterion. This leads to several inconsistent rankings for different criteria and finally leaves the decision about the optimal method to choose to the user. A multi-dimensional ranking score (e.g. [11]) provides a unique ranking for multiple criteria. In this subsection we introduce the different versions of ranking scores used in this paper.

Absolute Ranking Score. Let n_F be the number of evaluated forecasting methods F and m_E be the number of statistical error measures $E_i(F_n)$ with $1 \leq i \leq m_E$ calculated for each forecast output. Now, for each $E_i(F_n)$ the forecasting methods are ranked starting with $S_i(F_n) = 1$ for the lowest error value $\min(E_i)$ to $S_i(F_n) = n_F$ for the highest value $\max(E_i)$. The score $S_i(F_n)$ is the rank of F_n for its error measure E_i . The *Absolute Ranking Score* $RS(F_n)$ can then be described as the sum of $S_i(F_n)$ for all respective E_i as shown in equation 4:

$$RS(F_n) = \sum_{i=1}^{m_E} S_i(F_n) \quad \text{with} \quad 1 \leq S_i(F_n) \leq n_F \quad (4)$$

For example, in a setting using 5 different forecasting methods and 3 error measures, the best score that can be obtained is 3 (=first position for each error category), while the lowest score would be 15.

Normalized Ranking Score. As the RS only considers the absolute rank of a forecast in the result list, its scale is quite coarse. The RS provides no information about the magnitude of the distance between one position and the next, so two methods having very close error values would have the same score as two with a much wider spread. The *Normalized Ranking Score* $NRS(F_n)$ corrects this shortcoming. When determining $S_i(F_n)$, the error values of each category are normalized so that $\min(E_i) = 0$ and $\max(E_i) = 1$. This means that for the $NRS(F_n)$, the optimal value is 0 (lowest value for all E_i) while the worst result is n_F . This way any discrimination is eliminated and furthermore, the probability of having equal ranking scores for methods that simply alternate their absolute result positions is reduced.

$$NRS(F_n) = \sum_{i=1}^{m_E} S_i(F_n) \quad \text{with} \quad 0 \leq S_i(F_n) \leq 1 \quad (5)$$

Weighted Ranking Scores. When choosing number and type of error measures to be used for a ranking, there might be a need to over- or underweight a specific E_m in the final score. This requirement leads to the *Weighted Ranking Score* $WRS(F_n)$ which applies the weighting factor λ_i with \sum to the absolute score $S_i(F_n)$ for each E_i as described in Equation 6. Alternatively, λ_i can be applied to the normalized score as well and is denoted as *Weighted Normalized Ranking Score* $WNRS(F_n)$.

$$WRS(F_n) = \sum_{i=1}^{m_E} \lambda_i S_i(F_n) \quad \text{with} \quad \sum_{i=1}^{m_E} \lambda_i = 1 \quad (6)$$

Using static or variable weights allows for better adaption of the evaluation metric to specific characteristics of the underlying scenario, thus increasing its' overall flexibility. For example, continuously emphasizing the *MBE* criteria would lead to better scores for forecasts that do not have a high systematic error, although they might have strong absolute deviations in both directions during the whole evaluation period.

The usage of weighting factors introduces the problem of how to derive the optimal $\lambda_{i,t+1}$ for a given use case in advance. To determine $\lambda_{i,t}$ we use the current rankings for the individual error criteria $R_{i,t}(E_i)$ as the input vectors and one of the corresponding forecast benefit rankings $R_t(B)$ with $B \in \{FCL, FCV\}$ as the target values for the optimization function. For example, if a method F_n had the highest *FCV* in t , all $\lambda_{i,t}$ are set to values so that the weighted score ranking $R_t(E)$ with $E \in \{WRS, WNRS\}$ shall also return the first position for F_n . However, this will not always be solvable by the optimizer so the quality of the obtained results has to be verified in order to avoid misleading results for λ_i . This is done by calculating the accuracy A_j obtained from the weighted ranking according to Equation 7. The level j of A determines how many positions are relevant for the ranking accuracy, so e.g. setting $j = 5$ means that only the first 5 methods of the ranking are considered while all lower positions will be ignored.

$$a(E, B) = \begin{cases} 1 & \text{if } R(E) = R(B) \\ 0 & \text{others} \end{cases} \quad \text{and} \quad A_j = \frac{1}{j} \sum_{n=1}^j a_n \quad (7)$$

For static weights, setting the weighting factors once manually based upon expert knowledge might be suitable. However, as soon as the forecast environment changes in the long term (e.g. by a higher seasonal spread of electricity prices), regular adjustments are necessary to better reflect the new situation. Therefore, we use diurnal persistence thus assuming that the environment of the actual trading interval t is similar or equal to the day before so that for short terms we anticipate $\lambda_{i,t} = \lambda_{i,t-1}$. This decision depends on if $A_{j,t-1}$ is considered as satisfying, e.g. setting a threshold of $A_5 \geq 0.80$ signifies that at least 4 out of the top 5 methods in the list have to be ranked with correct positions, otherwise $\lambda_{i,t-2}$ is used and so on.

5 Evaluation

In this section we evaluate the forecast performance criteria introduced in the Sections 4.1 and 4.2. We describe the characteristics of the used data sets and the methodology of our experiments before we present and discuss the obtained results.

5.1 Methodology

For each of the markets presented in Section 3 we use a data set containing time series with the aggregated measured output from different local solar energy installations. In the Australian scenario, a single roof-top panel installed in Perth is used. For Denmark, the data is taken from a ground-based commercial solar farm close to Aalborg and for the third use case, we have an aggregated measurement of all mixed-type solar panels belonging to a local distribution network covering an area of 46km² in central Germany. Up to 4 external influences like e.g. global irradiation and air temperature are taken from an on-site or nearby weather station, and the corresponding prices from the day-ahead spot- and the regulation-market. Note that all time series are measured data, so any deviations possibly caused by unreliable weather predictions are excluded from the results. Furthermore, we extract additional numerical features like the Hour-of-the-day, the Day-of-the-year, and the Clear-sky value from the raw data in order to improve the expected forecast quality for those methods using them. All time series are equidistant without missing values and have a maximum resolution of 60min; at least two years of historical data are provided in all scenarios.

Symbol	Method
ARIMA	Auto-Regressive Integrated Moving Average.
ARIMAX	ARIMA with external regressors.
ETS	Exponential smoothing state space model.
GBM	Generalized Boosted Regression Model.
HW	Holt-Winters seasonal exponential smoothing.
KNN	Regression model using weighted k-Nearest Neighbors.
MARS	Multivariate Adaptive Regression Splines.
MLP	Multi-Layer Perceptron. A fully connected feedforward network.
MLR	Multiple Linear Regression
NAIVE_CS	Naive Clear-Sky. Values are taken from Clear-Sky feature (maximum model).
NAIVE_DP	Naive Diurnal Persistence. Values correspond to last day's observation.
NAIVE_ZE	Naive Zero. All values are zero (minimum model).
NNET	Neural Network with a single-hidden-layer.
RF	Random Forest. Regression based on a forest of trees using random inputs.
SVR	Regression model using Support Vector Machines.

Table 3. Forecasting models used for evaluation

Our objective is to measure the forecast benefit for different forecasting methods and compare the outcome to the ranking scores. Therefore, our experiments are organized as follows: For each use case, we apply a selection of 12 competing state-of-the-art forecasting methods (compare Table 3) to predict the future electricity output from the solar energy installations. We include a naïve persistence model *NAIVE_DP* to provide a benchmark as well as a minimum and maximum model (*NAIVE_ZE*, *NAIVE_CS*), to mark the upper- and lower bounds for the expected values. Initially, the first year of historical data is taken for training, and the second year for evaluating the forecasting models. Then, the forecasts are computed on a daily basis using a sliding window over the training data

according to the submission rules for the individual day-ahead spot markets. For example, if a forecast with hourly resolution has to be submitted at 12 AM before the next trading day, the available training history ends with the 11 AM observation and the forecast horizon is 36 hours ahead, of which the first values that still belong to the actual trading day will be discarded. Trading is simulated as day-ahead only, intra-day corrections are not considered. The errors are computed on the forecast output for each interval. First we use the statistical error measures *MAE*, *MBE*, *MSE*, *RMSE*, *MAPE* and *SMAPE* (compare Table 1) as individual criteria and then a combination of all of them to determine the ranking scores *RS*, *NRS*, *WRS*, and *WNRS*. To measure the obtained forecast benefit, *FCV* and *FCL* are calculated with the market prices P_{SPOT} , P_{R_UP} and P_{R_DW} .

5.2 Result discussion

The results listed in Table 4 show the monetary advantage when permanently choosing the optimal forecasting model. The lower bound is always marked by the Clear-Sky model *NAIVE_CS*, which means that constantly over-estimating the output would lead to the lowest benefit. On the other hand, for Australia and Germany the highest benefit is obtained when extremely under-estimating using *NAIVE_ZE*, so selling energy at the regulation market brings higher revenues than on the spot market. However, this strategy would not have worked for the Danish market. Although here the spread between best and worst result is the highest with 175%, no naïve model is found among the first ranks so the market rules clearly favor accurate forecast. In contrast, the top-ranked methods in terms of *FCL* are *GBM*, *RF*, *MARS*, *MLP* and *NNET*, which all are sophisticated forecasting methods that make use of external information. In a fluctuating environment, they are more likely to produce forecasts of higher accuracy than uni-variate or naïve methods.

#	Australia Method	Result	Denmark Method	Result	Germany Method	Result
1	NAIVE_ZE	656.27 \$	GBM	51,050.61 €	NAIVE_ZE	105,459.87 €
2	ETS	-5.9%	MARS	-2.7%	GBM	-4.2%
3	GBM	-6.9%	MLP	-3.3%	SVR	-4.4%
4	RF	-7.3%	RF	-3.4%	RF	-4.5%
5	MARS	-7.5%	NNET	-4.3%	MARS	-5.3%
...
15	NAIVE_CS	-64.7%	NAIVE_CS	-175.4%	NAIVE_CS	-110.8%
1	GBM	5.00 \$	GBM	1,053.29 €	RF	20,467.61 €
2	RF	+35.3%	RF	+2.6%	GBM	+0.8%
3	NNET	+39.4%	MARS	+12.8%	NNET	+4.5%
4	MLP	+42.4%	NNET	+13.5%	MARS	+5.5%
5	MARS	+49.3%	MLP	+33.4%	MLP	+9.2%
...
15	NAIVE_ZE	+3491.4%	NAIVE_CS	+908.5%	NAIVE_ZE	+512.2%

Table 4. Accumulated *FCV* in the upper- and *FCL* in the lower part. The first row of each block shows the absolute value, subsequently the percental deterioration is listed.

Figure 1 compares the outcome for the error criteria in terms of *FCV* on the left, and *FCL* on the right side. The method selection decision is reconsidered

in a daily interval, so when observation values are available at the end of each trading day, the optimal method from the last period is used again to predict the forthcoming day. It can be seen that for the *FCL*, the ranking scores (red) outperform most of the standard error criteria (light blue). Furthermore, the varying impact of the standard errors can be observed, e.g. basing on the *SMAPE* leads to a higher (avg. +30.6%) benefit for our examples than the *MAPE*. Using the *WNRS* with flexible weights works fine for Australia, for Denmark and Germany *NRS* and *RS* would be preferable. In contrast, for the *FCV* using the standard errors is more convenient than ranking scores with the exception of Australia. Average differences between the different options are much smaller in all scenarios than for the *FCL*.

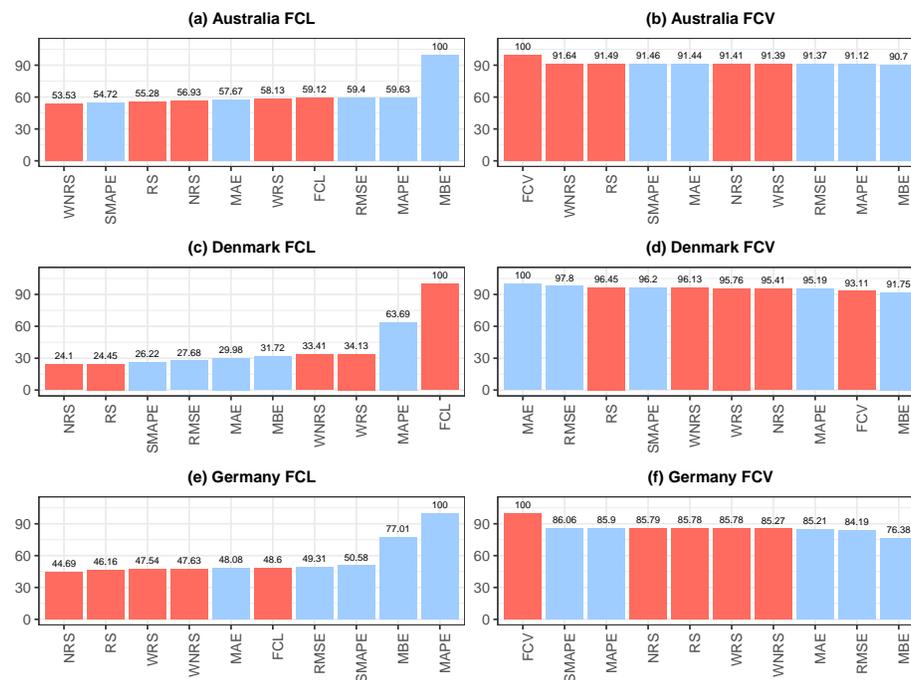


Fig. 1. Impact of selected error measure on total FCL in the left - and FCV in the right column. Standard error measures are displayed in light blue, while the red bars display the specific error criteria introduced in Section 4. Results are normalized and organized in the way that the left bar in each diagram represents the best result obtained.

6 Conclusions

Our findings show that *FCL* and *FCV* are justified context-aware output measures for forecast performance evaluation as both of them illustrate the economic benefit obtained from a specific method. However, they should not be used for

the same purpose. Basing method selection on multi-dimensional ranking scores leads to better forecasting results in terms of a minimized FCL than for most of the uni-dimensional criteria but the use of weights does not always outperform unweighted scores. Otherwise, the maximization of the FCV is preferably obtained by using the very same FCV for markets with one regulation energy price where over- and underestimations are equally fined. In contrast, for varying regulation prices MAE and $RMSE$ give better results. In fact, the definition of an appropriate evaluation metric strongly depends on the underlying scenario's business context information. Our future work will address refinements of the ranking accuracy measurement and the method selection strategy.

Acknowledgment

The work presented in this paper was funded by the European Regional Development Fund (EFRE) and the Free State of Saxony under the grant agreement number 100269304 and co-financed by Robotron Datenbank-Software GmbH.

References

1. N. Aparicio, I. MacGill, J. Rivier Abbad, and H. Beltran. Comparison of Wind Energy Support Policy and Electricity Market Design in Europe, the United States, and Australia. *IEEE Transactions on Sustainable Energy*, 3(4):809–818, oct 2012.
2. J. S. Armstrong. Evaluating Forecasting Methods. In J. S. Armstrong, editor, *Principles of Forecasting*, volume 30 of *International Series in Operations Research & Management Science*, pages 443–472. Springer US, 2001.
3. Z. Chen and Y. Yang. Assessing forecast accuracy measures. Technical report, Iowa State University, Department of Statistics & Statistical Laboratory, 2004.
4. R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
5. J. Luoma, P. Mathiesen, and J. Kleissl. Forecast value considering energy pricing in California. *Applied Energy*, 125:230–237, jul 2014.
6. H. Madsen, G. Kariniotakis, H. Nielsen, T. Nielsen, and P. Pinson. A Protocol for Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models. Anemos project, European Commission, 2004.
7. S. Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, 1993.
8. I. M. Operator. Wholesale Electricity Market Design Summary. Technical report, Independent Market Operator, 2012.
9. R. Raineri. Chile: Where it all started. In F. P. Sioshansi and W. Pfaffenberger, editors, *Electricity Market Reform: An International Perspective*, pages 77–108. Elsevier, 2006.
10. R. Ulbricht, A. Thoß, H. Donker, G. Gräfe, and W. Lehner. Dealing with uncertainty: An empirical study on the relevance of renewable energy forecasting methods. In *Lecture Notes in Computer Science*, volume 10097 LNAI, pages 54–66. Springer International Publishing, 2017.
11. B. Xu and J. Ouenniche. A multidimensional framework for performance evaluation of forecasting models: context-dependent DEA. *Applied Financial Economics*, 21(24):1873–1890, dec 2011.

ANALYSIS OF TERRESTRIAL WATER STORAGE VARIATIONS ON THE TERRAIN OF VISTULA AND ODRA BASINS IN POLAND

Zofia Rzepecka, University of Warmia and Mazury in Olsztyn, Poland

E-mail: zofia.rzepecka@uwm.edu.pl

Abstract. Nature and climate has been affected by increasing human activity nowadays. It is particularly observed from the middle of the 20th century. In this situation, various natural phenomena occurring in different areas should be monitored to observe if and how they are influenced by the changing climate. Water is especially important component of the natural environment. In this paper, time series of the total terrestrial water storage values, covering the period of 2002 to 2016, obtained on the basis of the GRACE mission satellite measurements, were examined. Results from different computing centers and obtained using different methods were compared. An attempt was made to determine the linear trend of the TWS data referring to the Vistula and Odra basins in Poland. Proper determination of the linear trend would provide the basis for determining whether TWS has been changing in recent years and whether these changes lead to an increase or decrease of the total water content in the studied area. Trend values were computed using the linear regression least squares method and then tested applying the Mann-Kendall trend tests. The analyzes have shown that the annual changes of TWS values have amplitude of the order of 4 cm, while the monotonic trends computed are so small that the total terrestrial water storage can be considered stable during the period studied.

Keywords: Terrestrial water storage TWS, GRACE, linear regression, Mann-Kendall trend test

1 Introduction

Terrestrial water storage (TWS) is defined as vertically integrated water of all forms above and below the Earth's Surface, e.g., surface water, soil moisture, groundwater, snow and ice [4]. Nowadays, observations of TWS can be acquired from the Gravity Recovery and Climate Experiment (GRACE), which observes time variations of the Earth gravity field potential. After taking into account atmospheric and oceanic effects, the remaining signal refers to TWS changes, in time scales of month. GRACE observations represent average values, both in time and space. The standard GRACE product is sets of spherical harmonic coefficients describing changes in the Earth's gravity field. They can be used for the surface mass variation computations. After filtration performed to reduce measurement errors, these data can be

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

converted to geographical coordinates and given in the form of discrete values assigned to the selected grid of geographical coordinates, usually in resolution of 1^0 by 1^0 . Two filters are used in this process. The first one aims at removing systematic errors connected with the existence of correlations between individual spherical harmonic coefficients. This stage of processing is called destriping [17]. The task of the next filter is data smoothing, and most often used for this purpose is the Gauss Filter, with the radius of 300 km. An additional feature of the GRACE data comes from cutting the solution to the selected degree and order of the potential series development. Usually the truncation at $l \leq 60$ is applied, which leads to the resolution of 300 km x 300 km of the resulting solution. Due to filtration and truncation the GRACE data presented in the domain of geographical coordinates, are a kind of averaging. This results in an impact of signals from one area to another (and vice versa). The errors that arise then are called the leakage errors [8]. These errors can be determined and eliminated from the solution. One possible way of doing this is to use a global assimilation system describing the state and changes of the land geosphere. The general goal of such systems is to integrate various satellite and terrestrial observations using advanced data modeling and assimilation techniques. One of such systems is GLDAS – Global Land Data Assimilation System [13], which consists of four sub-models – NOAH (National Centers for Environmental Prediction/Oregon State University/Air Force/Hydrologic Research Lab Model), MOSAIC, VIC (Variable Infiltration Capacity Model) and CLM (Common Land Surface model). To calculate the leakage errors the original GLDAS data, taken from one of the given submodels with 1 degree resolution, are converted to the domain of spherical harmonic factors and subjected to the same operations as the GRACE data during processing, that is, to the same filtering and trimming. A new hydrological signal is obtained, different from the original one. On the basis of a comparison of the original and new time series, the scale factors are calculated, which can be used for mitigation of the leakage errors.

An alternative approach to solving for gravity or mass variations in terms of spherical harmonics, which has recently become more and more popular, is to use mass concentration blocks (or “mascons”) as the basis function in the processing of the GRACE data. The most important advantage of using mascons instead of spherical harmonics is that each mascon has a specific known geophysical location, which is not the case with spherical harmonic coefficients. The latter are not connected to any particular localization, at least individually. This feature enables specifying a priori known constraints during the data inversion to internally remove the correlated error in the gravity solution. Therefore, unlike the unconstrained spherical harmonic solutions, the constrained mascon solutions typically do not need to be destriped or smoothed [1].

There are three main computation centers which make the GRACE spherical harmonic solutions data available, these are: GFZ (GeoforschungsZentrum in Potsdam), CSR (Center for Space Research at University of Texas, Austin) and JPL (Jet Propulsion Laboratory). Generally, they apply different computational strategies. The compatibility of independent calculations indicates the correctness of solutions. According to many authors - it is best to use the average of the three solutions [15].

2 Data admitted for analyzes

In this paper, TWS data referring to the area of Poland were analyzed. The data from all the three centers have been downloaded from the JPL server <http://grace.jpl.nasa.gov/>. All the data were previously processed according to the Release-05 GRACE Level-2 (RL05) standard by appropriate Centers. In the frame of post-processing procedures, they were destriped, the Gaussian Filter of 300 km radius was applied and the series were truncated to the maximum degree and order of 60. In all of the data, time mean for the dates 2004.000 to 2009.999 was removed. Leakage scaling factors, copied from the JPL website as ASCII files were applied to all the TWS values. TWS data for 14 years, for epochs from Nov. 2002 to Oct. 2016 were used, which after gaps patching, gave 168 (14x12) quantities in each time series. Missing values were computed using linear interpolation of neighboring entries. All TWS values were referred to the first epoch of the data, i.e. TWS value obtained for November 2002 was subtracted from all the successive values. Therefore, each time series starts from zero.

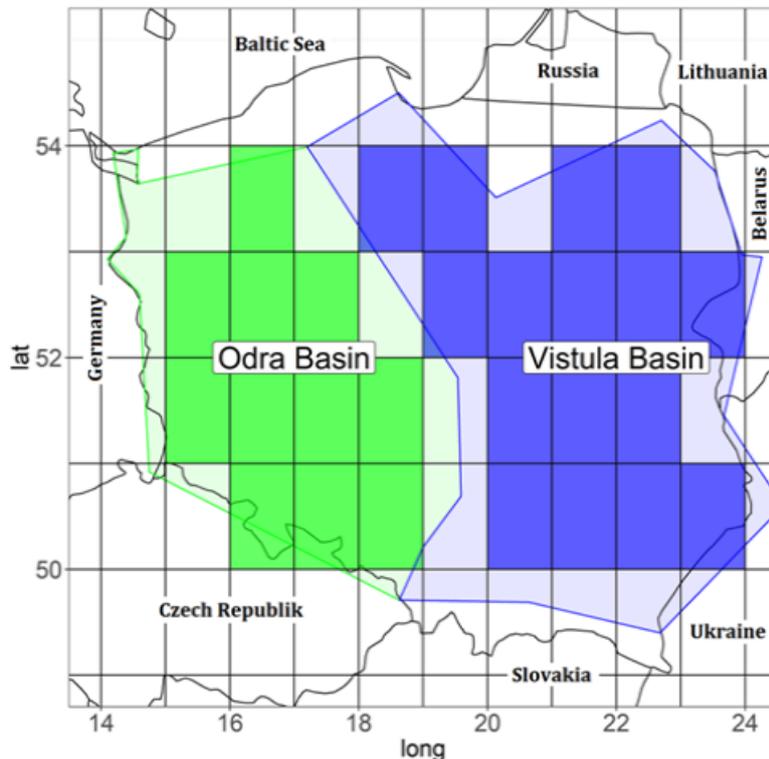


Fig. 1. Vistula and Odra basins: cells used in computations on the background of generalized basin borders

In Poland there are two main rivers: Vistula and Odra. Their basins cover almost whole Poland's area, see Fig. 1. Surface area of the Vistula basin is about 194 500 km² and that of the Odra basin is about 118 900 km². It is seen that the basins are not very big, but still their area is bigger than 300x300 km², thus the GRACE data can be used to analyze mean TWS values on them. All the computations and plots were performed using R statistical open software [11,19]. The grid data were averaged over cells shown in Fig. 1b. Eleven and sixteen cells were admitted respectively for the Odra and Vistula basins for averaging. Thus 8 time series have been obtained: from 3 centers plus mean of the three centers for two basins. The data are plotted in Fig. 2. Additionally, mascon solution TWS, acquired from JPL and GFSC computation centers, were used for comparisons. It can be seen from Fig. 2 that values of all the eight time series are very similar and have a similar course. General seasonal components are clearly seen in the data, the values are contained between minus and plus 15 cm.

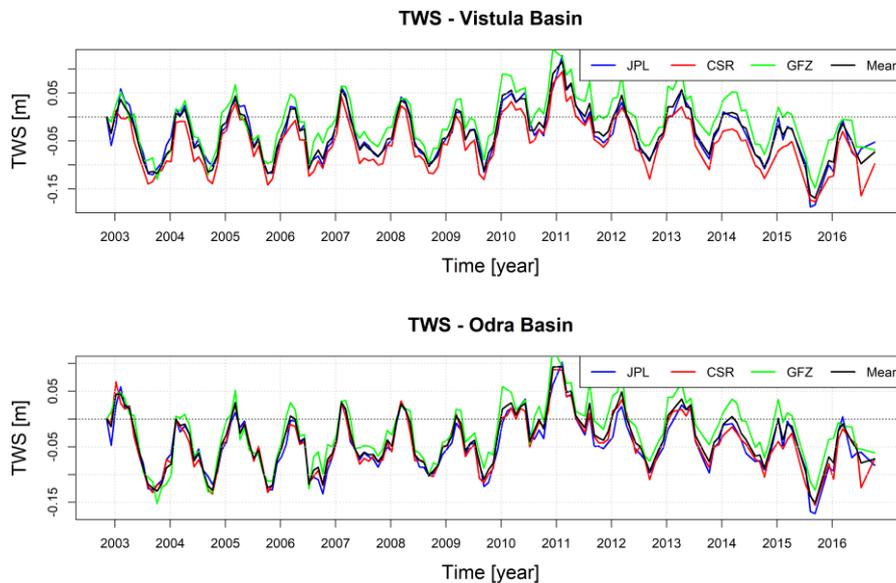


Fig. 2. Time series of TWS values obtained for Vistula and Odra basins

3 Analyses performed

The most important information that was wanted to get from the analysis was whether there were trends in the TWS data series, and if so, whether the total amount of water increases or decreases. Data for all tested time series look very similar, therefore it was expected that the obtained trends will be of the same type. Pearson's coefficients, which were used as a measure of the similarity of time series, are shown in Table 1, while the received trend values and the resulting total change in the amount of water are given in Table 2. The trends were computed using the linear regression

least squares method, implemented in the lm R function [2]. All the data are highly correlated, the lowest value of the Pearson's coefficient amounts to 0.87, and it occurs for the JPL solution for the Vistula basin and GFZ solution for the Odra basin. What more can be seen from the Table 1, is that solutions from a given computation center are more correlated between each other, even if they were taken between different basins (black bold numbers), than solutions obtained for the same basin but from different computation centers (dark-blue bold numbers).

Table 1. Pearson coefficients between time series data (upper triangle) and cross-correlation function values with one-year (12 months) lag (lower triangle); Vis refers to Vistula, Odr is abbreviation of Odra, J,C,G,A mean JPL, CSR,GFZ and average respectively

	J_Vis	C_Vis	G_Vis	A_Vis	J_Odr	C_Odr	G_Odr	A_Odr
J_Vis	1	0.94	0.92	0.97	0.97	0.92	0.87	0.94
C_Vis	0.61	1	0.94	0.98	0.92	0.97	0.89	0.95
G_Vis	0.59	0.59	1	0.97	0.91	0.92	0.96	0.95
A_Vis	0.60	0.60	0.60	1	0.96	0.96	0.93	0.97
J_Odr	0.57	0.56	0.57	0.58	1	0.95	0.91	0.98
C_Odr	0.59	0.59	0.60	0.60	0.53	1	0.93	0.98
G_Odr	0.57	0.56	0.60	0.59	0.53	0.53	1	0.97
A_Odr	0.59	0.58	0.60	0.60	0.54	0.54	0.55	1

Table 2. Trends and variations of water content

	J_Vis	C_Vis	G_Vis	A_Vis	J_Odr	C_Odr	G_Odr	A_Odr
Trend [cm/y]	-0.12	-0.15	0.10	-0.05	0.01	-0.01	0.24	0.08
Δ EWH [cm]	-1.65	-2.03	1.43	-0.75	0.15	-0.17	3.35	1.11
Δ Vol. [km ³]	-3.20	-3.94	2.79	-1.45	0.17	-0.21	3.9	1.32

Since correlation coefficients are symmetrical, they were given only in the upper triangle of the Table, while the lower triangle, so that it will not be empty, was filled with the same coefficient, but computed for time-shifted data, where the lag amounted to 12 months. All the coefficients obtained are greater than 0.5, which proves that the data contain strong seasonal component.

The trends given in Table 2 differ in both size and signs, also other values computed basing on them, of course, reflect this dispersion. Δ EWH is a total change of equivalent water height and Δ Vol. is a total change of water volume, obtained from Δ EWH multiplied by the area of the appropriate basin. In case of the Vistula basin, all trends, besides that of GFZ, are negative, so the outflow of water took place, while in case of the Odra basin, all trend, besides that of CSR, are positive. The trend for the Odra basin, computed on the basis of GFZ data, is the biggest of all and positive, the

total change of the equivalent water height in this case is over 3 cm. This results in almost 4 km³ of water inflow. In Fig. 3 there are presented data with the biggest and the smallest trends – GFZ solution for the Odra basin and CSR solutions for the Vistula basin.

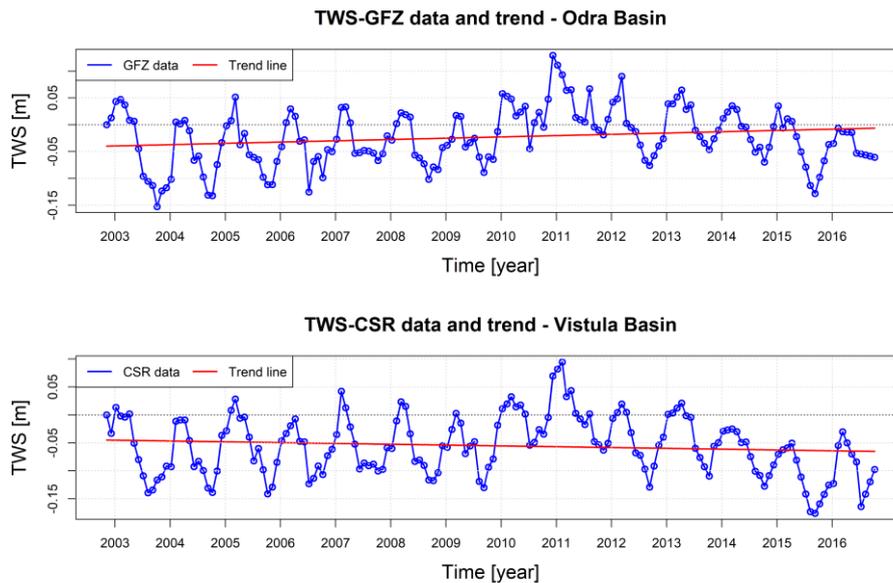


Fig. 3. TWS time series with the biggest and the smallest trends

It should be noted that fitting the annual frequency to the data and then analyzing the fitted values data for the occurrence of a trend, gives practically the same results as those given in Table 2. Example results of such an approach, applied for the data from Fig. 3, are given in Fig. 4.

Very similar looking data produce different linear trends – it gives rise to the suspicion that the results of trend computations are not very significant from statistical point of view. If we want to say something binding about the trends of TWS variations, the data must be subjected to further tests. Seasonal decomposition of the data shows, for the eight cases considered, that the determined trends are not monotonic, for an example of such decomposition, see Fig. 5. This decomposition was computed using the STL function of R [3,12].

Looking at Fig. 5 we can conclude that ranges of data are from -15 to +15 cm, the amplitude of the seasonal component amounts to about 8 cm (from -4 to +4 cm), the remainders of the decomposition also range from -4 cm to +4 cm, and the trend determined is not monotonic, it ranges from -10 cm/year to +2cm/year and it changes its sign every two or three years. Plots of seasonal decompositions performed for other data look very similar.

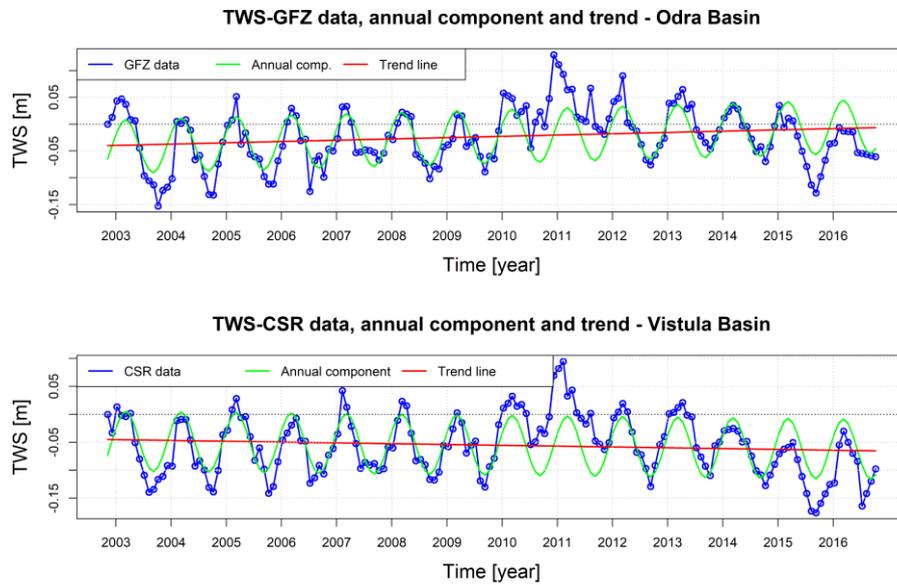


Fig. 4. Trends of deseasonalized data – examples

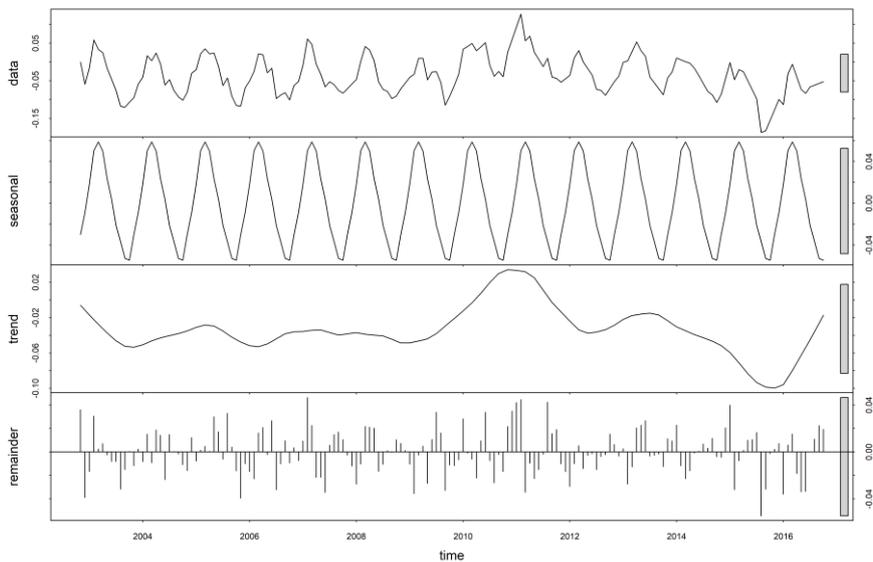


Fig. 5. Seasonal decomposition of the JPL TWS on the Vistula basin area

To further analyze the data in terms of the trends occurring in them, the Mann-Kendall's tests were applied [5,10]. This is a test for monotonic trend in a time series.

It is based on the Kendall rank correlation of the time series values and time [7]. Kendall's rank correlation measures the strength of monotonic association between two vectors, very often they are vector of successive epochs of time and the values occurring in this time series. This test was developed by Mann in 1945, and then improved by Kendall. It computes the Kendall rank correlation and its p-value on a two-sided test of the null hypothesis H_0 . The H_0 in this test claims that the two tested vectors are independent. As it is well known, small p-value (e.g. condition of p-value ≤ 0.05 is often admitted) means strong evidence against the null hypothesis H_0 , while bigger p-values mean weak evidence against H_0 . In case of seasonal data it is suggested to use the seasonal Mann-Kendall test. It was introduced for monthly water quality time series analyses [5,6]. For the purpose of this paper, the package Kendall from the R project was used to perform both the Mann-Kendall and the seasonal Mann-Kendall tests. The results of these tests are given in Table 3.

Table 3. Mann-Kendall (MK) and Seasonal Mann-Kendall (SMK) tests results

	J_Vis	C_Vis	G_Vis	A_Vis	J_Odr	C_Odr	G_Odr	A_Odr
t MK	-0.05	-0.07	0.04	-0.03	0.01	-0.01	0.09	0.04
p-val MK	0.38	0.17	0.40	0.59	0.82	0.91	0.06	0.48
t SMK	-0.02	-0.06	0.08	-0.01	0.06	0.03	0.18	0.11
p-val SMK	0.75	0.33	0.15	0.89	0.27	0.63	0.002	0.06

It can be seen from this Table, that there is only one case, in which we have enough evidence to reject the null hypothesis H_0 which states that there is no trend – these are GRACE TWS data obtained by GFZ for the Odra basin, treated with the seasonal Mann-Kendall test. Thus only for these data we can conclude that the trend exists in TWS values and it is positive (there is more water on average in the Odra basin).

As mentioned in the introduction, nowadays the mascon approach to the GRACE data elaboration gains recognition among scientists. Here, two mascon solutions for Vistula and Odra basins are presented for a comparison with the data that has been analyzed so far. These two solutions are: JPL RL05M v02 Mascons [18,20] and GSFC Global Mascons v02.3b [9]. Both the solutions were acquired from the data portal <http://ccar.colorado.edu/grace/>. The GFSC (NASA's Goddard Space Flight Center) solutions can be taken for given basins, the Vistula basin coincides with the GFSC mascon called „Eastern Poland”, while the Odra basin area coincides with the „Western Poland” mascon. In JPL mascon solutions the mascons are in the form of approximately 3-degree numbered blocks, the mascon No 471 is located on the area of the Vistula basin, and mascon 470 is on the area of the Odra basin. These four solutions are presented in Fig. 6. They were placed here as they were taken from the Colorado Center for Astrodynamics Research of the University of Colorado.

Table 4. Trend and annual amplitudes of GSFC and JPL GRACE mascon solutions

	GSFC_Vis	GSFC_Odr	JPL_Vis	JPL_Odr
Trend [cm/year]	0.22	0.41	-0.10	0.12
Amplitude [cm]	4.61	5.05	4.17	3.91

Trend and annual amplitudes for these two mascon solutions are in Table 4. It can be seen that the GSFC is the most similar to the harmonic solution of GFZ: the trends for both the basins are positive and GSFC mascons give a little bigger values of trends computed. The JPL mascon solutions well coincide with the JPL spherical harmonic solutions: for the Vistula basin the trend is almost the same, while for the Odra basin the trend for the mascon solution is bigger. Annual amplitudes for all the cases are about 4 cm (see Figures 4, 5).

4 Conclusions

Based on the analyzes carried out, it can be concluded that the TWS variations over territories of the Vistula and Odra basins in Poland, are not very big. Even if there are trends in the data, they are rather small and not very significant from statistical point of view. Attention should be drawn to the similarity of independent GFZ and GSFC solutions, especially that the GFZ solution, as the only one, has a trend that is statistically significant. All in all, we can be happy that the ongoing climate change, at least in the years 2002-2016, did not cause significant changes in the value of the average total water storage in Poland. It would be advantageous to include in the comparisons also data from field measurements of groundwater level [14]. Such data, combined with total water storage values obtained from a chosen assimilation system of land data, like GLDAS, could provide independent estimation of TWS variations.

5 Acknowledgments

The paper was supported by the National Science Center, UMO-2015/17/ST10/03927 dated from 16.03.2016

All computations and plots were prepared using R: A language and environment for statistical computing, at <https://www.R-project.org/>

GRACE data are available at <http://grace.jpl.nasa.gov>, supported by the NASA MEaSUREs Program.

6 Bibliography

1. Baur O, Sneeuw N (2011), Assessing Greenland ice mass loss by means of point-mass modeling: a viable methodology. *Journal of Geodesy*, 85:607–615, 2011. doi: 10.1007/s00190-011-0463-1

2. Chambers JM (1992) Linear models. Chapter 4 of *Statistical Models* in S eds J. M. Chambers and TJ Hastie, Wadsworth & Brooks/Cole
3. Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6, pp 3–73
4. Famiglietti JS (2004) Remote sensing of terrestrial water storage, soil moisture and surface waters, in *The state of the planet: Frontiers and challenges in geophysics*, *Geophys. Monogr. Sr.*, vol. 150, edited by RSJ Sparks and CJ Hawkesworth, pp 197 – 207, AGU, Washington D. C.
5. Hipel KW, McLeod AI (2005), *Time Series Modelling of Water Resources and Environmental Systems*. Electronic reprint of our book originally published in 1994. <http://www.stats.uwo.ca/faculty/aim/1994Book/>
6. Hirsch RM, Slack JR, Smith RA (1982), Techniques for trend assessment for monthly water quality data, *Water Resources Research* 18, pp 107-121
7. Kendall M, Gibbons JD (1990) [First published 1948], *Rank Correlation Methods*. Charles Griffin Book Series (5th ed.). Oxford: Oxford University Press. ISBN 978-0195208375
8. Landerer FW, Swenson SC (2012), Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, Vol 48, W04531, 11 PP, doi:10.1029/2011WR011453
9. Luthcke SB, Sabaka TJ, Loomis BD, et al. (2013), Antarctica, Greenland and Gulf of Alaska land ice evolution from an iterated GRACE global mascon solution, *J. Glac.* 59(216), 613-631. doi:10.3189/2013JoG12J147
10. Mann HB (1945), Nonparametric tests against trend, *Econometrica*, 13, 245-259
11. R Core Team (2018), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
12. Ripley BD, Fortran code by Cleveland et al (1990) from 'netlib'
13. Rodell M, et al. (2004), The Global Land Data Assimilation System, *Bull. Am. Meteorol. Soc.*, 85, 381–394
14. Rzepecka Z, Biryło M, Kuczynska-Sieghien J, Nastula J, Pajak K (2017), Analysis of Groundwater Level Variations and Water Balance in the Area of Sudety Mountains, *Acta Geodynamica et Geomaterialia*, Vol. 14, No. 3 (187), Prague 2017, DOI: 10.13168/AGG.2017.0014, pp 313–321
15. Sakumura C, Bettadpur S, Bruinsma S (2014), Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models, *Geophys. Res. Lett.*, 41, 1389–1397
16. Swenson SC (2012), GRACE monthly land water mass grids NETCDF RELEASE 5.0. Ver. 5.0. PO.DAAC, CA, USA. Dataset accessed at <http://dx.doi.org/10.5067/TELND-NC005>
17. Swenson S.C., Wahr J (2006), Post-processing removal of correlated errors in GRACE data, *Geophys. Res. Lett.*, 33, L08402, doi:10.1029/2005GL02528
18. Watkins MM, Wiese DN, Yuan DN, Boening C, Landerer FW (2015), Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons, *J. Geophys. Res. Solid Earth*, 120, doi:10.1002/2014JB011547
19. Wickham H, Chang W (2015), devtools: Tools to Make Developing R Packages Easier. R package version 1.8.0. <http://CRAN.R-project.org/package=devtools>
20. Wiese DN, Landerer FW, Watkins MM (2016), Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution, *Water Resour. Res.*, 52, 7490–7502, doi:10.1002/2016WR019344

**Fourier Analysis of Cerebral Metabolism of Glucose:
Gender Differences in Mechanisms of Color Processing in
the Ventral and Dorsal Streams in Mice**

**Short title: Gender Difference in Fourier Analysis of Cerebral
Metabolism for Color**

✉ Philip C. Njemanze^{1*}

Email: philip.njemanze@chidicon.com

¹Neurocybernetic Flow Laboratory, International Institutes of Advanced Research Training,
Chidicon Medical Center, Owerri, Imo State, Nigeria.

<https://orcid.org/0000-0003-3215-5114>

Mathias Kranz²

Peter Brust²

²Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiopharmaceutical Cancer Research,
Department of Neuroradiopharmaceuticals, Research Site Leipzig, Germany.

**Keywords: * chromatic opponency * sex differences * light wave * light
particle * blood flow * frequency * resonance ***

Abstract

Introduction

Conventional imaging methods could not distinguish processes within the ventral and dorsal streams. The application of Fourier time series analysis was helpful to segregate changes in the ventral and dorsal streams of the visual system in male and female mice.

Materials and Methods

The present study measured the accumulation of [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) in the mouse brain using small animal positron emission tomography and magnetic resonance imaging (PET/MRI) during light stimulation with blue and yellow filters compared to darkness condition. Fourier analysis was performed using mean standardized uptake values (SUV) of [^{18}F]FDG for each stimulus condition to derive spectral density estimates for each condition.

Results

In male mice, luminance opponency occurred by S-peak changes in the subcortical retino-geniculate pathways in the dorsal stream supplied by ganglionic arteries in the left visual cortex, while chromatic opponency involved C-peak changes in the cortico-subcortical pathways in the ventral stream perfused by cortical arteries in the left visual cortex. In female mice, there was resonance phenomenon at C-peak in the ventral stream perfused by the cortical arteries in the right visual cortex in female mice during luminance processing. Conversely, chromatic opponency occurred by S-peak changes in the subcortical retino-geniculate pathways in the dorsal stream supplied by the ganglionic arteries in the right visual cortex.

Conclusions

Fourier time series analysis uncovered distinct mechanisms of color processing in the ventral stream in male, while in female mice color processing was in the dorsal stream. It demonstrated that computation of colour processing as a conscious experience could have a wide range of applications including in artificial intelligence and quantum mechanics.

1 INTRODUCTION

Color processing as a conscious experience could serve as a model for unraveling mechanisms of color memory information processing. In a recent study, we applied conventional functional positron emission tomography and magnetic resonance imaging (fPET/MRI) technique to demonstrate gender-related cerebral metabolic changes during color processing in a mouse model [1]. However, the conventional methods have poor image resolution and could not be used to segregate the processes taking place in various parts of the visual system. The visual system originates from the primary visual cortex and is organized into a ventral occipitotemporal stream for representation of ‘what’ system, while the dorsal occipitoparietal stream demonstrates the ‘where’ [2, 3]. The ventral stream is implicated in hierarchical processing of object representations that culminate in object recognition regardless of changes in the surrounding environment. On the other hand, the dorsal stream is involved in hierarchical processing that leads to the computation of complex motion in three dimensions. It has been suggested that, there is integration of both dorsal and ventral stream information [4]. Color is a complex multidimensional stimulus implicated in object recognition as well as in complex computations of location in three-dimensional space. Therefore, elucidating the mechanism of color processing in the ventral and dorsal streams could have a wide range of applications.

One useful approach to study the two streams would be to segregate the arterial network of the blood flow supply system in the visual cortex. The visual pathways and extrastriate cortex ‘color centers’ [5] obtain blood supply from the territories of the posterior (PCA) and middle (MCA) cerebral arteries [6]. It has been established that, color processing takes place within cortico-subcortical circuits working through the basal ganglia via the ventromedial occipital region to the posterior inferior temporal (PIT) cortex, the latter is located along the anterior third of the calcarine sulcus [7]. A reversed pathway of subcortico-cortical circuit may also be possible. The vascular supply to the visual system as in other regions of the brain comes from the principal arteries of the circle of Willis that give rise to two different systems of secondary vessels called the ganglionic and cortical systems. The cortical and ganglionic systems are independent of each other and do not communicate at any point in their peripheral distribution. There is, between the parts supplied by the two systems, a borderline of diminished nutritive activity [6]. The ganglionic system perfuse mainly the dorsal stream from the subcortical region to the cortical region, while the cortical system perfuse the ventral stream from the cortical region to the subcortical region.

Color is a brain memory computational process of at least three primary qualities of hue, saturation (chroma) and lightness (value). The two main memory processes

associated with color vision are simultaneous color contrast and color constancy [8-15]. Simultaneous color contrast is the phenomenon that surrounding colors profoundly influence the perceived color [10]. Others have suggested that, the conditions for simultaneous color contrast, implies having a chromatic contrast detector subserving one area of the chromatic space, excite a chromatic detector of opposite type and/or inhibit a chromatic detector of the same type in neighboring areas of chromatic space [9]. According to a recent claim, the mechanism for simultaneous color contrast may involve wavelength-differencing [15].

Color processing has been studied using mean cerebral blood flow velocity (mCBFV) measurements indexed using transcranial Doppler and demonstrated selective response to colors of different wavelengths in humans [16]. Furthermore, Fourier time-series analysis has been applied to mCBFV to demonstrate changes related to color processing [17-19] and facial processing [20]. Blood flow and metabolism, therefore, have been considered virtually equivalent, indirect indices of brain function [21]. However, regional uncoupling of CBF and CMRO₂ was found, during neuronal activation induced by somatosensory stimulation [22]. Overall, rCBF has been found to correlate to mCBFV [23]. The rationale for application of Fourier analysis [24, 25] to characterize the periodicity of biological systems [25] and in particular the cerebrovascular system [26] has been studied. PET images rendered in units of standardized uptake values (SUV) of [¹⁸F]FDG in response to stimuli can be subjected to similar Fourier time series analysis.

Therefore, the application of Fourier analysis could separate the frequency peaks from the ‘dorsal stream’ supplied by the ganglionic branches, from those of the ‘ventral stream’ that obtain perfusion from the cortical branches of the brain arteries. We presume that, the vessels of the cortical arterial system are not so strictly “terminal” as those of the ganglionic system, and perfuse areas that could be mapped to retinotopic structures in the mouse visual cortex [27]. We postulate that, following the cortical arterial supply system there could be a cortico-subcortical top-bottom feedback mechanism through cortico-subcortical circuits [28] for color processing, and the reverse subcortico-cortical bottom-up feed-forward mechanism [28] through subcortico-cortical circuits perfused by ganglionic arteries.

Color is a memory process and could be characterized by known models of the synaptic and cellular events that may be associated with memory formation. We postulated and tested that, the Fourier time-series analysis of the frequency-domain of SUV mean values as a surrogate marker of cerebral metabolism may uncover the underlying memory mechanisms explained by the phenomena of long-term potentiation (LTP) [29] and long-term depression (LTD) [30], primarily because they exhibit numerous properties expected of a synaptic associative memory mechanism, such as rapid induction, synapse specificity, associative interactions, persistence, and dependence on correlated synaptic activity. Given these important features, LTP and LTD remains only models of the synaptic and cellular events that may underlie memory formation.

To further elucidate the mechanism implicated in processing of light stimulus, certain presumptions and several definitions of concepts were made in the present work related to the physical characteristics of light stimulus. The light stimulus has dual

wave and particle nature which is characterized by physical properties of amplitude, phase difference, wavelength, frequency and resonance phenomenon. Therefore, the mechanisms implemented in the processing of light stimuli must include mechanistic strategies for dealing with these physical properties. Five mechanistic strategies in response to the physical properties of light could include: (a) Changes in peak amplitude; (b) Phase difference; (c) Wavelength-differencing; (d) Frequency-differencing; and (e) Resonance.

The major aim of the present work is to demonstrate the application of Fourier spectral density analysis to determine overall and specific effects of visual stimulations in the dorsal and ventral streams of the visual cortex in male and female mice, respectively.

2 MATERIALS AND METHODS

2.1 Animals

All procedures were in compliance with the ‘Principles of laboratory animal care’ (NIH publication no. 85e23, revised 1985) and were approved by the Institutional Animal Care and Use Committee in the state of Saxony, Germany as recommended by the responsible local animal ethics review board (Regierungspräsidium Leipzig, TVV08/13, Germany). The experimental setup (Figure 1A-F) using a custom-made photostimulation device Chromatoscope in a mice model (Figure 1A-B) followed by small animal PET/MRI (Figure 1C) has been described in detail elsewhere [1].



Fig. 1. (A-F) shows the experimental setup (Figure 1A) in close view of the mice Chromatoscope (Figure 1 B), with the animal place within the gantry of the PET/MRI (Figure 1 C). The animals were housed under controlled conditions with free access to food and water (Figure 1D). The heart rate, respiration (Figure 1E) and anesthetic airflow were monitored (Figure 1F).

All experiments were performed under isoflurane anaesthesia and all efforts were made to minimize pain. Five male and five female CD-1 mice (10 - 12 weeks, 22 – 28 g) were housed under a 12 hour: 12 hour light:dark cycle (lights one at 7:00 am) at 24°C in a vented temperature-controlled animal cabinet (HPP108, MEMMERT GmbH & Co. KG; Germany) (Figure 1D), with free access to food and water. The heart rate, respiration and anesthetic airflow were monitored (Figure 1E-F). PET studies were conducted on the same animals repeatedly on consecutive days without randomization to keep the daytime of measurement (e.g. the glucose/insulin levels) constant. There was no significant change in weight of the animals over the several days of study in male and female mice. The weights in male mice were (Day 1 = 34.5 ± 2.8 g; Day 2 = 34.4 ± 2.4 g; Day 3 = 33.7 ± 2.3 g; Day 4 = 34.7 ± 2.3 g; Day 5 = $34.3 \pm$

2.5g; Day 6 = 34.6 ± 2.5 g; Day 7 = 34.1 ± 2.6 g) and in female mice were (Day 1 = 25.6 ± 1.7 g; Day 2 = 25.4 ± 1.3 g; Day 3 = 25.5 ± 1.5 g; Day 4 = 25.6 ± 1.2 g; Day 5 = 26.4 ± 1.4 g; Day 6 = 26.2 ± 1.4 g; Day 7 = 26.5 ± 1.3 g). The intraperitoneally injected radiotracer ($[^{18}\text{F}]\text{FDG}$) dose was in male mice (Day 1 = 12.05 ± 1.23 MBq; Day 2 = 12 ± 0.9 MBq; Day 3 = 11.7 ± 1.2 MBq; Day 4 = 10.6 ± 0.5 MBq; Day 5 = 12.1 ± 1.7 MBq; Day 6 = 10.8 ± 1.2 MBq; Day 7 = 11.9 ± 1 MBq) and female mice (Day 1 = 12.7 ± 1.23 MBq; Day 2 = 12.7 ± 1.3 MBq; Day 3 = 12.6 ± 0.9 MBq; Day 4 = 13.9 ± 0.7 MBq; Day 5 = 11.4 ± 0.9 MBq; Day 6 = 12.3 ± 1.2 MBq; Day 7 = 12 ± 1.4 MBq) and did not vary significantly over the several days of the study. The male (10.1 ± 1.5 mmol/L) and female (7.8 ± 1.8 mmol/L) mice random blood sugar levels were similar. All animals were at the end of the study euthanized by cervical dislocation under anesthesia.

2.2 Light Stimulation Studies

The experimental setup for the fPET/MRI study was displayed in Figure 1 (A-F). The mice was placed in prone position on a special heated mouse pad with head affixed to a mouth piece (Figure 1A). The eyes were positioned and fixed for 20 mins light stimulation through the double barrel of the light source Chromatoscope (Figure 1B) [1]. Subsequently, a whole body PET scan was started for a duration of 20 mins using a preclinical scanner (Figure 1C). The animals were housed in an animal cabinet controlled day-light regimen with free access to food and water (Figure 1D). The respiration and anesthetic gas flow were monitored (Figure 1E-F).

The stimulation device is a custom-made double barrel tunnel placed around both eyes and the nose ridge to separate both visual fields, and has been described in detail elsewhere [1]. Both eyes were open at all times. At the end is a white screen illuminated by a remote light source. There is a groove before the screen that allows insertion of filters into the right and left visual fields respectively. At onset before stimulation, the animal was positioned with both eyes open and fixed peeping through the double barrel tunnel connected to a light source behind the white screen. The light source was a tungsten coil filament, of a general service lamp ran at a constant 21 V and 150 W, with maximum light output of the bulb of 40,000 foot candles, (430,000 lux), power at tip of fibre at a maximum bulb intensity of 1.4 W/m^2 , with a color temperature of about 3200 K and approximately 20 lumens/watt. (OSL2 High-Intensity Fiber Light Source, Thorabs Inc., Newton, New Jersey, USA).

The mouse eye had a fully dilated pupil with a numerical aperture of 0.49. The light was presented to the eye over a circular region of $\sim 24^\circ$ diameter on the retina. The laboratory room illumination was by ceiling-mounted fluorescent lamps (150 lux). The stimulations were accomplished at about the same time of day in the same animal over the several days of study, to maintain synchronization (entrainment) to nature's biorhythm cycle of 24 hours. The study in all animals in one day lasted for 6 hours from about 9:30 AM to 3:30 PM, with most male mice studied in the morning hours and female mice in the afternoon hours.

2.3 Color vision testing in mice using PET/MRI

The two types of cone pigment found in the mouse retina have spectral absorption curves λ_{\max} as follows: UV (ultraviolet) (360 nm); M(510 nm) [32]. The Wratten gelatine filter 47 has a pass-band for UV+deep blue. However, mainly the deep blue reaches the retina since most of the UV may not penetrate beyond the anterior of the eye. The selection for the respective color stimulation, was the following Wratten filters: Deep Blue (No. 47B) with short dominant wavelength of 452.7 nm and Deep Yellow (No. 12) with medium dominant wavelength of 510.7 nm (Kodak Photographic Filters). The mice were placed in prone position in a special mouse bed (heated up to 37°C), with the head fixed to a mouth piece for the anaesthetic gas supply with 1.8% of isoflurane in 40% air and 60% oxygen (Anaesthesia unit U-410, AgnTho's AB, Lidingö, Sweden; Gas blender 100, MCQ Instruments, Rome, Italy) while the respiration was monitored for the duration of investigation.

The stimulation was performed with the anesthetized animal positioned with both eyes open and

fixed peeping through the double barrel optic connected to a light source behind the white screen. It is known that mice under narcosis had their eyes open, pupil maximally dilated and did not blink. One eye was occluded to achieve short-term monocular deprivation for the duration of the stimulation (20 minutes) to excite the contralateral eye. The closure of the eye was achieved by covering with 5% dexpantenol ointment (Bepanthen, Bayer, Germany). The animals received an i.p. injection of 12 ± 1 MBq [^{18}F]FDG (Supplier: Prof. M. Patt, Department of Nuclear Medicine, University Hospital Leipzig, Leipzig, Germany) immediately followed by one session of 20 min stimulation. Thereafter, a whole body PET scan was started for a duration of 20 min, using a preclinical Scanner (nanoScan1PET-MRI, Mediso Medical Imaging Systems, Budapest, Hungary) as shown in the timeline of the scan protocol demonstrated elsewhere [1].

Each animal investigation was performed only once a day with one stimulation and one [^{18}F]FDG injection was applied followed by a 24 h recovery period. The protocol was setup as a high-throughput experimentation with time-shift overlaid parallelization. This meant that rather than carrying out single experiments in one animal after another, the procedure overlaid several tasks with start of stimulation in one animal preceding the other by about 25 minutes, therefore shortening the overall time for experiments.

The following seven stimulation conditions were used:

- dark: both eyes (1) closed (dark)
- light: left (2) or right (3) eye open and subjected to standard light source (short: LightL, LightR)
- blue: left (4) or right (5) eye open and subjected to standard light source with blue filter (short: BlueL, BlueR)
- yellow: left (6) or right (7) eye open and subjected to standard light source with yellow filter (short: YellowL, YellowR).

2.4 Acquisition and analysis of PET and MRI data

Every PET image was corrected for random coincidences, dead time, scatter and attenuation, based on a whole body MRI scan (T1 weighted gradient echo sequence (GRE), $T_R=20$ ms; $T_E=6.4$ ms; matrix size: $160 \times 160 \times 62$, resolution: $0.04 \times 0.04 \times 0.05$ cm, slice thickness: 0.5mm acquisition duration 12 mins) immediately following the PET acquisition. The anatomic details were identified on the T1 images from this sequence. PET data were collected by a continuous whole body scan during the entire investigation in list-mode (scan duration 20 min). Thereafter, the data was reconstructed into 4 uniform time frames (5 min each). Parameters for reconstruction for the list mode data were 3D-ordered subset expectation maximization (OSEM) with 4 iterations and 6 subsets, energy window: 400-600 keV, coincidence mode: 1-5, ring difference 81.

Two observers performed the delineation of the volume of interest (VOI) and data analysis in consensus with ROVER (ABX advanced biochemical compounds, Radeberg, Germany, v.2.1.15). The right and left hemispheres were identified using the MRI information from the GRE scan. Presumptions were made for the localization of the respective brain region. For example, that VOI in the visual cortex with tracer concentration is a sample volume of a cylindrical mask in a space stretching from the primary visual cortex to the extrastriate cortex perfused by both the ganglionic branches (e.g. lenticulostriate arteries) and cortical arteries from the main stems of the middle cerebral artery (MCA) and posterior cerebral artery (PCA) [1,6].

For the definition of contour VOI, a stack of planar, closed polygons called region-of-interest (ROI) is applied. For the VOI statistics the contours are manually and semi-automatically outlined on the loaded images, and the pixels contained within the contour boundaries are considered. The contour vertices coordinates are defined as the (x, y, z) triples, of which the x , y and z offsets are in [mm] from the image origin. Data analysis delineated three separated VOIs (Fig. 2A-B). First, a whole cortex (Ctx) cylindrical mask with (x, y, z) pixel size $(20, 20, 20)$ or $(0.6 \text{ cm}, 0.6 \text{ cm}, 0.6 \text{ cm})$ was created to cover most arteries of the two parts of this brain area. It was centred at the midline in coronal view of the PET/MR image. Using the transverse plane for orientation of tracer accumulation, the VOI extends from the ventromedial occipital region through the posterior inferior temporal cortex. Two sub-volume VOIs with mask (x, y, z) pixel size $(10, 10, 10)$ or $(0.3 \text{ cm}, 0.3 \text{ cm}, 0.3 \text{ cm})$ were placed from the midpoint to the right border (visCtxR) and to the left border (visCtxL) of the Ctx mask.

The VOIs in each hemisphere were defined in homologous areas on both sides of the brain in coronal plane (Fig. 3). Hemisphere-specific [^{18}F]FDG accumulation was expressed as standardized uptake value (SUV) [33]. The SUV is defined as the ratio of the tissue radioactivity concentration c (kBq/g) at time point t , and the injected activity divided by the body weight. The investigation of specific tracer uptake was performed at four mid-frame time points: 29.5, 34.5, 39.5 and 44.5 min, after the radiotracer injection. The SUV values were obtained over time for group A (males) and group B (females) under the aforementioned color stimulation conditions, for the

both brain regions visCtxR and visCtxL. For a full quantification of cerebral metabolic rate of glucose (CMR_{Glc}) a dedicated kinetic modeling as well as arterial blood sampling would be required. Due to the small blood volume of mice, repeated arterial blood sampling is challenging. Furthermore the study design did not allow us to obtain PET data in this study immediately following the injection of [^{18}F]FDG, therefore we refrain from kinetic modeling, but used the SUV values as a surrogate marker for CMR_{Glc} .

2.5 Statistical Analysis

Results were given as mean \pm SD and plots represented as Mean/SE/1.96*SE where applicable. Stimulus effects were assessed by paired *t*-test statistics of stimulus condition compared to dark condition, and one-way analysis of variance (ANOVA), the *p*-value was set at $p < 0.05$. Multivariate Analysis of Variance (MANOVA) with repeated measures was applied. The latter was followed by planned *t*-tests to examine specific differences. The level of significance was at $p < 0.05$.

2.6 Fourier Analysis

The spectrum analysis was applied to examine the cyclical patterns of data of the mean \pm SD SUV values. The rationale for exploration of the cyclical components is that it may correlate to the frequency of neuronal discharges in a given region of the brain during the observed phenomenon. It is hoped that we could uncover just a few recurring cycles of different lengths in the time series of metabolic activity that may reveal the seemingly random noise of neuronal activity. Fourier transform algorithm was applied using standard software (Time series and forecasting module, Statistica for Windows, StatSoft, OK, USA). All other analyses were performed using the software packages Statistica for Windows (StatSoft, OK, USA) and SPSS Version 20 (IBM). The spectrum analysis was applied to the SUV values provided in Table 1, to obtain spectral density coefficients given in Table 2, in male and female mice, respectively.

The purpose of the Fourier analysis is to decompose the original time series into underlying sine and cosine functions of different frequencies, so as to identify the important frequency region. The wavelength of a sine or cosine function expressed as the number of cycles per unit time (*frequency*) is denoted as ν . The period T of a sine or cosine function is defined by the length of time required for one full cycle. Thus, the period T is the reciprocal of frequency, or $T = 1/\nu$. One approach used is to restate the matter as a linear multiple regression model where the dependent variable is the observed time series, and the independent variables are the sine functions of all possible (discrete) frequencies. Thus the multiple regression model could be expressed as:

$$x_t = a_0 + \sum_{k=1}^q [a_k * \cos(\lambda * t) + b_k * \sin(\lambda * t)]$$

The notation λ is the frequency expressed in radians per unit time, given by $\lambda = 2\pi\nu$, where $\pi = 3.1416$. The cosine parameters a_k and sine parameters b_k are regression coefficients that indicate the degree of correlation with the data. There are q different sine and cosine functions; of which there are $n/2+1$ cosine functions and $n/2-1$ sine functions. This would mean that there would be many different sinusoidal waves as there are data points that would be able to completely replicate the series from the underlying functions. If the number of data points in the series are odd, then the last data point is ignored; and there must be at least two data points of high peak and low trough for a sinusoidal function to be identified. In other words, the standard and most efficient Fourier algorithm requires that the length of the input series is equal to a power of 2 [24]. If this is not the case, additional computations have to be performed. Fourier analysis will identify the correlation of sine and cosine functions of different frequency within the observed data. If a large correlation is identified it could be said that there is strong periodicity of the respective frequency in the data.

The Fourier series states that, any periodic function (or signal) can be expressed as a summation of orthogonal pair of matrices with one fundamental frequency and infinite number of harmonics. In other words, the sine and cosine functions are mutually independent (or orthogonal), thus the squared coefficients of each frequency may be summed to obtain the periodogram:

$$P_k = \text{sine coeff.}_k^2 + \text{cosine coeff.}_k^2 * n/2.$$

Where P_k is the periodogram value at frequency ν_k and n is the overall length of the series. However, the periodogram values are subject to substantial random fluctuation that could yield many chaotic periodogram spikes. Hence, in practice, the plots utilized are the frequencies with the greatest spectral densities; that is, the frequency regions consisting of many adjacent frequencies, that contribute most to the overall periodic behaviour of the series. This is accomplished by smoothing the periodogram values via a weighted moving average transformation. In the Hamming window for each frequency, the weights for the weighted moving average of the periodogram values are computed as follows:

$$w_j = 0.54 + 0.46 * \cos(\pi * j/p)$$

(for $j = 0$ to p)

$$w_{-j} = w_j$$

(for $j \neq 0$)

All weight functions will assign the greatest weight to the observation being smoothed in the center of the window and increasingly smaller weights to values that are further away from the center. A ‘white noise’ input series will result in periodogram values that follows an exponential distribution.

2.7 Software Procedure for Data Analysis

To obtain the required time series, the 20 data points for each stimulus condition were analysed for males and females, respectively. The analysis begins in Fourier Analysis dialog window, by choosing spectral density estimates and the Hamming window, then *Plot* to display cyclical patterns in male and female mice, respectively. The spectral density estimates, derived from single series Fourier analysis were plotted, and the frequency regions with the highest estimates were marked as peaks. The spectral density estimates between two minima including the peak (as maxima) were analysed to examine the effects of stimuli on cortical and subcortical sites. The spectral density peaks were identified as cortical (C-peak) and subcortical (S-peaks) whose peaks occurred at regular frequency intervals of 0.2 and 0.4, respectively. For evaluation of stimulus responses, the area under the curve derived for a particular stimulus was compared to that derived from another stimulus. Further statistical analysis was carried out using the five data points from trough-to-peak-to-trough for the C-peak and S-peak, respectively, as shown in Table 2.

3 RESULTS

Figures 2 (A-B) show the *f*PET/MRI images (Fig. 2A) and volumes of interest (VOI) in mouse cerebral vasculature by X-ray micro-CT (Fig. 2B). In male mice, Fig. 2A (black circle) shows a tracer distribution in the left visual cortex (visCtxL) that is shaped like a Canadian ‘duckpin’ with a short small head at the top dorsal cortical region, a narrow neck (open arrow), and a long fat base into a wide area of spread in the central ventral subcortical region during Blue light stimulation. In male mice, the “upright Canadian duckpin” tracer spread may suggest a small focal brain area of ‘arousal’ at the head placed within the dorsal cortical region of the left visual cortex and at the base a secondary wide area of spread in the subcortical region. On the other side, in the right visual cortex (visCtxR) there is intense concentration of tracer and no peculiar organization of the tracer distribution. In female mice, Fig. 2B (white circle) shows an “inverted Canadian duckpin” tracer distribution in the right visual cortex (visCtxR) with a small area of “arousal” at the head within the subcortical region in the right visual cortex and at the base in the dorsal temporo-occipital cortical region. The contralateral left visual cortex (visCtxL) does not show any remarkable tracer distribution.

Figures 2C-D, show the MRI images in male (Fig. 2C) and female (Fig. 2D) mice used for anatomic orientation. Figure 2E, shows the VOI setup in mouse brain on micro-CT image [39] in coronal view. The figure illustrates the VOIs used for data analysis in relation to the brain vessel system, red: Ctx, black: visCtxL and green: visCtxR. Visual observation due to low image resolution could not clearly delineate vascular territories and borders of cortical and subcortical distribution. Hence, more reliable methods are needed. Figure 2F, shows the schematic diagram of the arterial tree of the Circle of Willis and the relationship to the ROI to the cortical and ganglionic branches of the middle cerebral artery in the mouse brain. The left side of the

Circle of Willis shows the ‘male model’ of radiotracer distribution, while the right side shows the ‘female model’.

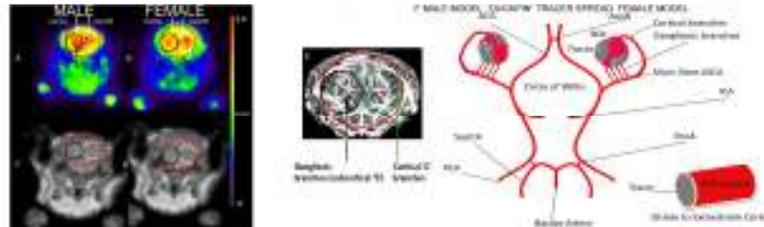


Fig. 2. (A-F) shows the β PET images of radiotracer accumulation in male (Fig. 2A) and female mice (Fig. 2B) and the MRI images in male (Fig. 2C) and female (Fig. 2D) mice. Figure 2E, shows the VOI setup in mouse brain on micro-CT image [39] in coronal view. Figure 2F, shows the schematic diagram of the arterial tree of the Circle of Willis and the relationship to the ROI to the cortical and ganglionic branches of the middle cerebral artery in the mouse brain. The radiotracer distribution shows a ‘male model’ and a ‘female model’.

We analyzed the mean \pm SD SUV data obtained in direct measurements in male (Table 1A) and female mice (Table 1B), respectively. A MANOVA with repeated measures was performed on the mean \pm SD SUV values, with a 7 x 2 x 2 design: seven levels of stimulation of the visual cortex, Stimulations (Dark, Light R, Light L, Blue R, Blue L, Yellow R, Yellow L), two levels of Visual Cortex (visCtxR, visCtxL) and two levels of Gender (male and female), with the mean \pm SD SUV during stimulations were analyzed as the dependent variable. There were main effects of Stimulations, $F(6,228) = 7.621$, $MS = 0.356$, $p < 0.0000001$; Visual Cortex, $F(1,38) = 7.157$; $MS = 0.026$; $p < 0.05$; and Gender, $F(1,38) = 15.15$, $MS = 2.065$, $p < 0.001$. There was a Stimulation x Gender interaction, $F(6,228) = 6.405$, $MS = 0.299$, $p < 0.0001$. There was also a Stimulations x Visual Cortex interaction $F(6,228) = 4.21$, $MS = 0.0141$, $p < 0.001$. The analysis of mean \pm SD values of SUV detected changes related to the overall effects of visual stimulations in the visual cortex but did not distinguish changes in the dorsal and ventral streams, respectively.

The spectral density data obtained from the times series analysis SUV values (Table 1A-B) are provided for male and female (Table 2). A MANOVA with repeated measures was performed with a 7 x 2 x 2 x 2 design: seven levels of Stimulations (Dark, Light R, Light L, Blue R, Blue L, Yellow R, Yellow L), two levels of Visual Cortex (visCtxR, visCtxL), two levels of Peaks (C-peak, S-peak) and two levels of Gender (male, female). There was a main effect of Stimulations, $F(6,102) = 8.65$, $MS = 0.094$, $p < 0.0000001$. There was a Stimulations x Gender interaction, $F(6,102) = 7.68$, $MS = 0.083$, $p < 0.000001$. There was a Stimulations x Visual Cortex x Gender interaction, $F(6,102) = 3.4$, $MS = 0.00098$, $p < 0.01$. There was a Stimulations x Visual Cortex x Peaks interaction, $F(6,102) = 3.4$, $MS = 0.00075$, $p < 0.01$. Fourier spec-

tral density analysis demonstrated the overall effects. Further analysis was undertaken to show specific effects of spatial opponency, luminance opponency and chromatic opponency during visual stimulations in the visual cortex in the dorsal (cortical C-peak) and ventral (subcortical S-peak) streams, respectively. Detection of these effects would dramatically improve the temporal and spatial resolutions of conventional PET/MRI imaging.

In male mice, Table 3 (top panel) demonstrates spatial, luminance and chromatic opponency determined by Fourier spectral density coefficients (mean \pm SD) values in paired *t*-test results in the right visual cortex through the left eye (LvisCtxR) and left visual cortex through the right eye (RvisCtxL), and percent changes from Dark condition ($\Delta\%$ Dark). Figure 3(A-H) shows the spectral density plots of Fourier coefficients. In male mice, under Dark condition, while there was no significant difference in amplitudes between C-peak and S-peak in the right visual cortex (visCtxR) (Fig. 3A), in the left visual cortex (visCtxL) (Fig 3B), the S-peak was significantly higher than the C-peak, which demonstrated spatial opponency, ($p < 0.05$). During white Light stimulation through the left eye in the right visual cortex (LvisCtxR) (Fig. 3C), the C-peak was attenuated by -66.7%, ($p < 0.05$), with no significant change in S-peak, but resulted in significant presence of spatial opponency, ($p < 0.05$). During stimulation with white Light through the right eye in the left visual cortex (RvisCtxL) (Fig. 3D), there was accentuation of S-peak by 210%, ($p < 0.05$), which in comparison to Dark condition induced luminance opponency, ($p < 0.05$), in the subcortical region in the ventral stream. Luminance opponency was present in the cortico-subcortical circuit of combined contrast using C-peak and S-peak, ($p < 0.05$). During stimulation with Blue light through the left eye in the right visual cortex (LvisCtxR) (Fig. 3E), there was attenuation of C-peak by -21%, ($p < 0.05$), but no significant change in S-peak. However, during stimulation with Blue light through the right eye in the left visual cortex (RvisCtxL) (Fig. 3F), there was no significant change in C-peak, but the S-peak was remarkable attenuated by -92.7%, ($p < 0.05$). During stimulation with Yellow light through the left eye in the right visual cortex (LvisCtxR) (Fig. 3G), the C-peak was attenuated by -21%, ($p < 0.05$), and S-peak was also attenuated by -9.8%, ($p < 0.05$). However, during stimulation with Yellow light through the right eye in the left visual cortex (RvisCtxL) (Fig. 3H), the C-peak did not change significantly, but there was a marked attenuation of the S-peak by -87%, ($p < 0.05$). In other words, during stimulation with the Blue/Yellow pairs, there was marked attenuation of C-peaks and S-peaks across brain regions but not in the C-peaks in the left visual cortex through the right eye (RvisCtxL), which resulted in significant differences between Blue versus Yellow for chromatic opponency, ($p < 0.05$) in the cortical region in the dorsal stream. The chromatic opponency was present in the cortico-subcortical circuit of combined C-peak and S-peak Blue/Yellow contrast, ($p < 0.01$). Spatial opponency was not elicited during stimulation with Blue and Yellow colors. Overall, in male mice, luminance opponency was implemented in the subcortical region of the left visual cortex in the ventral stream, while chromatic opponency was present in the cortical region of the left visual cortex in the dorsal stream.

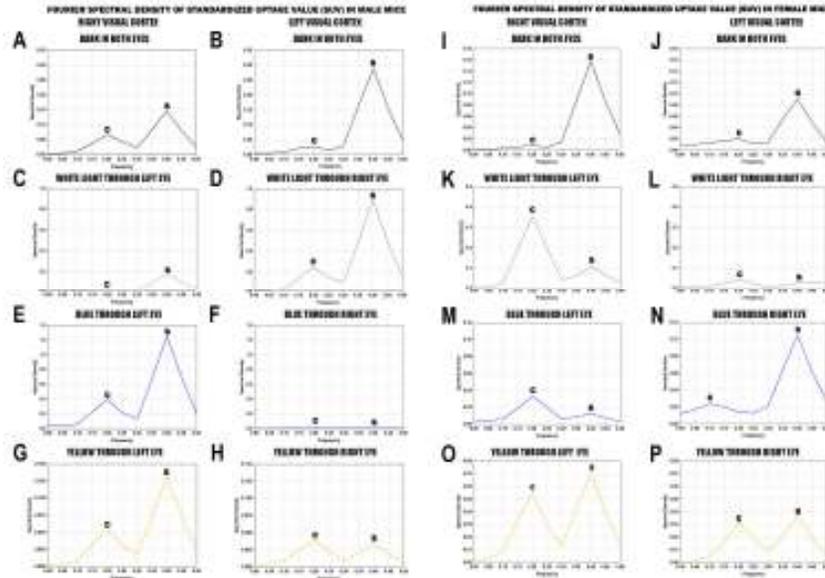


Fig. 3. (A-P) shows the spectral density plots of Fourier coefficients for male, Fig. 3(A-H), and female Fig. 3 (I-P) mice, respectively.

In female mice, Table 3 (bottom panel) demonstrates spatial, luminance and chromatic opponency determined by Fourier spectral density coefficients (mean \pm SD) values in paired *t*-test results in the right visual cortex through the left eye (LvisCtxR) and left visual cortex through the right eye (RvisCtxL), and percent changes from Dark condition ($\Delta\%$ Dark). In female mice, under Dark condition, there was significant difference in amplitudes in the right visual cortex (visCtxR) (Fig. 3I), the S-peak was significantly higher than the C-peak, which demonstrated spatial opponency, ($p < 0.05$). On the other hand, in the left visual cortex (visCtxL) (Fig. 3J), the C-peak and S-peak were not significantly different. During white Light stimulation in the right visual cortex through the left eye (LvisCtxR) (Fig. 3K), there was marked broad-spectrum accentuation of C-peak by 3020% with a wide range of standard deviation (SD) and was not significantly different from Dark condition. There was a wide range of spectral densities below the area of the curve (Fig. 3K), which in turn indicates a broad spectrum of constituent frequencies activated by the broad spectra of white Light photonic frequencies from low to very high with no selective response. In the contralateral left visual cortex through the right eye (RvisCtxL), both C-peak and S-peak were not significantly different from Dark condition. During stimulation with

Blue light through the left eye in the right visual cortex (LvisCtxR) in female mice (Fig. 3M), there was significant accentuation of C-peak by 220%, ($p < 0.05$), and attenuation of S-peak by -90.5%, ($p < 0.05$), suggestive of a selective response to the high-frequency Blue light. In the contralateral left visual cortex through the right eye (RvisCtxL) (Fig. 3N), the changes were not significantly higher than Dark condition but with a tendency for attenuation of C-peak and accentuation of S-peak which created a condition for spatial opponency, ($p < 0.05$). During stimulation with Yellow light through the left eye in the right visual cortex (LvisCtxR) (Fig. 3O), the C-peak and S-peak were not significantly different from that in Dark condition. During stimulation with Yellow light through the right eye in the left visual cortex (RvisCtxL) (Fig. 3P), there was no significant change in C-peak, but a significant attenuation of S-peak by -75%, ($p < 0.05$). The chromatic opponency of Blue/Yellow pairs was present in the subcortical region (S-peak) in the ventral stream in the right visual cortex, ($p < 0.05$). The chromatic opponency was implemented in the cortico-subcortical circuit of combined C-peak and S-peak Blue/Yellow contrast, ($p < 0.01$). Overall, in female mice, the response to Blue light stimulation was frequency-modulated which induced resonance phenomenon in the cortical region in the dorsal stream of the right visual cortex, while in the subcortical region in the ventral stream in the right visual cortex, chromatic opponency was implemented.

4 DISCUSSION

4.1 Gender Differences in Mechanisms for Color Processing

We accomplished our main objective to use Fourier analysis to identify gender-related white light and color processing mechanisms in the dorsal and ventral streams of the visual system. In male mice, luminance opponency occurred in the subcortical region in the left visual cortex, while chromatic opponency was implemented in the cortical region of the left visual cortex in the ventral stream. On the other hand, in female mice, luminance evoked a resonance response in the cortical region in the dorsal stream in the right visual cortex, while chromatic opponency was implemented in the subcortical region in the dorsal stream in the right visual cortex. The frequency-related resonance phenomenon while present in female mice was absent in male mice, that is, female mice had preference for processing the frequency of particles of light, while male mice has the preference of processing the wavelength of light. In other words, male mice differentiated color light stimulus by wavelength-differencing, while female mice implemented frequency-differencing [18-19]. The gender differential processing of light as a wave and as a particle to our knowledge has not been firmly established [17-19], and requires further studies, because of the potential implications for understanding the differences in processing in the visual sensory system.

We postulate that, there are two separate male (Fig. 4A) and female (Fig. 4B) models for color processing in the mouse brain, shown in the schematic diagram (Fig. 4 A-B).

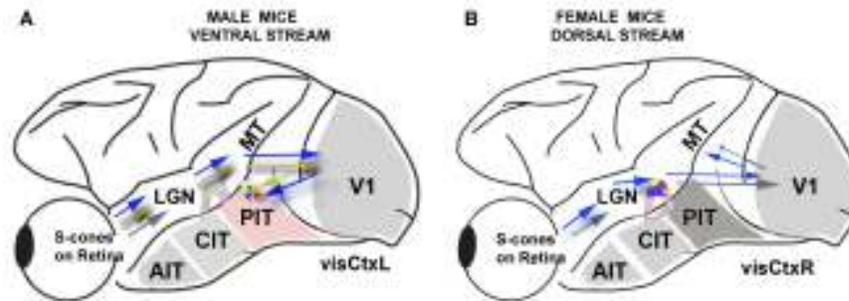


Fig. 4. (A-B). The schematic diagram of two separate male (Fig. 4A) and female (Fig. 4B) models for color processing in the mouse brain.

Figure 4A shows, the male mouse model, comprising a wavelength-differencing system that receive inputs from retinal S-cones on activation of the visual pathway from the retina, through the dorsal lateral geniculate nucleus of the thalamus (LGN) to the primary visual cortex (V1) traversing the associative visual areas (V2, V3) and areas like the fusiform gyrus and lingual gyrus with others collectively referred to as visual area 4 (V4), to the angular gyrus implicated in the higher order processing of colors along the ventral stream [34] through posterior (PIT), central (CIT) and anterior inferior temporal (AIT) cortex. The color processing is accomplished between two color spaces by wavelength-sensitive cortical neurons with axons in the cortical-subcortical circuits. The IT contains regions that are relatively more responsive to color (seen in Figure 4A as an area checkered with colors in PIT). The ventral pathway is thought to represent stable attributes of objects (object quality) [35, 36]. The processing of luminance takes place within the retino-subcortical circuits by light-sensitive subcortical neurons.

Female mice demonstrate a frequency-differencing system by neurons in the subcortical region that differentiate high frequency color such as blue from low frequency color such as yellow. Figure 4B, shows the schematic model in female mice, which demonstrates retino-subcortico-cortical circuits that receive inputs from the retinal S-cones on activation of the visual pathway from the retina to specialized frequency-selective neurons in the dorsal lateral geniculate nucleus of the thalamus (LGN), with axons ascending within the dorsal stream to the primary visual cortex (V1) and MT region of the brain. Conventionally, it is presumed that, the dorsal stream was implicated in encoding dynamic spatiotemporal relationships among visual objects (object action) [35, 36].

4.2 Fourier Analysis to Differentiate Cortical and Ganglionic Vessels

The primary questions were to resolve the origins of C-peak and S-peak, their anatomic correlates and functional significance. The C-peak and S-peak occurred at multiples of the first harmonic or the fundamental frequency, at the second and third harmonics, respectively. It has been demonstrated that, in the vascular system, the first five harmonics contain 90% percent of the pulsatile energy of the system [37]. These frequencies could be converted to cycles per second (Hz), assuming that the fundamental frequency of cardiac oscillation was the mean heart rate. The fundamental frequency f of the first harmonic was determined by the mean heart rate per second of CD-1 mice = 515 ± 30 bpm/60 seconds = 8.6 Hz [38]. Thus, the distance of the reflection site for fundamental frequency could be presumed to emanate from a site at $D_1 = \frac{1}{4}\lambda$ or $c/4f$, or $1.91 \text{ m/s} / (4 * 8.6 \text{ Hz}) = 0.055 \text{ m/s}$ or 5.5 cm; where $c = 1.91 \pm 0.44$ m/s, is the wave propagation velocity [38]. These distances are not physical measurements but approximate the actual arterial lengths. Taking into account vascular tortuosity, the estimated distance (5.5 cm) approximates that from the terminal vessels in the brain, to an imaginary site of summed reflections from the aorto-iliac junction of the mice. The C-peak occurred at the second harmonic, such that, the estimated arterial length given by $D_2 = 1/8\lambda$ or $c/8 \times 2f$, or $1.91 \text{ m/s} / (8 * (2 * 8.6 \text{ Hz} \text{ or } 17.2 \text{ Hz})) = 0.0139 \text{ m}$ or 1.4 cm, approximates the visible arterial length from the main stem of the major cortical arteries around the cerebral convexity to the end occipito-temporal junction as shown in mouse [39]. The cortical frequency of 17 Hz is within the beta rhythm range (~14-18 Hz) said to be implicated in cortical areas of higher visual hierarchy in top-down feed-back processing [40] in the visual system. It thus implies that, beta rhythms could predominate in cortico-subcortical patterns of activation in male mice. The S-peak occurred at the third harmonic, such that the estimated arterial length given by $D_3 = 1/16\lambda$ or $c/16 \times 3f$, or $1.91 \text{ m/s} / (16 * (3 * 8.6 \text{ Hz} \text{ or } 25.8 \text{ Hz})) = 0.0046 \text{ m}$ or 0.46 cm or 4.6 mm, which approximates the visible arterial length from the main stem of the major cortical arteries to the distal arterioles of the ganglionic branches [39]. The 25Hz is the frequency of the rhythm of gamma waves [40]. The ratio of the length of the ganglionic branches to the cortical branches is 1:3 in mice. The same ratio has been found in human subjects [20]. This may suggest that the cerebral vaso-architecture was optimized in mammals to facilitate harmonic oscillations within the cortico-subcortical networks.

4.3 Conclusion

The findings suggest that during color stimulation, the cortical C-peak correlates with 17 Hz beta-band synchronization in the left visual cortex in male mice, and the subcortical S-peak correlates with 25 Hz gamma-band synchronization in the right visual cortex in female mice. These present findings agree with recordings of neuronal cortical beta-band and subcortical gamma-band synchronization in neuronal computation [40], consistent with the input function from GABA modulation of col-

or-opponent bipolar cells in the retina [41]. Furthermore, the present work demonstrated that Fourier time-series analysis is a computational approach that could be implemented in the visual system for light and color processing. Fourier analysis of mean SUV of [^{18}F]FDG have improved the use of $\mu\text{PET}/\text{MRI}$ to index the coupling of rCBF, CMRO_2 and neuronal activity, with a wide range of applications in several areas of cognitive neuroscience, artificial intelligence and quantum mechanics.

5 FIGURE LEGENDS

Figure 1 (A-F) shows the experimental setup (Figure 1A) in close view of the mice Chromatoscope (Figure 1 B), with the animal place within the gantry of the PET/MRI (Figure 1 C). The animals were housed under controlled conditions with free access to food and water (Figure 1D). The heart rate, respiration (Figure 1E) and anesthetic airflow were monitored (Figure 1F).

Figure 2 (A-F) shows the μPET images of radiotracer accumulation in male (Fig. 2A) and female mice (Fig. 2B) and the MRI images in male (Fig. 2C) and female (Fig. 2D) mice. Figure 2E, shows the VOI setup in mouse brain on micro-CT image [39] in coronal view. Figure 2F, shows the schematic diagram of the arterial tree of the Circle of Willis and the relationship to the ROI to the cortical and ganglionic branches of the middle cerebral artery in the mouse brain. The radiotracer distribution shows a ‘male model’ and a ‘female model’.

Figure 3 (A-P) shows the spectral density plots of Fourier coefficients for male, Fig. 3(A-H), and female Fig. 3 (I-P) mice, respectively.

Figure 4 (A-B). The schematic diagram of two separate male (Fig. 4A) and female (Fig. 4B) models for color processing in the mouse brain.

6 TABLES

Table 1A. The SUV values mice obtained during dark condition, white light, blue and yellow light stimulation in time series in male mice.

Table 1B. The SUV values mice obtained during dark condition, white light, blue and yellow light stimulation in time series in female mice.

Table 2. Fourier Spectral Density obtained during Dark condition and stimulation with white light, blue and yellow color in male and female mice, respectively.

Table 3. Fourier Spectral Density (mean \pm SD) values and paired *t*-test results in the right visual cortex through left eye and left visual cortex through right eye, and percent changes (Dark $\Delta\%$) from Dark condition, to demonstrate spatial, luminance and chromatic opponency in male and female mice, respectively.

7 AUTHOR CONTRIBUTIONS

P.C.N. and P.B. designed the concept of the studies. P.C.N. and M.K. acquired and analyzed the PET/MRI data in mice. P.C.N. performed the Fourier analysis. P.C.N. M.K. and P.B. participated in writing the paper.

8 COMPETING INTERESTS

The authors declare no competing financial or non-financial interests.

9 REFERENCES

1. Njemanze, P. C., Kranz, M., Amend, M., Hauser, J., Wehrl, H., Brust, P.: Gender differences in cerebral metabolism for color processing in mice: A PET/MRI Study. *PLoS One*. **12**, e0179919 (2017). <https://doi.org/10.1371/journal.pone.0179919>.
2. Ungerleider, L.G., Haxby, J.V.: 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **4**, 157-165 (1994).
3. Ungerleider, L.G., Mishkin, M.: Analysis of visual behavior. In: Goodale, M.A., Mansfield, R.J.Q. (eds). pp. 549-586. MIT, Cambridge (1982).
4. Perry, C.J., Fallah, M. Feature integration and object representations along the dorsal stream visual hierarchy. *Front Comput Neurosci.* **8**, 84 (2014).

5. McKeefry, D.J., Zeki, S.: The position and topography of the human color centre as revealed by functional magnetic resonance imaging. *Brain*, **120** (Pt 12), 2229-2242 (1997).
6. Gray, H., Clemente, C.D.: *Gray's Anatomy of the Human Body*, 30th American Edition. Lippincott Williams & Wilkins, Philadelphia (1984).
7. Takechi, H., Onoe, H., Shizuno, H., Yoshikawa, E., Sadato, N., Tsukada, H., Y., Watanabe, Y.: Mapping of cortical areas involved in color vision in non-human primates. *Neurosci. Lett.* **230**, 17-20 (1997).
8. Buchsbaum, G., Gottschalk, A.: Trichromacy, opponent colors coding and optimum color information transmission in the retina. *Proc. R. Soc. Lond. B. Biol. Sci.* **220**, 89-113 (1983).
9. Gouras, P. *The Perception of Color: Vision and Dysfunction*. In: P. Gouras, (ed.), pp. 179-197. Macmillan, England (1991).
10. Daw, N.W.: Goldfish retina: organization for simultaneous color contrast. *Science*. 158, 942-944 (1967).
11. Livingstone, M.S., Hubel, D.H.: Anatomy and physiology of a color system in the primate visual cortex. *J. Neurosci.* **4**, 309-356 (1984).
12. Dufort, P.A., Lumsden, C.J.: Color categorization and color constancy in a neural network model of V4. *Biol. Cybern.* **65**, 293-303 (1991).
13. Kraft, J.M., Brainard, D.H.: Mechanisms of color constancy under nearly natural viewing. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 307-312 (1999).
14. Conway, B.R.: Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1). *J. Neurosci.* **21**, 2768-2783 (2001).
15. Zeki, S.: *A vision of the brain*. Plate 16. Blackwell Scientific, Cambridge MA. (1993).
16. Njemanze, P.C., Gomez, C.R., Horenstein, S.: Cerebral lateralization and color perception: a transcranial Doppler study. *Cortex*. **28**, 69-75 (1992). ISI:A1992HK41300005. PMID: [1572174](#)
17. Njemanze, P.C.: Asymmetric neuroplasticity of color processing during head down rest: a functional transcranial Doppler spectroscopy study. *J. Gravit. Physiol.*, **15**, 49–59 (2008).
18. Njemanze, P.C.: Gender-related asymmetric brain vasomotor response to color stimulation: a functional transcranial Doppler spectroscopy study. *Exp. & Transl. Stroke Med.* **2**, e21 (2010). Epub 2010/12/02. <https://doi.org/10.1186/2040-7378-2-21> PMCID: PMC3006356. PMID: [21118547](#).
19. Njemanze, P.C.: Gender-related differences in physiologic color space: a functional transcranial Doppler (fTCD) study. *Exp. & Transl. Stroke Med.* **3**, e1 (2011).
20. Njemanze, P.C.: Cerebral lateralisation for facial processing: gender-related cognitive styles determined using Fourier analysis of mean cerebral blood flow velocity in the middle cerebral arteries. *Laterality*. **12**, 31-49 (2007).
21. Yarowsky, P.J., Ingvar, D.H.: Symposium summary. Neuronal activity and energy metabolism. *Fed. Proc.* **40**, 2353-2362 (1981).
22. Fox, P.T., Raichle, M.E.: Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 1140-1144 (1986).
23. Dahl, A., Lindegaard, K-F., Russell, D., Nyberg-Hansen, R., Rootwelt, K., Sorteberg, W., Nornes, H.: A comparison of transcranial Doppler and cerebral blood flow studies to assess cerebral vasoreactivity. *Stroke*. **23**, 15-19 (1992).
24. Bloomfield, P.: *Fourier analysis of time series: An introduction*. Wiley, New York (1976).
25. Attinger, E.O., Anne, A., McDonald, D.A.: Use of Fourier series for analysis of biological systems. *Biophys. J.* **6**, 291-304 (1966).

26. Njemanze, P.C., Beck, O.J., Gomez, C.R., Horenstein, S.: Fourier analysis of the cerebrovascular system. *Stroke*, **22**, 721-726 (1991).
27. Schuett, S. Bonhoeffer, T. Hübener, M.: Mapping retinotopic structure in mouse visual cortex with optical imaging. *J. Neurosci.* **22**, 6549-6559 (2002).
28. Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., Fries, P.: Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*. **85**, 390-401 (2015).
29. Bliss, T.V., Lomo, T.: Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J Physiol.* **232**, 331-356 (1973).
30. Ito, M.: Long-term depression. *Annu. Rev. Neurosci.* **12**, 85-102 (1989).
31. Hutcheon, B., Yarom, Y.: Resonance, oscillation and the intrinsic frequency preferences of neurons. *Trends Neurosci.* **23**, 216-222 (2000).
32. Jacobs, G.H., Williams, G.A., Cahill, H., Nathans, J.: Emergence of novel color vision in mice engineered to express a human cone photopigment. *Science*, **315**, 1723-1725 (2007). <https://doi.org/10.1126/science.1138838> PMID: 17379811.
33. Thie, J.A.: Understanding the standardized uptake value, its methods, and implications for usage. *J. Nucl. Med.* **45**, 1431-1434 (2004). PMID: 15347707
34. V.S. Ramachandran, in *The Tell Tale Brain*. 500 Fifth Avenue, New York, NY 10110: W. W. Norton & Company, Inc. ISBN 9780393340624. January 17, 2011.
35. DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition?. *Neuron*. **73**, pp. 415-434, 2012.
36. Kravitz, D.J., Saleem, K.S., Baker, C.I., Mishkin, M.: A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* **12**, 217-230 (2011).
37. McDonald, D.A.: *Blood Flow in Arteries*, Edn. 2nd., Williams&Wilkins Co., Baltimore (1974).
38. Di Lascio, N. Stea, F., Kusmic, C., Sicari, R., Faita, F.: Non-invasive assessment of pulse wave velocity in mice by means of ultrasound images. *Atherosclerosis*. **237**, 31-37 (2014).
39. Ghanavati, S., Yu, L.X., Lerch, J.P., Sled, J.G.: A perfusion procedure for imaging of the mouse cerebral vasculature by X-ray micro-CT. *J. Neurosci. Methods*. **221**, 70-77 (2014).
40. Fries, P.: Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu. Rev. Neurosci.* **32**, 209-224 (2009).
41. Zhang, D.Q., Yang, X.L.: GABA modulates color-opponent bipolar cells in carp retina. *Brain Res.* **792**, 319-323 (1998).

Table 1A. The SUV values mice obtained during dark condition, white light, blue and yellow light stimulation in time series in male mice.

Stimulation	Dark	Dark	LightR	LightR	LightL	LightL	BlueR	BlueR	BlueL	BlueL	YellowR	YellowR	YellowL	YellowL
Visual Cortex	visCtxR	visCtxL												

Time	Male Mice													
150	1.48	1.28	1.58	1.81	1.53	1.35	1.41	1.37	1.32	1.29	1.48	1.31	1.16	1.08
150	1.41	1.23	1.12	1.04	0.94	1.03	1.34	1.49	1.22	1.12	1.24	1.32	1.01	0.91
150	1.52	1.55	1.63	1.56	1.67	1.79	1.36	1.52	0.88	1.00	1.54	1.50	1.42	1.46
150	0.99	0.97	0.69	0.72	1.14	1.06	1.58	1.65	1.43	1.46	1.48	1.43	1.19	1.13
150	1.38	1.34	1.05	1.02	1.51	1.41	1.37	1.52	1.21	1.22	1.48	1.55	1.16	1.18
450	1.44	1.20	1.73	1.84	1.43	1.39	1.41	1.50	1.33	1.31	1.50	1.36	1.20	1.18
450	1.35	1.21	1.12	1.08	1.00	1.07	1.39	1.56	1.27	1.28	1.26	1.37	1.07	0.96
450	1.52	1.73	1.71	1.61	1.66	1.75	1.45	1.46	1.03	1.06	1.57	1.62	1.49	1.46
450	1.06	1.07	0.80	0.78	1.13	1.14	1.56	1.63	1.43	1.33	1.47	1.45	1.21	1.27
450	1.49	1.47	1.12	1.15	1.54	1.49	1.56	1.55	1.26	1.24	1.56	1.49	1.29	1.24
750	1.53	1.32	1.79	1.64	1.40	1.41	1.52	1.43	1.35	1.28	1.51	1.25	1.26	1.29
750	1.38	1.21	1.08	1.02	1.05	1.12	1.50	1.48	1.20	1.23	1.27	1.41	1.00	0.93
750	1.45	1.66	1.70	1.58	1.58	1.71	1.50	1.49	1.14	1.17	1.64	1.60	1.50	1.47
750	1.12	1.15	0.78	0.80	1.24	1.24	1.66	1.59	1.51	1.30	1.53	1.48	1.33	1.22
750	1.46	1.45	1.27	1.18	1.50	1.53	1.45	1.68	1.30	1.24	1.46	1.57	1.29	1.32
1050	1.48	1.41	1.70	1.74	1.40	1.33	1.46	1.47	1.40	1.30	1.60	1.38	1.30	1.26
1050	1.29	1.24	1.02	1.02	1.05	1.05	1.32	1.47	1.27	1.25	1.21	1.36	0.99	0.97
1050	1.55	1.61	1.74	1.63	1.62	1.55	1.41	1.49	1.22	1.18	1.57	1.59	1.50	1.50
1050	1.19	1.14	0.88	0.92	1.30	1.35	1.65	1.56	1.44	1.46	1.53	1.44	1.31	1.29
1050	1.44	1.37	1.24	1.30	1.54	1.49	1.48	1.68	1.28	1.25	1.52	1.46	1.40	1.43

Table 1B. The SUV values mice obtained during dark condition, white light, blue and yellow light stimulation in time series in female mice.

Stimulation	Dark	Dark	LightR	LightR	LightL	LightL	BlueR	BlueR	BlueL	BlueL	YellowR	YellowR	YellowL	YellowL
Visual Cortex	visCtxR	visCtxL	visCtxR	visCtxL	visCtxR	visCtxL	visCtxR	visCtxL	visCtxR	visCtxL	visCtxR	visCtxL	visCtxR	visCtxL
Time	Female Mice													
150	1.14	0.99	0.94	0.97	1.13	1.03	1.23	1.26	1.29	1.22	1.21	1.24	1.28	1.21
150	1.31	1.22	1.00	0.92	1.11	1.06	1.44	1.49	1.25	1.30	1.35	1.23	1.32	1.39
150	1.06	1.11	1.15	1.11	1.40	1.36	1.05	1.22	1.41	1.38	1.15	1.00	1.01	1.01
150	1.39	1.37	1.19	1.08	1.58	1.60	1.21	1.24	1.38	1.40	1.16	1.18	1.26	1.28
150	1.21	1.25	1.06	0.99	1.13	1.18	1.34	1.25	1.37	1.18	1.16	1.18	1.23	1.32
450	1.17	1.16	0.95	0.99	1.13	0.98	1.08	1.16	1.29	1.27	1.25	1.27	1.38	1.28
450	1.26	1.28	1.10	1.07	1.15	1.13	1.52	1.40	1.30	1.24	1.40	1.25	1.41	1.41
450	1.03	1.12	1.11	1.14	1.44	1.38	1.14	1.12	1.44	1.43	1.21	0.98	1.08	1.01
450	1.40	1.31	1.19	1.23	1.71	1.75	1.26	1.37	1.40	1.35	1.19	1.23	1.23	1.29
450	1.18	1.17	0.97	1.02	1.06	1.12	1.35	1.35	1.31	1.13	1.19	1.23	1.21	1.29
750	1.11	1.17	0.92	0.98	1.14	0.98	1.03	1.14	1.29	1.28	1.25	1.19	1.30	1.35
750	1.33	1.33	1.16	1.15	1.12	1.11	1.53	1.53	1.27	1.23	1.47	1.23	1.52	1.51
750	1.01	1.10	1.16	1.12	1.44	1.35	1.14	1.22	1.51	1.54	1.16	1.01	1.05	1.13
750	1.46	1.49	1.18	1.24	1.71	1.63	1.34	1.29	1.43	1.38	1.18	1.19	1.27	1.22
750	1.16	1.14	0.93	0.98	1.08	1.19	1.30	1.33	1.26	1.10	1.18	1.19	1.25	1.30

1050	1.17	1.05	0.91	0.98	1.04	1.01	1.04	1.05	1.34	1.34	1.27	1.20	1.28	1.25
1050	1.35	1.32	1.22	1.18	1.19	1.08	1.44	1.39	1.29	1.28	1.41	1.34	1.44	1.53
1050	0.97	1.09	1.13	1.06	1.43	1.39	1.14	1.11	1.44	1.39	1.23	1.08	1.14	1.09
1050	1.37	1.33	1.21	1.23	1.68	1.66	1.29	1.30	1.36	1.36	1.18	1.17	1.24	1.28
1050	1.11	1.14	0.92	0.88	1.03	1.11	1.39	1.35	1.20	1.12	1.18	1.17	1.24	1.20

Table 2. Fourier Spectral Density obtained during Dark condition and stimulation with white light, blue and yellow color in male and female mice, respectively.

Stimulation	Dark	Dark	LightR	LightR	LightL	LightL	BlueR	BlueR	BlueL	BlueL	YellowR	YellowR	YellowL	YellowL
--------------------	------	------	--------	--------	--------	--------	-------	-------	-------	-------	---------	---------	---------	---------

Visual Cortex	visCtxR	visCtxL												
Male Mice														
C-peak	0.005	0.006	0.012	0.014	0.002	0.003	0.014	0.004	0.001	0.002	0.002	0.005	0.002	0.005
	0.012	0.010	0.026	0.033	0.005	0.006	0.015	0.007	0.006	0.008	0.004	0.010	0.008	0.009
	0.038	0.023	0.105	0.141	0.014	0.023	0.021	0.016	0.029	0.022	0.019	0.025	0.033	0.038
	0.068	0.029	0.185	0.255	0.023	0.039	0.030	0.027	0.055	0.030	0.034	0.039	0.055	0.068
	0.043	0.015	0.103	0.148	0.013	0.021	0.017	0.017	0.037	0.016	0.019	0.023	0.031	0.038
S-peak	0.023	0.030	0.044	0.102	0.017	0.037	0.006	0.008	0.022	0.013	0.013	0.010	0.017	0.023
	0.081	0.162	0.170	0.485	0.095	0.224	0.014	0.011	0.061	0.039	0.051	0.021	0.073	0.098
	0.147	0.290	0.312	0.896	0.177	0.420	0.022	0.014	0.102	0.058	0.088	0.035	0.132	0.180
	0.080	0.157	0.169	0.490	0.101	0.239	0.013	0.011	0.058	0.030	0.048	0.020	0.073	0.100
	0.025	0.048	0.051	0.152	0.036	0.083	0.007	0.007	0.021	0.010	0.016	0.007	0.025	0.033
Female Mice														
C-peak	0.002	0.010	0.002	0.010	0.003	0.006	0.003	0.009	0.004	0.002	0.002	0.001	0.003	0.006
	0.004	0.013	0.011	0.009	0.030	0.038	0.009	0.017	0.006	0.008	0.006	0.005	0.009	0.012
	0.007	0.017	0.049	0.025	0.193	0.233	0.014	0.012	0.019	0.037	0.028	0.017	0.034	0.033
	0.009	0.020	0.082	0.044	0.358	0.429	0.010	0.005	0.033	0.066	0.050	0.031	0.055	0.048
	0.005	0.012	0.046	0.025	0.197	0.234	0.006	0.003	0.020	0.038	0.027	0.020	0.030	0.025

Table 3. Fourier Spectral Density (mean \pm SD) values and paired *t*-test results in the right visual cortex through left eye and left visual cortex through right eye, and percent changes (Dark $\Delta\%$) from Dark condition, to demonstrate spatial, luminance and chromatic opponency in male and female mice, respectively.

Stimulation Through (R,L)	Brain Area	Mean \pm SD		$\Delta\%$		<i>p</i> -value		<i>Spatial Opponency</i> <i>p</i> -value	<i>Luminance Opponency</i> <i>p</i> -value			<i>Chromatic Opponency</i> <i>p</i> -value		
		C-peak	S-peak	C-peak	S-peak	C-peak	S-peak	C-peak vs S-peak	C-peak vs C-peak	S-peak vs S-peak	C-peak vs S-peak	C-peak vs C-peak	S-peak vs S-peak	C-peak plus S-peak
Male Mice														
*Dark	visCtxR	0.033 \pm 0.255	0.071 \pm 0.05	-	-	-	-	NS						
*Dark	visCtxL	0.016 \pm 0.009	0.137 \pm 0.140	-	-	-	-	0.05						
LightL	visCtxR	0.011 \pm 0.008	0.085 \pm 0.063	-66.7%	19.7%	0.05	NS	0.05	NS	NS				
LightR	visCtxL	0.118 \pm 0.098	0.425 \pm 0.319	638%	210%	NS	0.05	NS	NS	0.05	0.01			
BlueL	visCtxR	0.026 \pm 0.022	0.053 \pm 0.033	-21%	-25%	0.05	NS	NS				NS	NS	
BlueR	visCtxL	0.014 \pm 0.009	0.01 \pm 0.003	-12.5%	-92.7%	NS	0.05	NS				0.05	NS	0.01
YellowL	visCtxR	0.026 \pm 0.021	0.064 \pm 0.046	-21%	-9.8%	0.05	0.05	NS				NS	NS	
YellowR	visCtxL	0.02 \pm 0.013	0.019 \pm 0.01	25%	-87%	NS	0.05	NS				0.05	NS	0.01
Female Mice														
*Dark	visCtxR	0.005 \pm 0.003	0.075 \pm 0.057	-	-	-	-	0.05						

*Dark	visCtxL	0.014±0.004	0.045±0.032	-	-	-	-	NS						
LightL	visCtxR	0.156±0.144	0.06±0.032	3020%	-18.9%	NS	NS	NS	NS	NS				
LightR	visCtxL	0.023±0.014	0.02±0.01	64%	-55.5%	NS	NS	NS	NS	NS	NS			
BlueL	visCtxR	0.017±0.012	0.008±0.003	220%	-90.5%	0.05	0.05	NS				NS	0.05	0.01
BlueR	visCtxL	0.009±0.005	0.054±0.039	-35.7%	20%	NS	NS	0.05				NS	0.05	
YellowL	visCtxR	0.026±0.02	0.036±0.024	420%	-52%	NS	NS	NS				NS	0.05	0.01
YellowR	visCtxL	0.015±0.012	0.02±0.012	180%	-75%	NS	0.05	NS				NS	0.05	

*Dark in both eyes.

Correspondence between the NIST tests, bifurcation diagrams and Lyapunov exponents for the evaluation of chaos-based pRNGs

Octaviana Datcu and Radu Hobincu

University POLITEHNICA of Bucharest, Faculty of Electronics, Telecommunications
and Information Technology,
Bucharest, Romania
{octaviana.datcu,radu.hobincu}@upb.ro
<http://www.electronica.pub.ro>

Abstract. This investigation reveals how some chaos theory tools - the Lyapunov largest exponent and the bifurcation diagram - can predict the seed-space available for chaos-based pRNGs. A discrete three dimensional Hénon map system was used in a previous work to implement a pseudo-random number generator. While evaluating the randomness of the resulted pRNG, the authors observed a close correspondence between the results of the NIST test suite and the above mentioned chaos-specific metrics. The present paper gives a sample of such an analysis, when only one of the component of the seed is varied while the other four are kept fixed. Numerical results and graphics sustain the correlation between the NIST tests results for randomness and the (non)chaotic behavior proven for the underlying system by the bifurcation diagrams and the Lyapunov exponents.

Keywords: chaotic Hénon map, pRNG, NIST tests, Lyapunov exponents, bifurcation diagrams

1 Introduction

The generalized three dimensional Hénon map [1] in (1) was used to propose a chaos-based pseudo random number generator (pRNG) in [2].

$$\begin{aligned}x^+ &= a - y^2 - bz \\y^+ &= x \\z^+ &= y\end{aligned}\tag{1}$$

where the values of the parameters a and b will be discussed later in this paper. The symbol $+$ stands for the future value of the state, *i.e.* $x^+ = x[k + 1]$, with k the discrete time.

The last significant byte of each of the three states is used in a bitwise xor operation. For example, for $(a, b) = (1.9, 0.03)$ and initial values (0.814723686393179,

0.905791937075619, 0.126986816293506), the value for each time series at iteration $k = 1000$ is given by $x[k] = 1.118346703158675$, $y[k] = -1.821411758554415$, $z[k] = 0.850893376791295$, with the 64-bits binary representations equivalent to the floating-point above-mentioned values:

```
00111111 11110001 11100100 10111111 10000011 00111010 01111110 10101111
10111111 11111101 00100100 10000000 10100111 11111000 10101000 00101111
00111111 11101011 00111010 10000100 10111111 00110110 01110011 10100110
```

Adding modulo 2 without carry the least significant byte of each of the three values results in 00100110.

The seed-space of the pRNG is given by the tuple $S = \{a, b, x[0], y[0], z[0]\}$. Randomness for the numbers generated was proven for some values of the seed, using NIST tests in [2].

2 Strange attractors' evaluation specific tools for pseudo-randomness (chaoticity)

2.1 Bifurcation diagrams

Bifurcation diagrams are a specific tool for analyzing the evolution of chaotic systems [5]. The values of $x[k]$, when plotting the last $5 \cdot 10^3$ values out of 10^4 iterations, for some meaningful fixed values for a , and varying b with step 10^{-2} are shown in Fig. 1. The initial conditions are irrelevant for this analysis, as long as they belong to $(-2, 2)$. Nevertheless, here, the initial conditions vector was the one mentioned in Section 1. The implementation is done in Matlab, the code being available at <http://gitlab.dcae.pub.ro/research/chaos/HenonMapPRNG.git>, together with other software used in this paper. The top left image describes the behavior for $a=0.15$, where one can see that the system has a unique solution for a large interval of values, $b \in (-0.76, 0.9)$. The use of these pairs (a, b) , for encryption is not recommended, the Hénon map being stable. The top right diagram shows the richest behaviour of the map, for $a = 1.4$, where stable solutions alternate with chaotic, or even hyperchaotic behaviour. The bottom left picture sustains the frequent use of the value $a = 1.76$, almost (except $(a, b) = (1.76, 0)$) each pair giving rise to an extremely large number of solutions for the Hénon map (1), thus the system being (hyper)chaotic. The bottom right case is a situation to be considered too, because of the pair $(1.8, 0.01)$ which generates an apparently chaotic behaviour for the first iterations, only to settle down to three values as it can be seen in Fig. 2.

2.2 Lyapunov exponents

The Lyapunov exponents are usually computed to estimate the local predictability around a point (x, y, z) in the state space of the considered system. They are the eigenvalues of the Jacobian matrix $J(x, y, z)$. The algorithm used by this work to calculate these values [6] sorts them from the largest to the smallest.

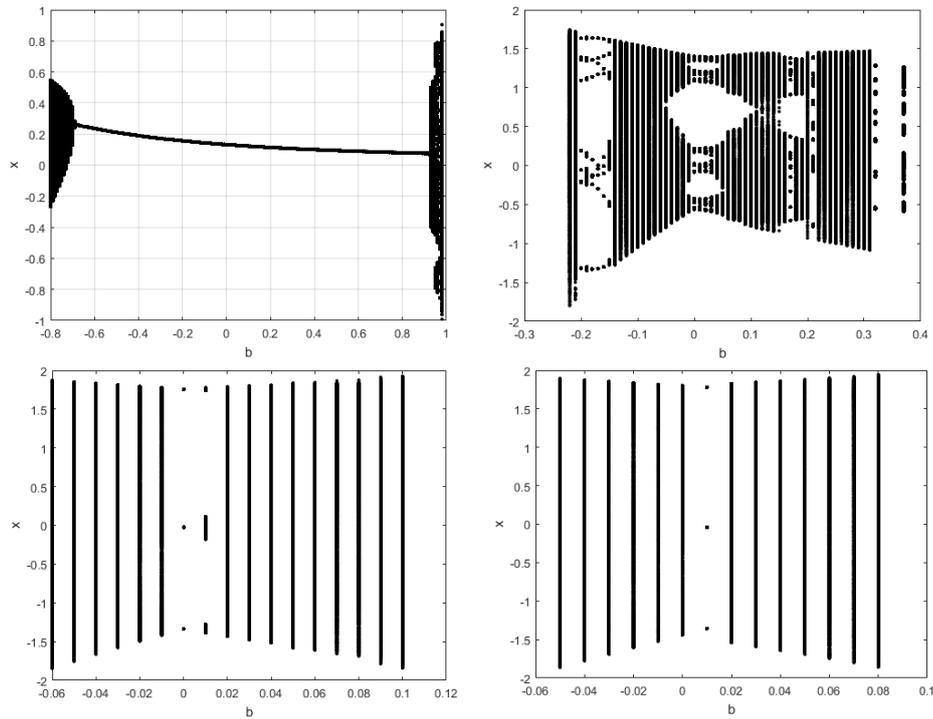


Fig. 1. Bifurcation diagrams for the parameter a kept fixed at 0.15 (top, left), 1.4 (top, right), 1.76 (bottom, left) or 1.8 (bottom, right) and varying b .

For our system, $\lambda_1, \lambda_2, \lambda_3$ are graphically represented in Fig. 3 and Fig. 4 for $a \in \{0.15, 1.4, 1.76, 1.8\}$ and varying b with step of 0.01. Table 1 gives the numerical values for the three Lyapunov exponents for $b = 0.1$ kept fixed and varying $a \in [1.35, 1.45]$, with step 0.01. We observe λ_1 for each pair. A negative value of the exponent indicates stability of the system, while a positive value indicates a so-called chaotic behavior, a long-term unpredictable evolution. When also λ_2 is positive, the system has two unstable directions, it is even more unpredictable, dynamics known as hyper-chaoticity [3]. A chaotic or a hyperchaotic system generates seeds suitable for the pRNG.

3 NIST test suite for randomness

The NIST randomness tests [4] are another metric to test the quality of the proposed pRNG. The C code for the NIST tests can be found at [7]. There are fifteen statistical tests which are thoroughly described in [4] and [7]. The present paper gives the results when running the NIST test suite for fixed parameter $b = 0.1$ and varying a in Table 1.

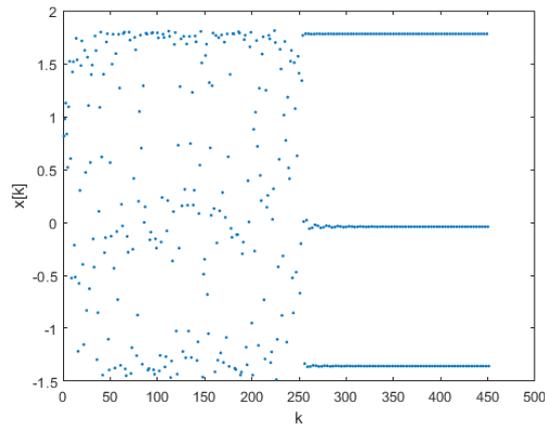


Fig. 2. Temporal evolution of the system for $(a, b) = (1.8; 0.01)$ and randomly chosen $x[0] = 0.814723686393179$, $y[0] = 0.905791937075619$, $z[0] = 0.126986816293506$.

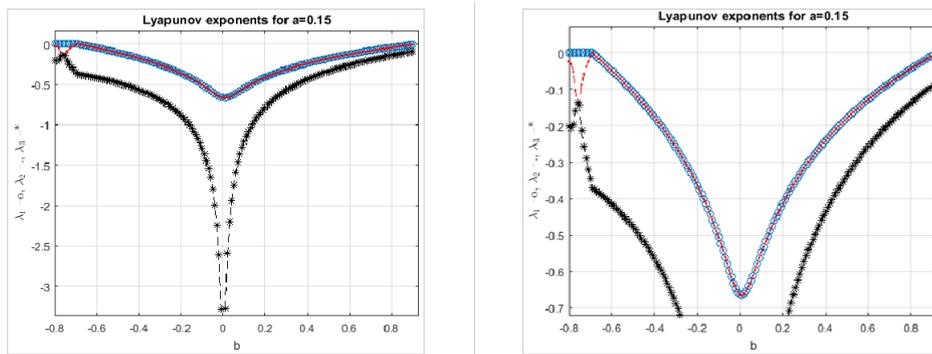


Fig. 3. Lyapunov exponents for fixed parameter a and varying b and zoom (right) on the two largest exponents.

4 Case study

The proposed pRNG is evaluated for fixed $b = 0.1$, $x[0] = 0.814723686393179$, $y[0] = 0.905791937075619$; $z[0] = 0.126986816293506$ and varying a . The same randomness of the dynamics of the Hénon map is revealed by the NIST tests applied to the bytes obtained by the summing modulo 2 without carry of the generated states as well as the Lyapunov exponents and the bifurcation diagram. The pairs for which the number of failed tests is close to 0 are suitable for encryption.

The range $a \in [1.35; 1.45]$ was chosen for investigation in table 1. It is obvious there is a correlation between the number of failed test and the signum of the

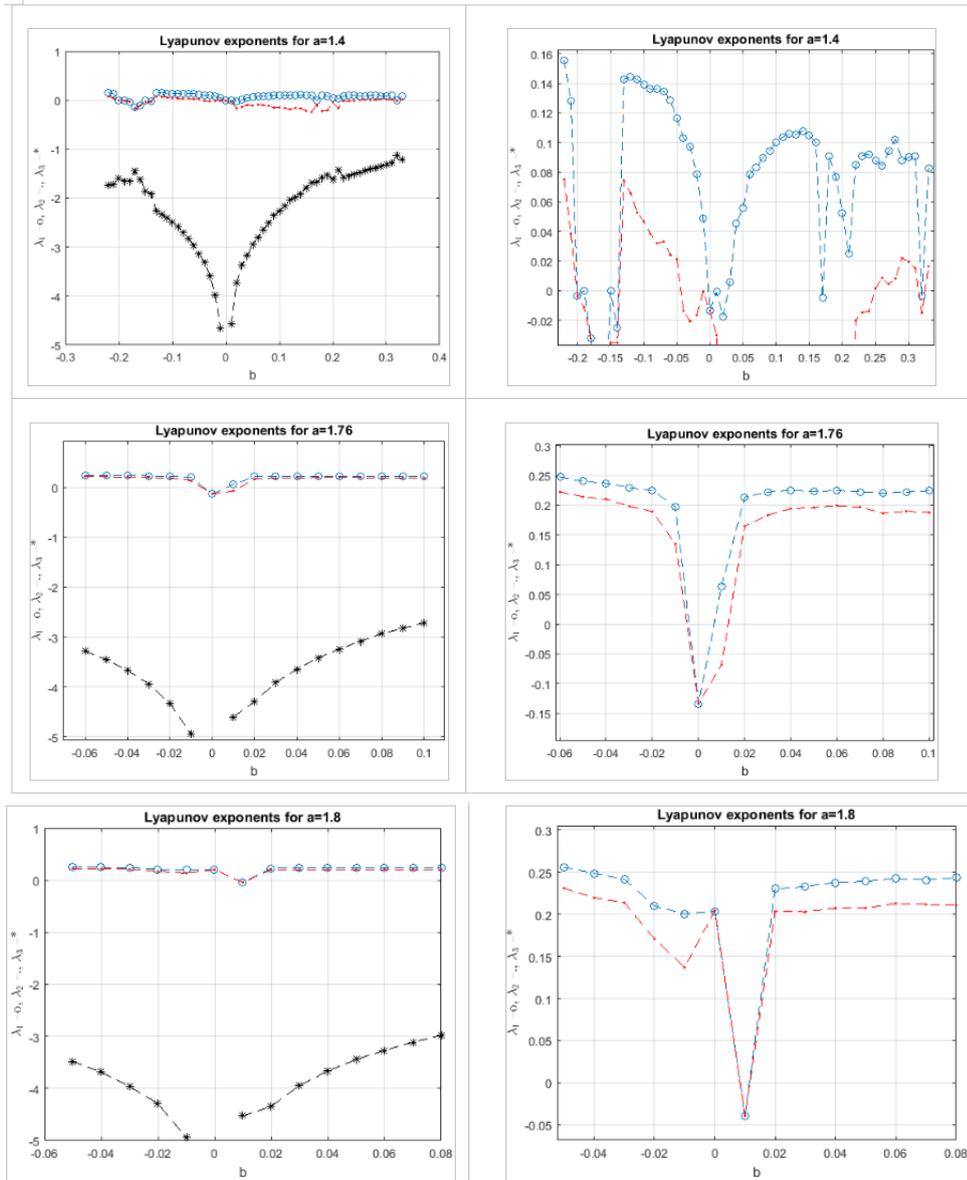


Fig. 4. Lyapunov exponents for fixed parameter a and varying b and zoom (right) on the two largest exponents.

greatest Lyapunov exponent. The full table can be found at <http://gitlab.dcae.pub.ro/research/chaos/HenonMapPRNG.git>.

Table 1: Parallel between the Lyapunov exponents of the 3D Hénon Map and the number of failed NIST tests (out of 162 [4]).

a	λ_1	λ_2	λ_3	No. of failed NIST
1.35	-0.0676145077	-0.1662534673	-2.068717118	158
1.3501	-0.0659004988	-0.1682743943	-2.0684101999	162
...
1.3597	-0.0193103016	-0.3314144084	-1.951860383	162
1.3598	-0.0212302902	-0.3085722679	-1.972782535	125
1.3599	0.0232750774	-0.2940394427	-1.9852705729	0
...
hline 1.3603	0.0332313676	-0.2575153642	-2.0118383612	0
1.3604	-0.0363284566	-0.2501730785	-2.0160835579	160
1.3605	0.0397667208	-0.2429825709	-2.0198358012	0
1.3606	0.043654664	-0.2357105453	-2.0232198836	0
1.3607	-0.0480871492	-0.2282225086	-2.0262754351	161
...
1.3657	-0.0054165485	-0.2216907723	-2.0754777722	134
1.3658	-0.0067873953	-0.2200592435	-2.0757384542	111
1.3659	0.0047640361	-0.231546584	-2.0758025452	0
1.366	0.0058567449	-0.2317957684	-2.0766460695	0
1.3661	0.0086015009	-0.2334653236	-2.0777212703	1
...
1.3673	0.0215470675	-0.2919010521	-2.0322311084	0
1.3674	0.0217868257	-0.2888364904	-2.0355354283	0
1.3675	-0.0221574555	-0.2846550147	-2.0400875338	127
1.3676	0.0209173687	-0.2870543433	-2.0364481184	0
1.3677	-0.0226785132	-0.2765923859	-2.0486712203	122
1.3678	0.0229810516	-0.2748081302	-2.0507580144	0
1.3679	0.0240589897	-0.2808801748	-2.0457639079	1
...
1.4038	0.1027297759	-0.1388422687	-2.2664726002	0
1.4039	0.1019081097	-0.1392937935	-2.2651994091	0
1.404	-0.0026220373	-0.1182015184	-2.1817615373	136
1.4041	0.1028707067	-0.137832284	-2.2676235157	0
1.4042	0.1033055646	-0.1375011018	-2.2683895557	0
...
1.4111	0.105487349	-0.1215661286	-2.2865063134	0
1.4112	0.1056496251	-0.1214847932	-2.2867499248	0
1.4113	-0.0521167591	-0.0835752972	-2.1668930367	129
1.4114	0.1047319211	-0.1216269874	-2.2856900267	0
1.4115	0.104764906	-0.1217029356	-2.2856470634	0
...
1.4186	0.0968266246	-0.116457645	-2.2829540726	1

Table 1 continued from previous page

a	λ_1	λ_2	λ_3	No. of failed NIST
1.4187	0.0961054683	-0.1155501673	-2.283140394	0
1.4188	-0.0060027634	-0.0513761481	-2.2452061815	161
1.4189	0.0913510397	-0.1127120699	-2.2812240627	0
1.419	0.0936529906	-0.112884383	-2.2833537006	0
...
1.4379	0.1267784975	-0.0638727714	-2.3654908191	0
1.438	0.1265972388	-0.0636485545	-2.3655337772	0
1.4381	-0.0125720176	-0.0506843259	-2.2393287495	161
1.4382	-0.010125316	-0.0204739992	-2.2719857779	121
1.4383	0.1037222672	-0.0537626287	-2.3525447315	0
1.4384	0.1266121703	-0.0623409312	-2.3668563321	0
...
1.4492	0.1345145322	-0.0449513758	-2.3921482494	0
1.4493	0.1349026788	-0.0445188058	-2.3929689661	0
1.4494	0.0082976594	-0.0043917833	-2.3064909691	160
1.4495	0.1335393672	-0.0432459886	-2.3928784716	0
...
1.45	0.1337718289	-0.0420068804	-2.3943500415	0

5 Conclusions

This investigations shows the possibility of using chaos-specific metrics in order to evaluate chaos-based pseudo random number generators instead of using time consuming test batteries like NIST.

The presented analysis needs to be detailed by changing all seed components $(a, b, x[0], y[0], z[0])$, and decreasing the sweep step. Nevertheless, we used the pseudo-random numbers generated by the pRNG from [2] in a secret communication scheme. A bitwise xor was performed between the intensity of each pixel (a byte) and such a pseudo-random number, resulting in a stream cipher. The original image and its histogram are given in Fig. 5.

By observing the bifurcation diagram in Fig. 1 for $a = 1.76$, one can conclude that the dynamics converges to three points only. For $b = 0.01$ the evolution converges to three regions of the attractor.

The Lyapunov exponents in Fig. 4 are all three negative for $b = 0$: $\lambda_1 = -0.134188912709540$, $\lambda_2 = -0.134175939778293$, $\lambda_3 = -36.572538874864669$. For $b = 0.01$ the largest Lyapunov exponent is positive attesting the chaoticity of the system ($\lambda_1 = 0.062811122373326$, $\lambda_2 = -0.068389807227143$, $\lambda_3 = -4.599591501134266$).

The number of failed tests for $(a, b) = (1.76, 0)$ is 162, while for $(a, b) = (1.76, 0.01)$ all NIST tests are passed. Although the Hénon map folds to three regions of the attractor only, the behavior is complex, as Fig. 6 and Fig. 7 show.

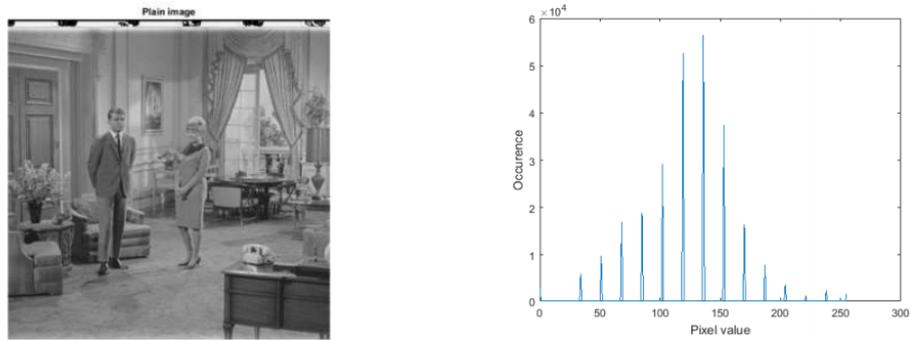


Fig. 5. The image to be encrypted and its histogram.

The results of the encryption in Fig. 8 and Fig. 9 highlight once again the lack of randomness induced by the seed $(a, b) = (1.76, 0)$ and the uniformity of the encrypted pixels when using $(a, b) = (1.76, 0.01)$.

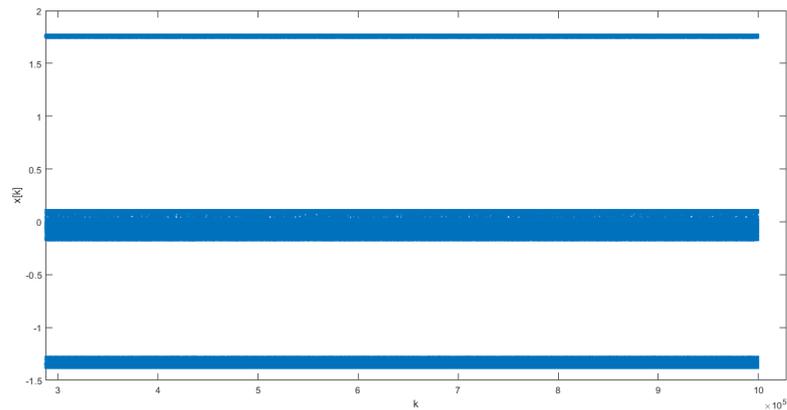


Fig. 6. Temporal evolution of $x[k]$ for parameters $(a, b) = (1.76, 0.01)$.

As future work, the authors intend to propose an automatic system that can generate valid seeds so that the proposed pRNG can have real-world applications.

Acknowledgment

This work has been funded by University Politehnica of Bucharest, through the Excellence Research Grants Program, UPB GEX 2017. Identifier: UPB-GEX2017, Ctr. No. 09-17-06/2017 (36/25.09.2017; ID 118).

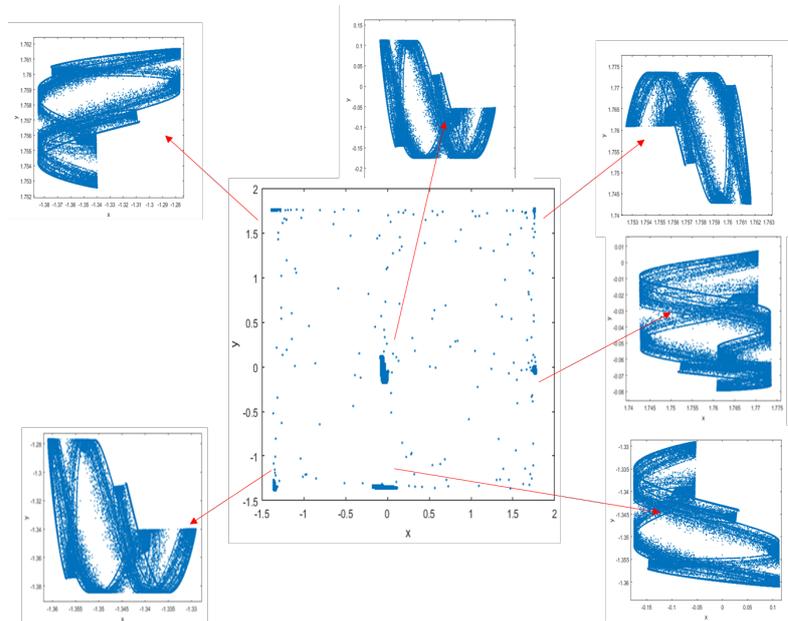


Fig. 7. The complexity of the 3D Hénon map attractor for parameters $(a,b)=(1.76,0.01)$.

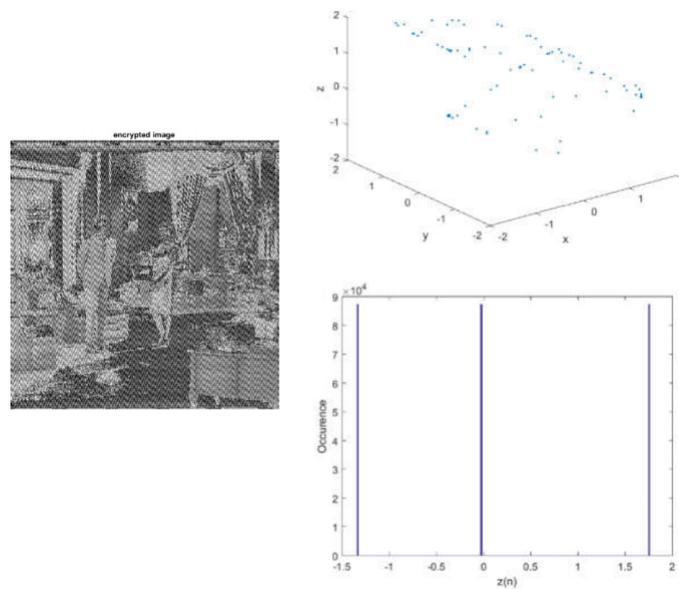


Fig. 8. The encrypted image and the dynamics of the Hénon map for parameters $(a, b) = (1.76, 0)$.

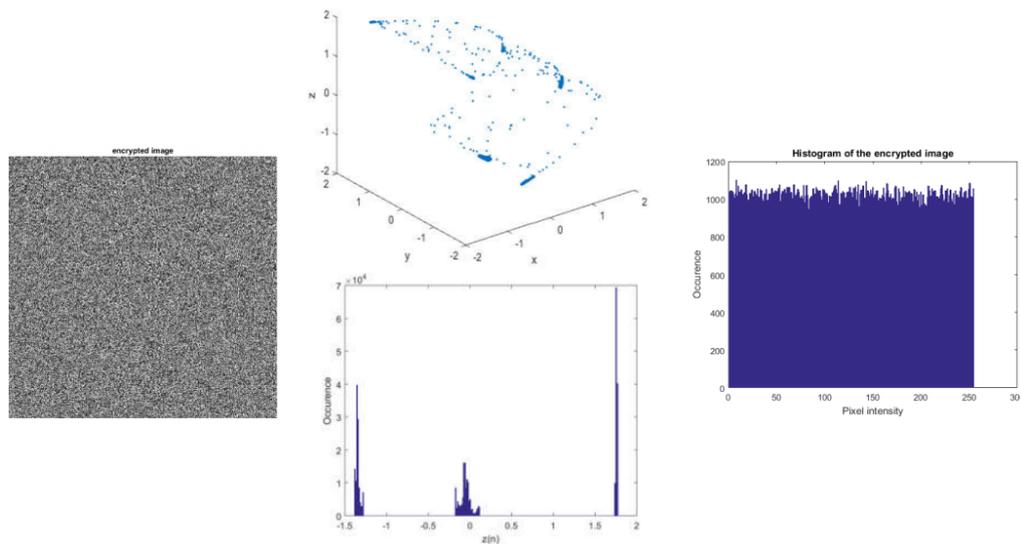


Fig. 9. The encrypted image, the dynamics of the Hénon map for the considered key, and the histogram of the encrypted image for parameters $(a, b) = (1.76, 0.01)$.

References

1. G. Grassi and D. A. Miller, "Theory and Experimental Realization of Observer-Based Discrete-Time Hyperchaos Synchronization", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no 3. 2002.
2. R. Hobincu, O. Datcu, "A novel Chaos Based PRNG Targeting Secret Communication", *COMM 2018*, Bucharest, Romania, June 2018.
3. D. A. Miller & G. Grassi, "A discrete generalized hyperchaotic Hénon map circuit", *Midwest Symposium on Circuits and Systems*, Vol. 1, pp. 328-331, 2001.
4. A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Application", *SanVo*, Revised: April 2010 Lawrence E Bassham II.
5. S. Strogatz, "Non-linear Dynamics and Chaos: With applications to Physics, Biology, Chemistry and Engineering", Perseus Books, 2000.
6. A. Wolf, J. B. Swift, H. L. Swinney, J. A. Vastano, "Determining Lyapunov exponents from a time series", *Physica D: Nonlinear Phenomena*, Vol. 16, Issue 3, pp. 285-317, 1985, ISSN 0167-2789, DOI: 10.1016/0167-2789(85)90011-9.
7. NIST http://csrc.nist.gov/groups/ST/toolkit/rng/documentation_software.html, accessed on 15th of June 2017.

Risk Assessment Approach to Support IT Collaboration Network

D.CHIKHAOUI ¹, MS. BENQATLA ², B. BOUNABAT ³

^{1,2} ALQUASADI TEAM ENSIAS,
ADMIR LABORATORY,
RABAT IT CENTER, MOHAMMED V University,
RABAT, Morocco

dikra.chikhaoui@um5s.net.ma, b.bounabat@um5s.net.ma

Abstract. This paper addresses an IT Governance improvement approach based on Risk Assessment in Network Collaboration context. Designing methodologies that allow IT risk analysis in changing and heterogeneous business climate is increasingly critical in order to ensure effective governance. In this perspective, we investigate the contribution of Social Network Analysis (SNA) to conduct a quantitative risk assessment and IT Collaboration improvement opportunities, by studying actors and relationships. SNA is an area which has been researched in many different disciplines, we examine it in our approach, focusing on IT Governance suppliers' selection process.

Keywords: risk assessment, IT governance, collaboration network, social network analysis, suppliers management.

1 Introduction

Business major goal is to create value for shareholders as well as for stakeholders. This goal is perceived as realizing benefits at optimal resources cost while optimizing risks [3]. By collaborating together, organizations can achieve demanding and complex finalities they cannot attain on their own [1].

Effective IT Governance supports generating real business benefits such as enhanced reputation, trust, product leadership, and reduced costs [2]. However, managing the 21st Century IT organization is a highly complex task. Information Technology has become deeply entrenched as a critical dependency for many internal and external actors. Effective governance of the broader context (of Consultants, Contractors, Sub-Contractors and Suppliers) has become increasingly critical to strategic business success.

On the other hand, almost every business decision involves executives and managers to strike the balance between risk and benefits. Effectively managing the business risks is essential to an enterprise's success. Risk Management is becoming increasingly important in various domains. Most project management disciplines prescribe it as a best practice. Even in new types of contracts and in innovative, collaborative forms of

organization and management, we still perceive risk management failing to deal with the dynamics business climate changing.

Consequently, it is necessary to use appropriate techniques that allow a diagnosis as exhaustive as possible, to detect risks in just appropriate time, depending on its type, and accordingly advance treatment options that minimize the impact. In this paper is investigated how through the application of social network theory is possible to identify and quantify existing or potential risk among several IT actors in a collaborative network context. And to explore the role of quantitative risk assessment in the success of suppliers selection.

The structure of this paper is as follows. Section 2 presents the relevant concepts. Section 3 introduces brief literature background of Collaboration Network and Risk Assessment using SNA. Section 4 and 5, present the proposed risk-based approach to support IT Collaboration though exploring its application on the suppliers' management process of IT Governance. Conclusions and extensions of the research work are addressed in section 6.

2 Concepts

2.1 IT Governance in Collaboration Network Context

During the past two decades, researchers have studied network management and the associated value creation process; they have attested to the innovation strength of networks and their ability to react flexibly to changing conditions [6]. Studies of IT Governance and Collaboration Network have not cohered around a common theoretic framework, and many lack grounding in an established theoretical tradition. Some focus on network structures and its effectiveness in general way including IT[7]; others on network management and coordination processes[8]; still others on institutional factors such as coalition membership, resources, and decision making[9]. Several recent analyses construct their own frameworks using a blending of constructs. But there is little research examining IT Governance from a Collaboration Network viewpoint.

Research on IT governance in Network context has been attracting more and more attention, as the importance of IT in increasing business performance. Researchers have attested to the innovation strength of networks and their ability to react flexibly to changing conditions [6]. this area, however, remained limited [21]. Although, it only recently that it becomes fashionable.

2.2 SNA, Concepts & Indicators

2.2.1 SNA: Concepts

The volume of research related to social networks is growing exponentially within the management and organizational sciences [10]. Actually, SNA has been researched in many different fields, it uses graphical and mathematical methods to analyze the social structures of networks (Scott, 2000), is above all a toolbox for visualizing and modeling social relations as nodes (Actors, individuals, organizations ...) and links (relationships

between these nodes), using results from sociometry, mathematics and anthropology. This structural approach which focuses specifically on the description and analysis of the different possible modes of relationship: interdependence, centrality, Holes, frequency... The strength of this structural analysis lies in its ability to represent in a simplified way the complexity and diversity of relations between actors.

SNA has already proven useful in economic, political, social, professional, and medical contexts. In contrast to these other fields, IT governance has only recently begun to be addressed using SNA.

2.2.2 SNA: principal indicators:

Network metrics can be calculated at two levels—the node level and -the network level. Node-level network metrics reveal how a node is implanted in a network from that individual node’s perspective. Freeman (1979) provides a measure of the prominence of a given actor within a network, it’s defined what is called centrality, in different aspects. Most important are:

Degree centrality as defined by Freeman, provides a measure of communication activity, when a node is connected to a large number of other nodes, the node has high degree centrality. This metric was adopted in favour of the other two types of centrality identified by Freeman:

Betweenness: The extent to which an actor acts as a posits “broker” or intermediary “gatekeeper” in the network, means that other nodes are dependent on this node to reach out to the rest of the network.)

Closeness: this metric focuses on how close a node is to all the other nodes in the network. An actor is “close” when it has the shortest paths to all others, he is central if it can quickly reach all the others.

Network level metrics compute how the overall network ties are organized:

Network density is the number of total ties in a network compared with the number of potential ties. It is measure of the overall connectedness of a network (Scott, 2000)—density of one means that all nodes of the network are connected with each other.

Network centralization captures the extent to which the overall connectedness is organized around particular nodes in a network (Provan and Milward, 1995). Highest possible centralization is of a network with a star structure. Network with lowest centralization is when all nodes have the same number of connections.

Network complexity is defined as “the number of dependency relations within a network”; it depends on the number of nodes in the network and the degree to which they are interlinked [13].

2.3 Risk Management.

IT Governance frameworks like COBIT 5 cover the general concept of the IT Risk through specific processes, one of which focuses on stakeholder risk-related objectives: EDM03 Ensure risk optimization. Risk management frameworks and standards (Risk IT ISACA, ISO 31000, ISO 27005...) are designed to identify, analyze, mitigate, manage, monitor and communicate IT-related business risks and determine how to reduce,

manage and supervise them. IT risk management is in alignment with enterprise risk management [20], and involve four key steps : Context establishment Risk identification ; Risk assessment ; Risk treatment; Risk management plan evaluation/review :

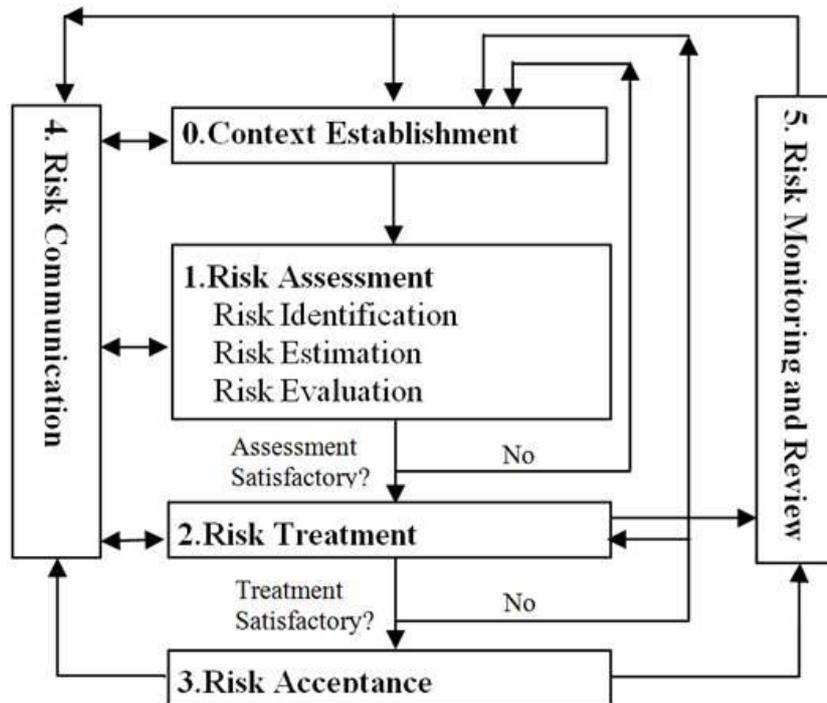


Fig. 1. Guidelines for information security risk management by ISO/IEC 27005.

The literature distinguishes two risk assessment approaches: Qualitative risk analysis consists of assessing identified risks using a pre-defined rating scale, risks will be scored based on their probability or likelihood of occurring and the impact on project objectives should they occur; and Quantitative risk by analyzing real data and exploit data bases to provide a quantitative approach to making decisions.

Qualitative risk analysis should generally be performed on all risks, quantitative risk depends on the project risks nature, and the availability of data to use to conduct the quantitative analysis.

3 LITERATURE REVIEW of Collaboration Network and Risk Assessment using SNA

The analysis of collaboration network is far from new. Borgatti et al (2009) cite the development of Moreno's sociometry in the 1930 as the precursor of SNA. In the 1950s, experimental studies of networks aimed to show that centralized decision making was more effective than decentralized [22]. They studied network using relational algebra used in SNA and research of causal power. In the mid-1970s Granovetter (1973) advanced an influential theory of the strength of weak ties. SNA is often used to study leaders and could in theory tell us much about the relational dimension, but it wasn't used explicitly to analyze IT Risk within IT Governance Network.

There are several methods in Literature that can be used to analyze risk [11] [15] [18]. However, none of them were address directly risk based approach to support the IT governance in a collaborative Network. In order to address this problem, the issue of risk analysis in collaborative network seems to be a useful instrument for IT governance and sustainability of the collaborative network.

On the other hand, Collaboration Network and SNA was explored in researches to view networks as analytical tools that encompasses and explains relationship, hierarchies as variations of network structures, but no study combined this too theories to deal with risk-based approach to support IT Collaboration Network problematics.

4 Risk assessment to support IT Collaboration

The successful implementation of IT governance and risk management implies the participation of all involved in the process from the management team of the company to the operational level.

In the Collaboration Network perspective, effective risk management program requires internals and externals partners consideration and cooperation, that involves all levels of network collaboration in conducting risk assessments and formalizing the treatment plans. Focusing on the IT Collaboration's strategic objectives, risk management should be acknowledged as being directly connected to the development of strategic project and business plans. Some critical strategic decisions may impact on the Collaboration's objectives and visibly the significance of such events require a vigilant appreciation of the associated risks. Collaborative Risk Management supports decision making processes internally and externally, at a governance level management and operational level to achieve a joint consensus on results.

Therefore, to highlight the importance of this risk-based approach to place an effective collaborative IT governance structures, processes, operations and relational mechanisms, and to increase our understanding of the risk factors influencing IT governance networks. We propose in our approach some tools, especially social analysis mechanism, to conduct qualitative risk-assessment process.

SNA is defined as any object, individual or group of objects or individuals that are involved in the network of Relations, which is well-suited with the basics of collaboration as a network. The motivation to use SNA is that is a tool which can be used just before the construction, during and right after the construction of the collaboration network, so permits to assess, reduce and monitor risks.

5 Quantitative Risk assessment approach using SNA applied to suppliers' selection process

5.1 Suppliers management

Suppliers are an extension of organization's operations and so they must reflect their values and work towards their objectives. An effective IT Governance includes management of integrated technology suppliers. For example, according to COBIT 5 [14], the organization should have a fair and formal practice of suppliers' selection to ensure a viable best fit based on specified requirements, it should identify evaluation criteria and evaluate the overall portfolio of existing and alternatives suppliers.

In our use case applied to Collaboration Network within IT Governance frame, particularly on the process of supplier selection, SNA is used to enable better suppliers' selection, as a tool for managing risk and value creation. In fact; a system of interconnected buyers and suppliers is better modeled as a network than as a linear chain. Suppliers selection process is the investigative process by which a company or other third party is reviewed to determine its suitability for a given task.

According to a Gartner study [16] IT organization use four categories of vendors. Hardware vendors sell or lease physical products and sell related support such as repair and maintenance. Software vendors sell commercial off the shelf (COTS) products, or offer licenses to an existing product line. Service vendors which can be either long term, such as an outsourcing annuity-based contract, or short term, such as staff augmentation services used to supplement enterprise personnel. Telecommunications vendors supply telephones, networks, circuits and network services. We use this categorization to classify the suppliers network.

5.2 SNA to improve suppliers' management within IT Collaboration Network

An organization who collaborates with several partners is looking to strengthen the IT governance of its information system. For this purpose, the management decides to select a contractor to conduct an external IT security control audit. Supplier management process should help to select the appropriate supplier to perform this task.

The principal objective fixed for this project is the objectivity of the audit deliverables. Risk is defined as the possibility of an event occurring that will have an impact on the achievement of the objective aimed by this IT Security audit. the risk of violating the objectivity of the audit is closely linked to the chosen contractor. In fact, we face in

this case the Critical Employee Risk – related to supplier who may have critical information or resources and he is assigned to a critical task.

The management decide to exclude the supplier who may not drive objectively the IT Security audit.

The main question addressed to analyze suppliers is: Which suppliers among the tenderers are involved in IT collaboration network and what are their collaborative relationships?

To answer this question the following matrix is addressed:

Supplier / Materials and services	S1	S2	S3	S4
Hardware	1	0	0	1
Software	0	1	0	0
Services	1	1	1	0
Telecom	0	0	0	0

Table 1. T1- incidence matrix indicating the intervention area of each supplier

The second question to answer is: What is the level of integration of each supplier in IT Governance network?

	S1	S2	S3	S4
S1	-	1	1	1
S2	1	-	1	0
S3	1	1	-	0
S4	1	0	0	-

Table 2. T2- Adjacency suppliers' matrix derived from T1

Then the degree of each supplier is calculated to determine its level of integration. In fact, to ensure audit objectivity, the organization should choose the least involved supplier in its information system.

Degree is calculated by the following formula [17]:

$$D_a = \left(\frac{\sum_1^N d(a,j)}{N-1} \right)$$

Where D_a is the centrality degree of the actor a ; $D(a,j)$ represents the value of a relation from the a actor directed to the j th actor in the network. N is the number of actors in the network.

	Degree	standarized degree
S1	3	1
S2	2	2/3
S3	2	2/3
S4	1	1/3

Table 3. T3- Suppliers' Degree

In T3, the degree of S4 is “1” because Actor S4 has only one direct contact with another actor, Actor S1. The degree of an actor can be interpreted as the “point centrality” of the actor. A point is central then, if it has a high degree. The actor can be said to be “well-connected,”.

By dividing the measure of degree with 3 (the total number of other actors 4-1), a relative measure called the “standardized degree” of an actor measures how connected a supplier relative to others.

S4 has the smallest centrality degree (1/3), it means that it is the least implicated supplier in the IT collaboration network of this organization, therefore in the best position to conduct the external audit mission objectively.

6 Conclusion

In business today, risk plays a critical role. IT Governance supports organization to execute new IT projects in a controlled manner to deliver value for the business while minimizing the risk of change. Managing risk is a part of several IT Governance frameworks and standards. Organizations seek for good governance, but then again to achieve it in a collaborative context, more systematic and quantitative analysis for risks is required. Reaching a better characterization and understanding of the potential risks in

collaborative processes is an important pre-condition to avoid project failure. Through increasing partners integration, companies have attempted to manage this enlarged level of complexity. IT Service vendors integration has been identified as a key practice to achieve superior IT Governance performance. Given the risk exposure and costs involved, quantitative risk assessment can be a cost-effective methodology that a company can implement and which may contribute in many significant ways.

From the beginning, this paper has highlighted how the social network aspect of information technology -- interactions among actors-- can strongly influence IT Collaboration decision. This has led to apply network theories, particularly, Social Network Analysis (SNA) to IT governance. We have investigated SNA concepts especially, degree centralities to answer questions that help designing, implementing, assessing and improving IT Governance suppliers' selection process.

This research is expected to help organizations in a number of ways. It provides insights that steering committees can use to establish an effective collaborative IT Governance, we gave supplier management process example. As not all supplier relationships should be the same, it is necessary to establish the level of importance of individual suppliers not just in terms of the relationship but also the risk, the criticality of the contract as illustrated in the application case of the audit mission.

REFERENCES

1. Actor network theory a framework of IT collaboration 2017 DOI : 0.1109/WINCOM.2017.8238192
2. Enhancing IT governance practices: A model and case study of an organization's efforts Paul L. Bowen a, May-Yin Decca Cheung b, Fiona H. Rohde b, doi: 10.1016/j.accinf.2007.07.00
3. Geoff Harmer 2014, Governance of Enterprise IT based on COBIT 5: A Management Guide
4. e.g., Powell and al., 1996; Möller, 2006
5. rovan, K.G. & Milward, H.B. (1995). A preliminary theory of interorganizational network effectiveness: A comparative study of four community mental health systems. *Administrative Science Quarterly*, 40, 1-33
6. McGuire 2002; Goldsmith and Eggers 2004; Herranz 2007
7. e.g., Stone et al. 2001; Callahan 2007

8. Borgatti, S.P. & Foster, P.C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*, 29, 991–1013.
9. Edwards, Peter, and Paul Bowen (2013). *Risk management in project organisations*. Routledge.
10. Frenken, 2000 p. 260; Kauffman, 1993
11. *A Business Framework for the Governance and Management of Enterprise IT*, ISACA.
12. Kendrick, T. (2015). *Identifying and managing project risk: essential tools for failure-proofing your project*. AMACOM Div American Mgmt Assn.
13. *The Complete Guide to Effective Vendor Management*, 2011 Gartner
14. Freeman, L.C., 1979. Centrality in social networks: conceptual clarification.
15. Rosas, J., Macedo, P., Tenera, A., Abreu, A., and Urze, P. (2015). Risk Assessment in Open Innovation Networks. In *Risks and Resilience of Collaborative Networks* (pp. 27-38). Springer International Publishing.
16. *CGEIT Risk Optimization*, ISACA
17. de Haes and van Grembergen, 2012
18. Borgatti et al chart, how society consists of patterned relationship

Enhancing Stock Index Forecasting With Ensemble-based Techniques

Dhanya Jothimani¹ and Surendra S. Yadav²

¹ Post doctorate research fellow, Data Science Laboratory
Ryerson University, Toronto, Canada,
dhanya@ryerson.ca

² Professor, Department of Management Studies,
Indian Institute of Technology Delhi, India
ssyadav@dms.iitd.ac.in

Abstract. Complex dynamics of stock market and non-linear and non-stationary characteristics of stock index data make its prediction a challenging task. Limitations of statistical and computationally intelligent models led to use of ensemble models for forecasting. In this paper, an ensemble model combining various non-classical decomposition techniques, namely, Empirical Mode Decomposition (EMD), Ensemble Empirical Mode Decomposition (EEMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), and machine learning techniques, namely, Artificial Neural Network (ANN) and Support Vector Regression (SVR), is proposed. The framework is tested on Nifty index data. Performance of the proposed models was analyzed and validated statistically. The decomposition models helped to improve the performance of machine learning models. CEEMDAN-SVR outperformed other models.

Keywords: Ensemble forecasting, Financial Time Series, Stock Index Prediction, Empirical Mode Decomposition, Artificial Neural Network, Support Vector Regression

1 Introduction

Prediction of stock prices and stock indices is, often, considered to be a difficult task due to its non-linear and non-stationary data characteristics. Complex dynamics of stock market can be attributed to various factors ranging from impact of financial news to macroeconomic factors to behaviour of investors. Fundamental and technical analysis are the most widely discussed approaches in Finance literature for stock price prediction [7]. Fundamental analysis uses fundamental indicators of the company such as Price to Equity (P/E ratio), Return on Equity (ROE) and Earnings per Share (EPS).

Technical analysis uses patterns of past stock price data to forecast the future values. Technical analysts use various tools such as technical indicators, charts and other models to monitor the patterns in stock price and volume data for a long period. The focus areas of technical analysts are: (1) to develop a model to predict the dynamics of stock market with a reasonable accuracy, and (2) to determine the market timings.

Random Walk theory states that the prices move in a random way and hence, they are unpredictable. It substantiates Efficient Market Hypothesis (EMH) and states that

tomorrow's stock price reflects all publicly available information and is independent of today's price [6].

In last two decades, few studies have shown the predictability of stock prices using macro-economic factors [16, 4, 3], technical indicators (Relative Strength Indicator (RSI), Moving Average and Momentum) and lagged price index [21] and combination of both fundamental and technical indicators [2].

Both statistical and computationally intelligent models have been used for predicting stock prices [1]. Statistical models such as AutoRegressive Moving Average (ARMA) and AutoRegressive Integrated Moving Average (ARIMA) assume the data to be linear, stationary and normally distributed. Though non-linear data can be modelled using Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and its extensions but irregularities in the capital market are not captured.

Recently, computationally intelligent techniques such as Support Vector Regression (SVR) and Artificial Neural Network (ANN) gained popularity due to their ability to model non-linear data. But they are prone to parameter sensitiveness and are subject to over-fitting. This led to the concept of ensemble models or hybrid models where more than one model is combined. It works on the idea that use of multiple predictors improves the forecast accuracy. Ensemble of decomposition models with machine learning models is one such example. Ensemble or hybrid approaches are considered to be one of the major contributions in the domain of Soft Computing [17].

The paper aims to present a comparative analysis of ensemble (Empirical Mode Decomposition (EMD), Ensemble Empirical Mode Decomposition (EEMD), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN))-based forecasting models for short-term prediction of stock index and to determine whether use of decomposition models improves the forecast accuracy.

The organization of the paper is as follows: Proposed ensemble model, data and methodology adopted in this paper are discussed in Section 2. Performance analysis of the models is provided in Section 3. Section 4 concludes the paper.

2 Ensemble Forecasting Model

The proposed ensemble models incorporate the advantages of both non-classical decomposition models and machine learning models for predicting the financial time series $F(t)$. Six ensemble models presented in the study are EMD-ANN, EMD-SVR, EEMD-ANN, EEMD-SVR, CEEMDAN-ANN and CEEMDAN-SVR. The steps of the framework are shown in Figure 1 and are detailed below:

Phase I

1. Decompose the original financial series $F(t)$ into various subseries (Intrinsic Mode Functions (IMFs) and residual component) using non-classical decomposition models (i.e., EMD, EEMD or CEEMDAN). For details on non-classical decomposition models, refer [11, 10, 8, 12, 18].

Phase II

2. Determine the model parameters for each IMF.
3. Forecast each IMF independently using ANN and SVR for (EMD, EEMD, CEEMDAN)-based ANN and (EMD, EEMD, CEEMDAN)-based SVR models, respectively.

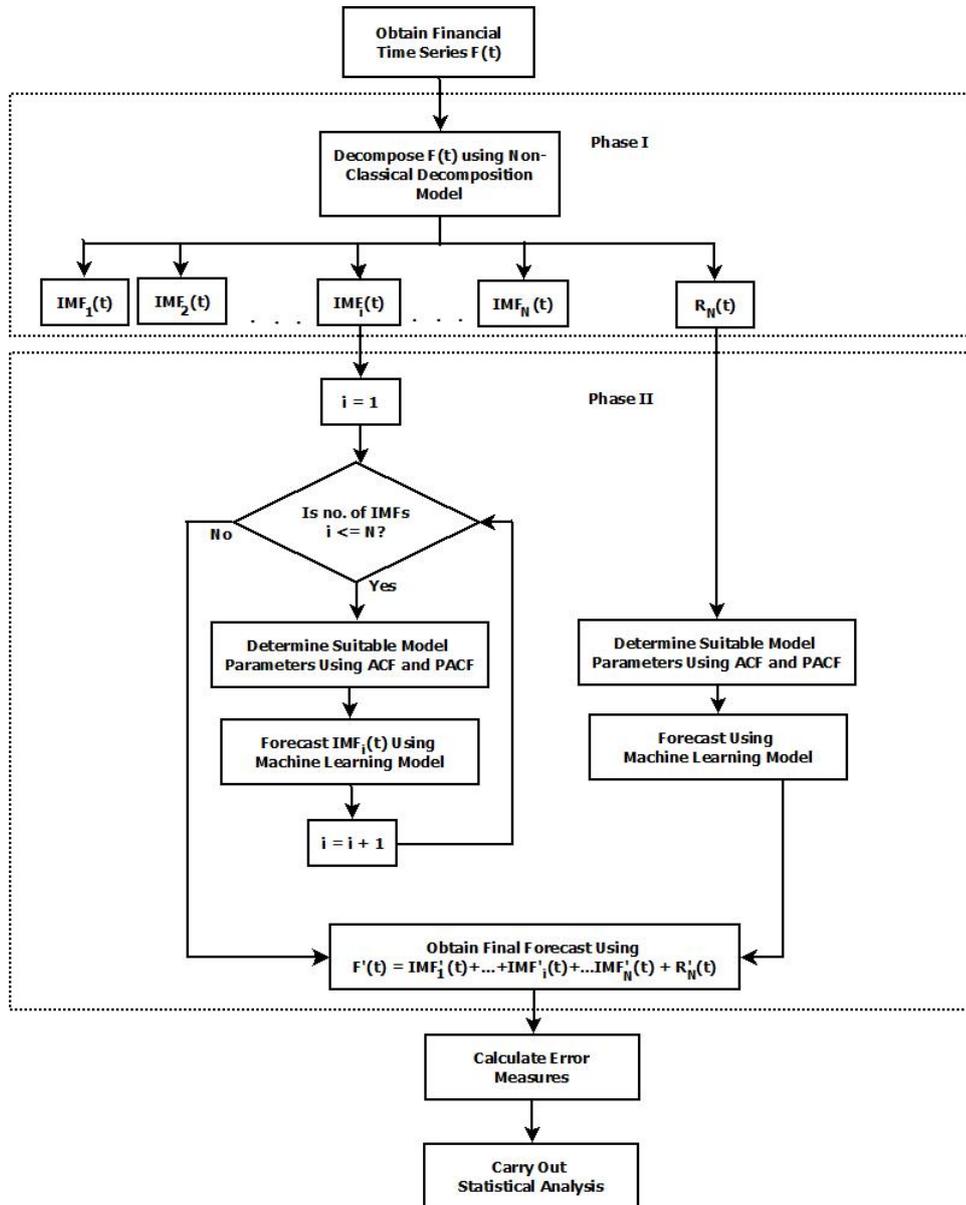


Fig. 1: Flow Chart of Ensemble Forecasting Framework

4. Similarly, determine the model parameters of the residual component and forecast the same using ANN and SVR.
5. Aggregate the forecasts of IMFs and residual component to obtain the forecast of original series, $F'(t)$.

6. Calculate error measures (namely, Root Mean Square Error (RMSE) and Directional Accuracy (DA)) for all the models.
7. Carry out Wilcoxon Signed Rank Test (WSRT) to analyze the predictive capabilities of the proposed models.

2.1 Data

There are two major stock indices for the Indian capital market. (1) Bombay Stock Exchange Sensitive Index (also known as SENSEX), and (2) National Stock Exchange Fifty Index (also known as NIFTY 50 or NIFTY). SENSEX and Nifty are the benchmark indices of Bombay Stock Exchange and National Stock Exchange, respectively. SENSEX consists of only 30 stocks while Nifty comprises of 50 stocks representing 22 sectors making it a better representative of Indian stock market.

Financial series considered for this study comprises of weekly close prices of Nifty index ranging from July 2007 to December 2015 covering a period of 8 years 6 months.

2.2 Methodology

Phase I: Non-classical Decomposition of Nifty Index

EMD was used to decompose the original series into a total of seven relatively stationary IMFs and a residual component (as shown in Figure 2). Similarly, EEMD and CEEMDAN decomposed the series into a total of seven relatively stationary IMFs and a residual component. High frequencies of IMF_1 to IMF_4 show the randomness present in the data. IMF_5 to IMF_7 represent the periodic component in the data. IMF_8 depicts the trend component of the financial series.

Phase II: Prediction of Subseries Using Machine Learning Models

In this phase, appropriate model parameters are determined and machine learning models (ANN and SVR) are used to predict each IMF and residual component independently.

- (a) **Determination of Lag Parameter:** Initially, each IMF and residual component obtained is checked for stationarity using Augmented Dickey-Fuller (ADF) test. In case of a non-stationary sub-series, first difference of the sub-series is checked for stationarity using ADF test. The process is repeated till either the sub-series becomes stationary or when maximum of number of iterations is reached. Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) are used to identify the relationship between a series and its past values. This relationship is known as lag parameter and it helps to determine the inputs for both SVR and ANN models. Figure 3 shows the ACF and PACF plot for subseries IMF_3 obtained using EEMD. Since IMF_3 cuts off at lag 4, and it has significant spikes at lags 1,2 and 3. Hence, it can be expressed mathematically as:

$$IMF_3(t) = f [IMF_3(t-1), IMF_3(t-2), IMF_3(t-3)] \quad (1)$$

Similarly, the lag values of other IMFs and residual component are determined.

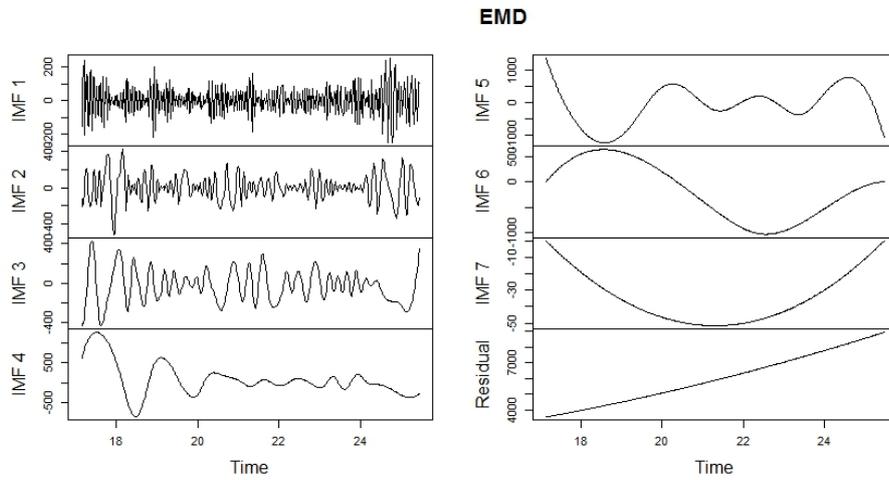


Fig. 2: Decomposition of Nifty Index Using Non-Classical Decomposition Models

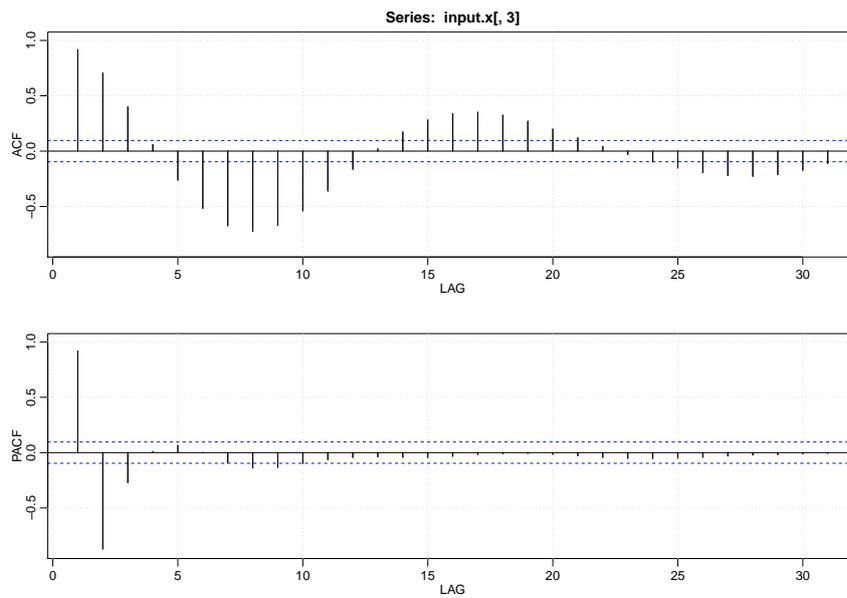


Fig. 3: ACF and PACF plot for IMF_3 of EEMD

- (b) **Prediction:** Once the lag parameters are determined, ANN and SVR are used to obtain 1-step ahead forecasts for respective (EMD, EEMD, CEEMDAN)-based ANN and (EMD, EEMD, CEEMDAN)-based SVR models. Since both ANN and SVR

are supervised machine learning algorithms, first 70% of data is used for training the model and rest 30% is used for testing the model.

The data is normalized between [0,1], [-1,1] and using z-scores. This data preprocessing step helps to reduce the probability of overfitting and the probability of the predictive models getting trapped in local minima. Further, it speeds up the training process [9, 20]. The performance of the model preprocessed using z score was found to be better than the remaining models.

- **ANN Model:** Network structure, input data format and training algorithm are three main components of a neural network. Feed forward neural network consisting of three layers, namely, input layer, hidden layer and output layer is used in this study. The number of input neurons is determined based on the lag parameter (as discussed in the previous section). The number of hidden neurons is determined based on best performances of the model. Since it is a regression problem, the number of neurons in the output layer is one. The network structure of all three (EMD, EEMD, CEEMDAN)-based models is shown in Table 1.

The input data format of IMF_4 of EMD-ANN model based on ACF and PACF is as follows:

$$IMF_4(t) = f [IMF_4(t - 2), IMF_4(t - 3), IMF_4(t - 4), IMF_4(t - 6)] \quad (2)$$

Table 1: Network Structure of ANN Used in All Three Ensemble Models

IMFs	EMD-ANN			EEMD-ANN			CEEMDAN-ANN		
	Input	Hidden	Output	Input	Hidden	Output	Input	Hidden	Output
IMF_1	3	8	1	5	10	1	4	9	1
IMF_2	3	9	1	5	9	1	6	10	1
IMF_3	4	10	1	3	9	1	6	8	1
IMF_4	4	8	1	6	8	1	3	10	1
IMF_5	5	10	1	3	9	1	4	9	1
IMF_6	4	11	1	4	10	1	6	10	1
IMF_7	6	10	1	5	9	1	3	8	1
Residue	7	7	1	7	9	1	6	10	1

Resilient Propagation algorithm [19] is used for training the model because of its superior performance over the widely used backpropagation algorithm. In addition, it quickens the training process and does not require specification of parameters (momentum and learning rate) during training process [15].

- **SVR Model:** Past values of IMFs and residual component are taken as input parameters for SVR. Grid search proposed by [14] is used to minimize the parameter sensitiveness (C and ϵ) of SVR. In grid search, values of C and ϵ are increased exponentially (for instance, $C = 2^{-5}, 2^{-3}, 2^{-1}, \dots, 2^{15}$) to determine their best values. This helps to reduce the mean square error of the model.

The transformations applied on the sub-series such as first difference, second difference and normalization should be removed before obtaining the final forecasts.

(c) **Obtaining Final Forecasts:** The final forecasts of each model are obtained by adding up the forecasts of their respective subseries. The final forecasts of the respective models can be obtained using the following equation:

$$F'(t+1) = IMF'_1(t+1) + IMF'_2(t+1) + \dots + IMF'_N(t+1) + R'_N(t+1) \tag{3}$$

where, $F'(t+1)$ is the 1-step ahead forecast of original series $F(t)$, $IMF'_1(t+1), \dots, IMF'_N(t+1)$ are the 1-step ahead forecasts of IMFs, and $R'_N(t+1)$ is the 1-step ahead forecasts of residual component.

To compare the effectiveness of ensemble-based models, the forecasts of series were obtained using traditional ANN and SVR models (both without decomposition of the series).

3 Results and Discussion

3.1 Error Measures

Two measures, namely, Root Mean Square Error (RMSE) and Directional Accuracy (DA) are used to measure and analyze the forecasting performances of the proposed models. Error is defined as the deviation of the forecasted values from the original values. RMSE is obtained by taking the square root of mean of errors and is expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n E_i(t)^2}{n}} \tag{4}$$

$$\text{where, } E_i(t) = F_i(t) - F'_i(t), \tag{5}$$

$$\tag{6}$$

n is the length of dataset,

$F(t)$ is the original series, and $F'(t)$ is the forecasted series

Directional Accuracy, represented in percentage, gives the number of times the direction of forecasted values matched that of the original series. Low values of RMSE and high values of DA indicate that the forecast model is good.

Table 2 represents the error measures for both proposed ensemble and traditional models. It can be observed that ensemble models have performed better than the traditional models. This can be attributed to decomposition of series of various relatively stationary sub-series. Decomposition aided in improving the accuracy of the machine learning models. The CEEMDAN-based ensemble models have superior performance in comparison to both EMD-based and EEMD-based models. In general, (EMD, EEMD, CEEMDAN)-based SVR models have outperformed (EMD, EEMD, CEEMDAN)-based ANN models.

Table 2: Error Measures

	RMSE DA (%)		RMSE DA(%)	
ANN	165.38	40.00	SVR	158.34 47.00
EMD-ANN	103.37	56.00	EMD-SVR	101.14 67.76
EEMD-ANN	70.31	85.60	EEMD-SVR	53.65 90.00
CEEMDAN-ANN	50.35	89.00	CEEMDAN-SVR	42.85 93.00

3.2 Statistical Tests

Though error measures indicate the superiority of ensemble models over traditional models, performances of proposed models should be validated statistically. Wilcoxon-Signed Rank Test (WSRT), a non-parametric and distribution-free technique, helps to compare and analyze the predictive capability of two models [5]. Null hypothesis of WSRT states that there is no difference in forecasting accuracy of two models. It compares the sign and ranks of the forecasted values [13].

Performance Comparison of Ensemble Models with Traditional Models: Two-tailed WSRT on RMSE values was carried out to compare the performance of traditional models and ensemble models (SVR vs (EMD, EEMD, CEEMDAN)-SVR and ANN vs (EMD, EEMD, CEEMDAN)-ANN). The results are tabulated in Table 3. Since the z statistics is beyond (-1.96,1.96), the null hypothesis of no difference in forecasting accuracy of two models is not accepted. The results are significant at 99% confidence level ($\alpha = 0.01$).

Table 3: WSRT Results

	ANN		SVR	
	z	WSRT	z	WSRT
EMD-ANN	-4.188	+	EMD-SVR	-3.255 +
EEMD-ANN	-5.990	+	EEMD-SVR	-4.880 +
CEEMDAN-ANN	-5.780	+	CEEMDAN-SVR	-5.210 +

+: EMD-ANN > ANN, EEMD-ANN > ANN, CEEMDAN-ANN > ANN, EMD-SVR > SVR, EEMD-SVR > SVR, CEEMDAN-SVR > SVR
=: EMD-ANN = ANN, EEMD-ANN = ANN, CEEMDAN-ANN = ANN, EMD-SVR = SVR, EEMD-SVR = SVR, CEEMDAN-SVR = SVR
-: EMD-ANN < ANN, EEMD-ANN < ANN, CEEMDAN-ANN < ANN, EMD-SVR < SVR, EEMD-SVR < SVR, CEEMDAN-SVR < SVR

Comparison of Performance among EMD-based Ensemble Models: Table 4 shows the performances of SVR and ANN in the EMD-based ensemble models. The statistics show SVR (EMD-SVR, EEMD-SVR, CEEMDAN-SVR) has performed than ANN (EMD-ANN, EEMD-CEEMDAN-ANN) in all three ensemble models.

4 Conclusion

The paper discussed a non-classical decomposition model (EMD) and its improved versions (EEMD and CEEMDAN) for analyzing financial time series. These decom-

Table 4: WSRT Results of ensemble EMD-based ANN vs ensemble EMD-based SVR Models

Models	z	WSRT
EMD-ANN vs. EMD-SVR	-2.564	-
EEMD-ANN vs. EEMD-SVR	-3.065	-
CEEMDAN-ANN vs. CEEMDAN-SVR	-2.748	-
+ : EMD-ANN > EMD-SVR, EEMD-ANN > EEMD-SVR, CEEMDAN-ANN > CEEMDAN-SVR		
= : EMD-ANN = EMD-SVR, EEMD-ANN = EEMD-SVR, CEEMDAN-ANN = CEEMDAN-SVR		
- : EMD-ANN < EMD-SVR, EEMD-ANN < EEMD-SVR, CEEMDAN-ANN < CEEMDAN-SVR		

position models have been integrated with two machine learning algorithms: ANN and SVR, resulting in a total of six ensemble models: EMD-ANN, EMD-SVR, EEMD-ANN, EEMD-SVR, CEEMDAN-ANN and CEEMDAN-SVR. Machine learning algorithms aided in the forecasting of time series. The models were evaluated using Nifty stock index data. The performance of models were compared using two error measures (RMSE and DA) and were validated statistically using a non-parametric Wilcoxon Signed Rank Test. The concluding remarks are summarized below:

- Ensemble forecasting models outperformed both ANN and SVR models (without decomposition).
- Among the ensemble EMD-based ANN models, CEEMDAN-ANN outperformed both EMD-ANN and EEMD-ANN models. Similarly, the performance of CEEMDAN-SVR model was found to be superior among ensemble EMD-based SVR models.
- Despite similar performances of ANN and SVR, ensemble EMD-based SVR models outperformed ensemble EMD-based ANN models. It can be concluded that EMD and its versions aided in improving the performance of SVR model.

The model was tested on stock index data of an emerging economy. It can be tested on stock index of a developed economy. The models can also be tested for intraday data. The models can be further improved by using boosting and bagging algorithms for prediction.

References

1. Atsalakis, G., Valavanis, K.: Surveying stock market forecasting techniques- Part I: Conventional methods. In: Zopounidis, C. (ed.) *Computation Optimization in Economics and Finance Research Compendium*, pp. 49–104. Nova Science Publishers, Inc, New York (2013)
2. Bettman, J.L., Sault, S., Schultz, E.: Fundamental and technical analysis: Substitutes or complements? *Accounting & Finance* 49(1), 21–36 (2009)
3. Bilson, C., Brailsford, T., Hooper, V.J.: Selecting macroeconomic variables as explanatory factors of emerging stock market returns. *Pacific-Basin Finance Journal* 9(4), 401–426 (2001)
4. Chen, N.F.: Financial investment opportunities and the macroeconomy. *The Journal of Finance* 46(2), 529–554 (1991)
5. Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253–265 (1995)

6. Fama, E.F.: Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25(2), 383–417 (1970)
7. Fischer, D., Jordan, R.: *Security Analysis and Portfolio Management*. Prentice-Hall International, US, 4th edn. (1987)
8. Jothimani, D.: *Portfolio optimization in Indian stock market: A study using Financial Analytics*. Ph.D. thesis (2017)
9. Jothimani, D., Shankar, R., Yadav, S.S.: Discrete wavelet transform-based prediction of stock index: A study on National Stock Exchange fifty index. *Journal of Financial Management and Analysis* 28(2), 35–49 (2015)
10. Jothimani, D., Shankar, R., Yadav, S.S.: A comparative study of ensemble-based forecasting models for stock index prediction. In: *MWAIS 2016 Proceedings* (2016), paper 5, <http://aisel.aisnet.org/mwais2016/5>
11. Jothimani, D., Shankar, R., Yadav, S.S.: A hybrid emd-ann model for stock price prediction. In: Panigrahi, B.K., Suganthan, P.N., Das, S., Satapathy, S.C. (eds.) *Swarm, Evolutionary, and Memetic Computing*. pp. 60–70. Springer International Publishing, Cham (2016)
12. Jothimani, D., Shankar, R., Yadav, S.S.: Ensemble of non-classical decomposition models and machine learning models for stock index prediction. In: *MWAIS 2017 Proceedings* (2017), paper 17, <https://aisel.aisnet.org/mwais2017/17/>
13. Kao, L.J., Chiu, C.C., Lu, C.J., Chang, C.H.: A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems* 54(3), 1228–1244 (Feb 2013)
14. Lin, C., Hsu, C., Chang, C.: *A practical guide to support vector classification*. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University, Taipei (2003)
15. Liu, H., Chen, C., Tian, H., Li, Y.: A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. *Renewable Energy* 48, 545 – 556 (2012)
16. Lo, A.W., MacKinlay, A.C.: Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1(1), 41–66 (1988)
17. Magdalena, L.: What is soft computing? revisiting possible answers. *International Journal of Computational Intelligence Systems* 3(2), 148–159 (2010)
18. Ren, Y., Suganthan, P., Srikanth, N.: A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. *IEEE Transactions on Sustainable Energy* 6(1), 236–244 (Jan 2015)
19. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: *IEEE International Conference on Neural Networks*. vol. 1, pp. 586–591 (1993)
20. Wu, G., Lo, S.: Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. *Expert Systems with Applications* 37(7), 4974 – 4983 (2010)
21. Yao, J., Tan, C., Poh, H.L.: Neural networks for technical analysis: A study on KLCI. *International Journal of Theoretical and Applied Finance* 02(02), 221–241 (1999)

GARCH-VMD Based Forecasting for Volatile Time Series of Indian Small Car Sales

Rajeev Pandey

Department of Management Studies,
Indian Institute of Technology Delhi, India
rajeev.mech@gmail.com

Abstract. The study presents a novel approach based on combination of statistical and functional decomposition techniques to model univariate time series with conditional volatility. GARCH is fitted to the original or detrended data based on stationarity in the first step. In the second step, variational mode decomposition is used to decompose the stationary series. Multistep ahead forecasting based on machine learning techniques is done to obtain forecast for multiple horizons. The individual forecasts are then aggregated to obtain final forecasts. The method outperforms benchmark statistical and machine learning techniques for both short and medium forecasting horizons.

Keywords: forecasting, volatility, functional decomposition, variational mode decomposition

1 Introduction

Time series data is characterized by stationarity, seasonality, volatility and length. Statistical techniques such as AutoRegressive Integrate Moving Average (ARIMA) and Seasonal AutoRegressive Integrate Moving Average (SARIMA) model non-stationarity and seasonality in the data, respectively. ARIMA/SARIMA is a linear model and it assumes constant variance of error. AutoRegressive Conditional Heteroskedasticity (ARCH) and its variations such as Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and Exponential Generalised AutoRegressive Conditional Heteroskedasticity (EGARCH) are used to model conditional volatility. Researchers [1, 12] have integrated ARIMA/SARIMA with ARCH/GARCH to form a non-linear time series model. It helps to take advantages of both the techniques, where ARIMA models AR, MA and the non-stationarity (through differences) in the data while GARCH models the variance in the error. They have observed that performance of ARIMA-GARCH is better than univariate time series modelling using only ARIMA or GARCH [1, 12].

Though there has been an increase in the use of machine learning techniques to forecast univariate time series in various domains [14, 3, 13] but very few studies have modelled volatility in the time series data using machine learning techniques. To overcome the limitations of GARCH type models such as inability to model multiscale nature of volatility and non-linear effects, and limited capability to capture direct long memory effects [5]; researchers developed machine learning based volatility models [5,

10, 6]. Extensions of GARCH such as nonlinear exponential GARCH, threshold GJR-GARCH and quadratic GARCH models that were developed to address the issue of non-linear volatile series suffered from limitations of being parametric and they assume specific functional form about data generation process, which limits the overall modeling capacity [2]. The advantages of using machine learning techniques such as SVM for volatility modelling are that they do not assume probability density function over the time series data and the parameters of the models are adjusted based on the principle of empirical risk minimization. An attempt to develop non-parametric GARCH models led to the development of machine learning based volatility models, which is an integration of ARCH models and its extensions with machine learning models, namely, ANN and Support Vector Machine (SVM). These studies conclude that hybrid GARCH-machine learning models perform better than GARCH techniques for modelling volatility in time series data [5, 10, 6].

Few limitations of classical decomposition models are unavailability of trend values for first and last few observations, assumption of repetition of seasonal component throughout the period of analysis and sensitivity to outliers or unusual pattern in the data [8]. This led to the use of functional decomposition models such as Discrete Wavelet Transform [9], Empirical Mode Decomposition and its variations, and Variational Mode Decomposition [11] along with machine learning models to obtain final forecast values. Each of the components is forecasted either using statistical techniques or machine learning techniques and forecasted components are combined to generate final forecasted values. Individual components are easier to forecast than the original series. Above mentioned studies focussed on improving the forecast accuracy of machine learning models based on decomposition models.

Volatility based forecasting has been addressed in Finance domain but few studies focussed on volatility based forecasting in other sectors like automotive sales. Hence, the objective of this study is to propose a framework for volatility-based sales forecasting. The study also aims to study the effect of decomposition in forecasting the time series data. There are limited studies using ensemble method, namely, Random Forest for forecasting timeseries as well as for modelling volatility. This study also tries to address this gap.

The paper is organised as: Section 2 discusses methodology followed by Data description and analysis in Section 3. Section 4 concludes the chapter.

2 Methodology

The framework developed for modelling volatility in automotive sales forecasting is shown in Figure 1. The steps are enumerated below:

1. Since the tests for checking volatility cannot be applied on non-stationary data, first, the data is checked for stationarity using correlogram and Unit Root test. If the series is not stationary, it is made stationary by differencing the series.
2. Data is checked for ARCH effects using ARCH-LM test.
3. If the data shows ARCH effects, which shows that the data contains conditional volatility, then it is fitted using GARCH model.

4. Now, the fitted data is decomposed into various subseries using Variational Mode Decomposition (VMD).
5. Each of the subseries obtained is forecasted using machine learning models such as SVM-LK, SVM-PK, ANN and RF for 1-, 2-, 3-, 6- and 12-steps ahead forecast horizon.
6. Final forecast is also obtained using the four above mentioned machine learning models for combining the individual forecasts.
7. Performance of the models is evaluated using error measures such as RMSE.

2.1 Functional Decomposition

Functional decomposition is the process of dividing a series to a number of subseries in such a way that original series can be reconstructed by subseries. Signal processing techniques such as Fourier transform (FT), wavelet transform (WT) and Empirical Mode Decomposition (EMD) have been used widely to capture underlying structure and information in a complex non-stationary and non-linear data. FT transforms the data in frequency domain, hence, the information on time is lost. FT cannot be used for analysing the local properties of the signal. One of limitations of WT is the identification of scale factor (or decomposition level) to be used in Wavelet Transform. Both FT and WT work on the series based on predefined functions, independent of the data [16]. This led to development of data adaptive models such as Empirical Mode Decomposition (EMD) and its variations.

Proposed by [7], EMD is a data-driven method and uses Huang-Hilbert Transform (HHT) to express non-stationary signals as a sum of Intrinsic Mode Functions (IMFs). EMD is an adaptive technique as it does not require any prior information regarding the number of the components (IMFs) to be obtained. It allows perfect reconstruction of original time series by taking summation of all the IMFs. Its ability to separate the signal into stationary and non-stationary components; and adaptive and perfect reconstructive properties of EMD has contributed to its wide applicability in time series denoising and forecasting. However, it suffers from few limitations. It lacks strong mathematical background. Decomposition of signals is highly dependent on methods adopted for finding the local minima and maxima, interpolation of these points into envelopes and the stopping criteria. Another major limitation of EMD is that it suffers from of mode mixing problem [17, 15]. Mode mixing is caused by signal intermittency, which could affect the physical meaning of IMF. These limitations led to development of Variational Mode Decomposition (VMD).

Variational Mode Decomposition Variational Mode Decomposition (VMD), developed by [4], is an adaptive quasi-orthogonal signal decomposition model to decompose multi-component signal to a set of finite band-limited intrinsic mode functions (IMFs). The IMFs generated are called modes and are different from those obtained using Empirical Mode Decomposition (EMD). The band-limited mode $y_k(t)$ is defined as:

$$y_k(t) = B_k(t)\cos(\phi_k(t)) \quad (1)$$

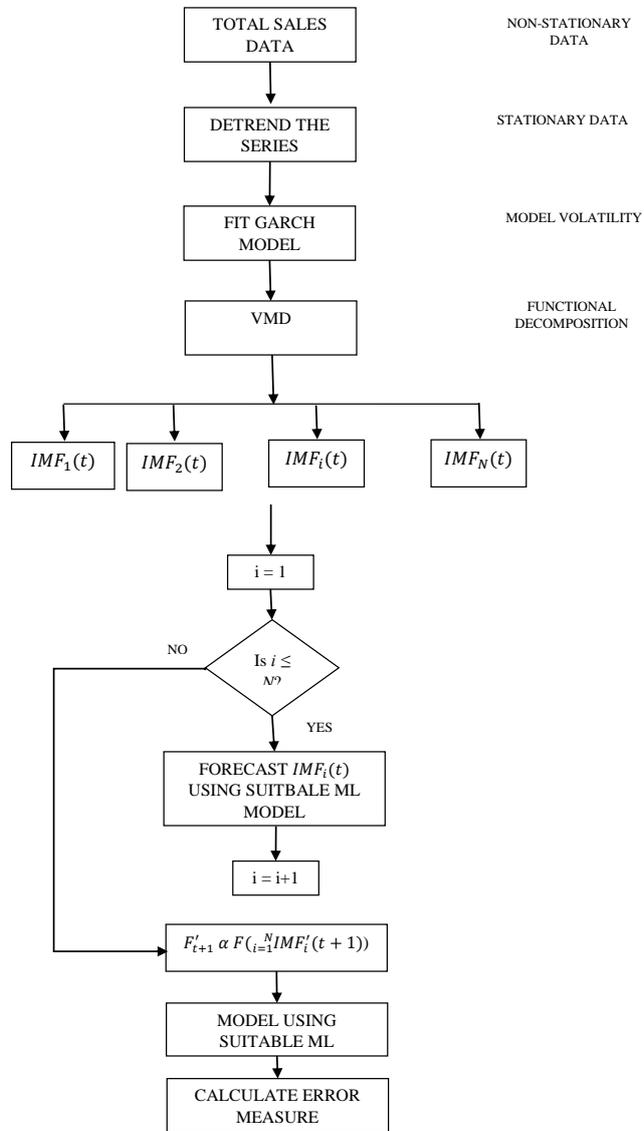


Fig. 1. Flowchart of Methodology

where, $\phi_k(k)$ represents the phase and is a non-decreasing function. Similarly, the envelope $B_k(t) \geq 0$ is also a non-decreasing function. Further, it is also considered that both the instantaneous frequency $\dot{\phi}_k(t) = \phi'_k(t)$ and the envelope vary much slowly as compared to the phase.

VMD can be represented as a constrained problem using the following equation:

$$Min_{y_k, w_k} = \sum_k \|\partial_t [(\delta(t) + j/\pi t) * y_k(t)] e^{-jw_k t}\|_2^2 \tag{2}$$

subject to

$$\sum_k y_k = f$$

where, w_k is the center frequency, y_k is the k^{th} mode, * represents convolution, k is number of modes, $j^2 = -1$, δ is the dirac distribution and f is the original signal.

The above constrained problem is converted to an unconstrained problem using Lagrange multiplier and the augmented Lagrangian is represented as:

$$L(y_k, w_k, \lambda) = \beta \sum_k \|\partial_t [(\delta(t) + j/\pi t) * y_k(t)] e^{-jw_k t}\|_2^2 + \|f - \sum_k y_k\|_2^2 + \langle \lambda, f - \sum_k y_k \rangle \tag{3}$$

where, β represents the balancing parameter of data fidelity constraint and λ is the Lagrange multiplier.

The equation 3 is solved using Alternate direction method of multipliers (ADMM) approach to obtain different decomposed modes and the center frequency during each shifting operation. The obtained modes are represented as:

$$\hat{y}_k(w) = \left[\hat{f}(w) - \sum_{i \neq k} \hat{y}_i(w) + (\hat{\lambda}(w)/2) \right] / [1 + 2\beta(w + w_k)^2] \tag{4}$$

The steps of VMD can be enumerated as:

1. *Updating the modes:* The modes are updated using Wiener filtering in the Fourier domain.

$$\hat{y}_k^{n+1}(w) = \left[\hat{f}(w) - \sum_{i < k} \hat{y}_i^{n+1}(w) - \sum_{i > k} \hat{y}_i^n(w) + (\hat{\lambda}(w)/2) \right] / [1 + 2\beta(w + w_k)^2] \tag{5}$$

2. *Updating the center frequencies:* The center of gravity approach for power system is used for updating the center frequencies as shown below:

$$w_k^{(n+1)} = \int_0^\infty w |\hat{y}_i^{n+1}(k)|^2 dw / \int_0^\infty |\hat{y}_i^{n+1}(k)|^2 dw \tag{6}$$

3. *Updating dual ascent*: Here, the Lagrange multiplier is updated as dual ascent to ensure exact reconstruction of signals till the condition $\sum_k \|\hat{y}_k^{n+1} - \hat{y}_k^n\|_2^2 / \|\hat{y}_k^n\|_2^2 < \epsilon$ is satisfied.

$$\hat{\lambda}^{n+1} = \hat{\lambda}^n + \tau(\hat{f} - \sum_k \hat{y}_k^{n+1}) \quad (7)$$

Since VMD uses Wiener filtering, it is robust to noise in the signal. Unlike EMD, it is non-recursive in nature.

3 Analysis

Total monthly sales data of small car segment in India for a period of 19 years ranging from April 1998 to March 2017 is considered for analysis. This results in a total of 228 data points for analysis. Analysis is carried based on the proposed framework depicted in Figure 1. ADF test is used for checking stationarity in the series. Based on the results (refer Table 1), it can be concluded that the series is non-stationary. First difference of the series is obtained to make the series stationary¹. Then it is tested for ARCH effects using ARCH-LM test. Based on p-value, the null hypothesis of there is no ARCH effects, is rejected (refer Table 2), hence, it can be concluded that the series is volatile in nature.

Table 1. Results of ADF Test

Dataset	Optimal lag length	Unit Root	
	p*	τ_μ	$Pr < \tau_\mu$
Original series	13	-0.872361	0.7955

Test critical values: 1% (-3.45923), 5% (-2.87414) and 10% (-2.57356).

Table 2. Results of ARCH-LM Test

Variable	Co-efficient	Std. Error	t-statistics	Prob.	F-statistics	Prob. Chi-sq.(-1)
C	1.12E+08	17401491	6.417792	0	4.27598	0.0396
<i>RESID</i> ² (-1)	0.136836	0.066173	2.067844	0.0398		

The detrended series is fitted using GARCH model (Figure 2), which is, then, decomposed to three IMFs using VMD (refer Figure 3). Each IMF is forecasted using both statistical techniques such as ARIMA and SARIMA and machine learning techniques such as Support Vector Machine using Polynomial Kernel (SVM-PK), Support Vector Machine using Linear Kernel (SVM-LK), Artificial Neural Network (ANN) and

¹ Obtained series is also known as detrended series.

Random Forest (RF), for 1-step, 2-, 3-, 6- and 12-steps ahead forecast horizons. The final forecast is obtained by combining the IMFs using machine learning models. The proposed framework of modelling volatility with GARCH and forecasting the series using functional decomposition models and machine learning techniques is compared with statistical models and machine learning models. The performance of the models is evaluated using Root Mean Square Error (RMSE).

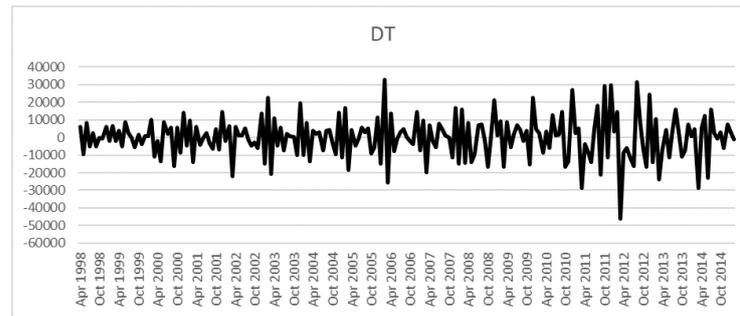


Fig. 2. Detrended Series Before VMD

Results for multistep ahead forecasting as obtained by the four different machine learning techniques are presented in Figure 4. SVM-PK consistently provides better results than all the other three methods with SVM-LK and ANN giving comparable results. Random forest performs moderately in this scenario.

Figure 5 presents the forecasting results for machine learning and statistical techniques. Two volatility based statistical models, namely, ARIMA + GARCH and SARIMA + GARCH are also used to model volatility along with autoregressive and moving average components. For 1-step ahead, SVM-PK gives the best results and for the remaining step ahead forecasts, both SVM-PK and SVM-LK outperform other techniques.

RMSE values of forecasts obtained with and without VMD are compared to understand the effect of functional decomposition model. For this purpose, best models in terms of low RMSE values in both cases, i.e., series modelled using GARCH-VMD-Machine learning/statistical techniques are compared. Percentage difference in the RMSE value is calculated for all five forecast horizons. Table 3 displays the effect of functional decomposition for detrended series. It can be observed that all the models using VMD performed better than the models without VMD for all forecast horizons. It can be concluded that the decomposition model aided in improvement of forecast accuracy as it is easier to forecast the individual components than the original series.

4 Conclusion

The study developed a volatility-based framework for forecasting automotive sales. The framework consisted of modelling volatility using GARCH followed by decomposing the series using VMD and forecasting each component using both machine learning as

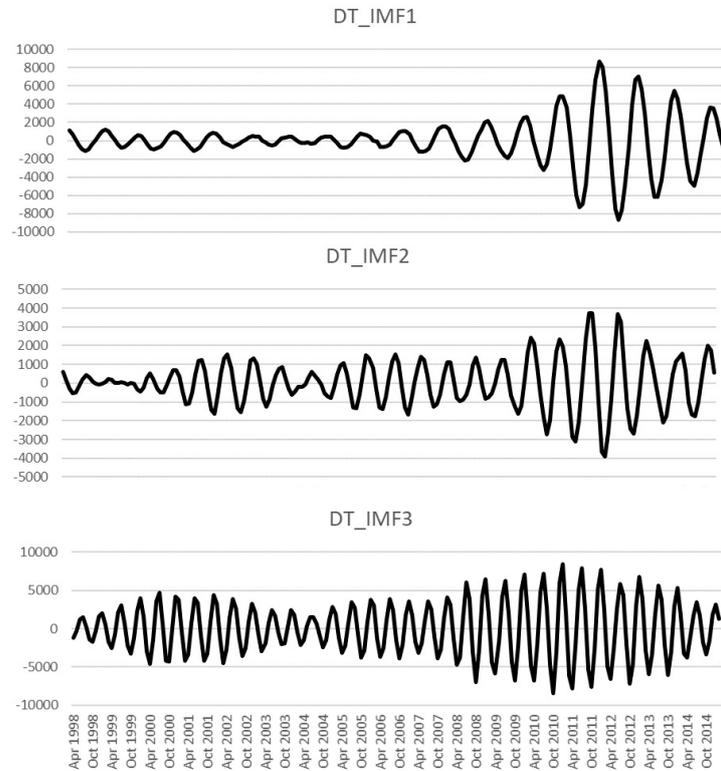


Fig. 3. Detrended Series: IMF_1 Obtained Using VMD

Table 3. Effect of functional decomposition in detrended series

Forecast Horizon	GARCH-VMD	GARCH-ML/SL	%Difference
1	4646.6197	6886.8306	48.212
2	7400.3064	7758.8407	4.845
3	7530.8866	7926.6379	5.255
6	8479.5251	8848.2900	4.349
12	8765.2273	9410.2068	7.358

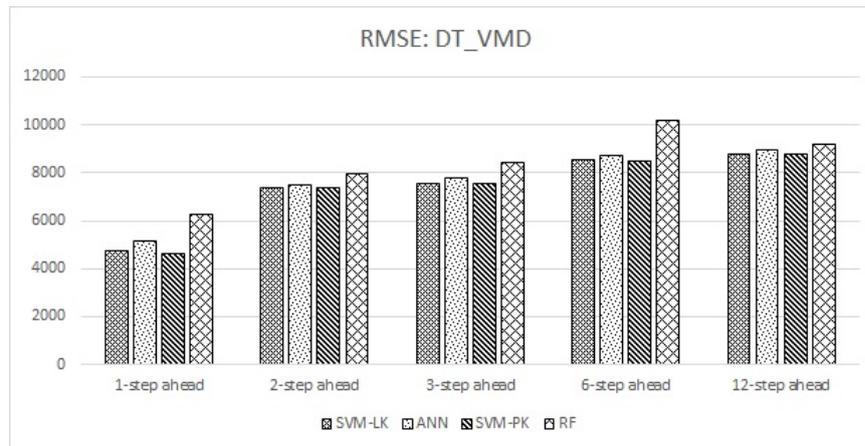


Fig. 4. RMSE of forecasts obtained using VMD and machine learning algorithms for the detrended series

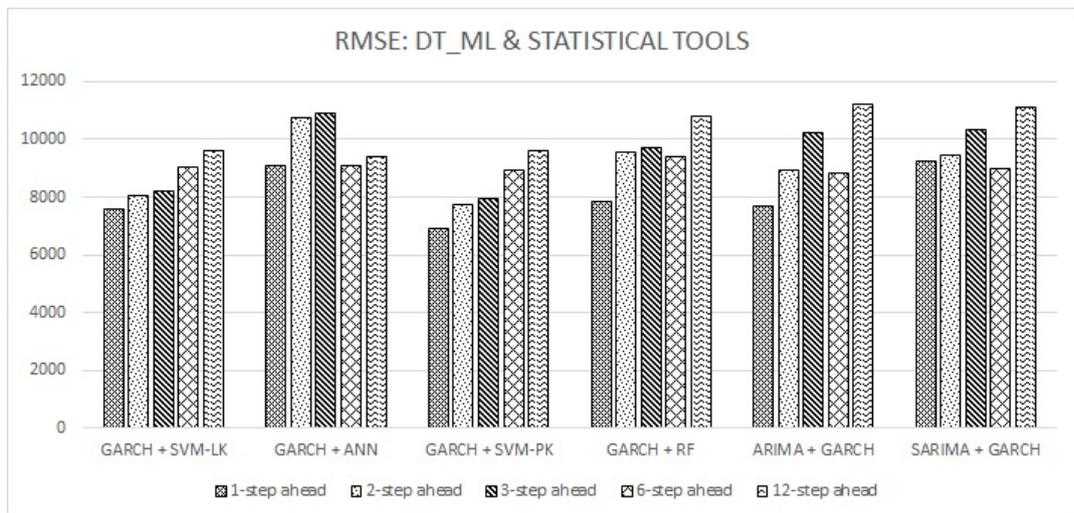


Fig. 5. RMSE of forecasts obtained using machine learning algorithms and statistical tools (without decomposition of series) for the detrended series

well as statistical techniques. This study also explored the use of ensemble machine learning method, namely, random forest for forecasting the time series data. Machine learning models performed better than statistical techniques in case of decomposition. Out of all machine learning techniques, SVM-PK performed better in most of the cases. Random Forest did not outperform other techniques. This could be attributed to the fact that it can handle complex data and the individual components generated after decomposition are low (or medium) in data complexity. Further, it can be concluded

that functional decomposition model used, namely, VMD aided in improvement in the forecast accuracy.

References

1. Chen, C., Hu, J., Meng, Q., Zhang, Y.: Short-time traffic flow prediction with ARIMA-GARCH model. In: 2011 IEEE Intelligent Vehicles Symposium (IV). pp. 607–612 (June 2011)
2. Chen, S., Haedle, W.K., Jeong, K.: Forecasting volatility with support vector machine-based GARCH model. *Journal of Forecasting* 29, 406–433 (2010)
3. Dong, Y., Wang, J., Guo, Z.: Research and application of local perceptron neural network in highway rectifier for time series forecasting. *Applied Soft Computing* 64, 656–673 (2018)
4. Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. *IEEE Transactions on Signal Processing* 62(3), 531–544 (Feb 2014)
5. Gavrishchaka, V., Banerjee, S.: Support vector machine as an efficient framework for stock market volatility forecasting. *Computational Management Science* 3(2), 147–160 (2006)
6. Hernández: Volatility of main metals forecasted by a hybrid ANN-GARCH model with regressors. *Expert Systems with Applications* 84, 290–300 (2017)
7. Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., Liu, 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), 903 – 995 (1998)
8. Hyndman, R.J., Athanasopoulos, G.: *Forecasting: principles and practice* Paperback. OTexts., 1st edn. (2013)
9. Jothimani, D., Shankar, R., Yadav, S.S.: Discrete wavelet transform-based prediction of stock index: A study on National Stock Exchange fifty index. *Journal of Financial Management and Analysis* 28(2), 35–49 (2015)
10. Kristjanpoller, W., Minutolo, M.C.: Forecasting volatility of oil price using an artificial neural network-GARCH model. *Expert Systems with Applications* 65(C), 233–241 (Dec 2016)
11. Lahmiri, S.: Comparing variational and empirical mode decomposition in forecasting day-ahead energy prices. *IEEE Systems Journal* 11(3), 1907–1910 (Sept 2017)
12. Liang, Y.H.: Forecasting models for taiwanese tourism demand after allowance for mainland china tourists visiting taiwan. *Computers & Industrial Engineering* 74, 111 – 119 (2014)
13. Lu, X., Que, D., Cao, G.: Volatility forecast based on the hybrid artificial neural network and garch-type models. *Procedia Computer Science* 91, 1044 – 1049 (2016), promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016)
14. Ruiz, L., Rueda, R., Cuéllar, M., Pegalajar, M.: Energy consumption forecasting based on elman neural networks with evolutive optimization. *Expert Systems with Applications* 92, 380 – 389 (2018)
15. Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P.: A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4144–4147 (May 2011)
16. Wang, Y., Markert, R.: Filter bank property of Variational Mode Decomposition and its applications. *Signal Processing* 120(C), 509–521 (2016)
17. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis* 1(1), 1–41 (2009)

Solar Irradiance forecasting of Ahmedabad based on Ant Colony Optimization and Neural Network

Md. Janibul Alam Soeb¹, Md. Irfanul Hasan², Md. Shahid Iqbal³

^{1, 2, 3} Department of EEE, Sylhet Engineering College, Bangladesh

Email: mjasoeb@gmail.com

Abstract.

This paper presents, a new prediction model for solar irradiance of Ahmedabad, India based on ant colony optimization (ACO) and neural network (NN), called as ACOSIP. In ACOSIP, the most salient climatological features are selected to enhance the solar irradiance prediction (SIP) accuracy. To implement such idea, ACO search technique utilizes the advantages of the combined activities of the features by considering the correlation information among the features and the outcome of NN. Thus, ACOSIP introduces the wrapper and filter approaches in its feature selection process. To make an effective ACO search, two sets of new rules have been designed for pheromone update and heuristic information measurement. To evaluate the performance of ACOSIP, 12 solar irradiance data samples in between the year of 2000-2016 were collected for Ahmedabad. Experimental results show that ACOSIP can select six most salient features easily with increasing the prediction accuracy, which are longitude, latitude, day light hour, max temp., min temp., and humidity. In addition, the averaged prediction accuracy of ACOSIP for Ahmedabad in testing case is 99.84% including the MAPE of 0.27%. The proposed ACOSIP also represents high correlation of 99.95% in between the actual and forecasted data.

Keywords: Solar radiation forecasting, Artificial neural network, Ant colony optimization and Feature selection.

1 Introduction

Solar energy research is based on solar radiation data but, is not available for most of the sites due to non-availability of solar radiation measuring equipments at the meteorological stations. Therefore, it is essential to predict solar radiation for a location using several climatic variables. Furthermore, the application of solar energy among the different types of renewable energy sources has more attentions nowadays throughout the whole world that might be increasing in the following days.

Several works have already been published in the literature related to modeling and prediction of solar radiation for various solar energy applications [1]. Analytic

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

formula [3], numeric simulation or statistical approaches [4] are developed two or three decades before to predict solar irradiance. But, these prediction models resulted in large errors. To overcome these limitations, NN-based models [3-9] has been introduced in the recent years for predicting the solar irradiance.

Mohandas proposed a model in [5] where solar irradiance data was collected from different 41 stations in Saudi Arabia to train the NN for finding the irradiance prediction accuracy. The incorporated input variables are latitude, longitude, altitude and sunshine duration. The results for the testing stations are within 16.4%. In [6], NN-based model used 12 various Indian weather stations with different climatic conditions. The results of RMSE in NN model is in between 0.0486-3.562 for Indian region. On the other hand, S. Rahman's proposed model [7] used data for Abha city of Saudi Arabia in NN training where the result of MAPE is 4.49%. S. Quaiyum proposes a SIP model [8] for Bangladesh obtained the result of MSE is 0.29%. Hasni proposes a SIP model [9] for south-western region of Algeria where the MAPE obtained is 2.9971%. In [3], Kalogirou has reviewed the use of NN in the applications of renewable energy systems. In addition, Mellit [4] has shown NN for sizing of PV systems.

According to the afore-mentioned proposed models for predicting the solar irradiance, it has been observed that, NN-based models cannot perform well in predicting the solar radiation using all the available features. The reason is that, the performance of NN always degrades if some irrelevant features exist in the available feature set [10]. Therefore, selecting those irrelevant features from the original set is a crucial work in the NN-based SIP process. In this regard, will tries in [11] to select climatic variables for 14 weather stations in North Argentina using niching genetic algorithm. However, in this paper, an ACO-based SIP model has been proposed to select the most relevant features to increase the prediction performance of NN. ACO here incorporates two techniques in its weight update rules, which are wrapper and filter approaches.

The remainder of this paper is organized as follows. ACOSIP are broadly discussed in Section 2. Simulation results, comparison to other methods, are reported in Section 3. Short conclusions with few remarks are given in Section 4.

2 Proposed ACOSIP Model

To determine the best climatological features in increasing the solar irradiance accuracy, ACOSIP uses hybrid ACO based search technique. Furthermore, to obtain a small subset of salient features, the ants are guided at its traversal period upon the feature space. For comprehensibility, the proposed ACOSIP model can be discussed by the flowchart shown in Fig. 1, which is described as follows:

i) Let D is a given data set with n distinct features. Initialize the pheromone trails τ and the heuristic information η of all n features by giving equal values to τ and η .

ii) Measure the correlation data of individual n features using correlation measurement scheme. The reason is to select more distinct features by ants. Thus, the ants are guided to select the features which are less correlated among the original features.

iii) Motivate the ants to choose location information (i.e., latitude and longitude) at each iteration.

iv) Generate a set of artificial A ants equivalent to n , i.e. $A = n$.

v) Subset size, s is determined prior to SC for each of the A ants. By using the conventional probabilistic transition rule [13] for selecting features with which to construct the subsets as follows:

$$P_i^A(t) = \begin{cases} \frac{[\tau_i(t)]^\alpha [\eta_i(t)]^\beta}{\sum_{u \in j^A} [\tau_u(t)]^\alpha [\eta_u(t)]^\beta} & \text{if } i \in j^A \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where j^A is the set of effective features, τ_i and η_i are the pheromone and heuristic values for the feature i ($i = 1, 2, \dots, n$). The relative importance of the pheromone value and heuristic information is decided by α and β . Eq. (1) shows random behavior in subset construction initially since the initial value of τ and η for all individual features are equal.

vi) After completion of subset construction by all ants, then system move forward, on the other hand system will back to Step 4.

vii) The subsets $R^A(t)$ is evaluated and measure the performance using MAPE $M(R^A(t))$. Here, $R^A(t)$ refers to the effective sets constructed by A ants at iteration t .

viii) The $R^l(t)$ (local best subset) among all $R^A(t)$ and the R^g (global best subset) among all $R^l(t)$ is selected with the searching mythology. Here, $t = 1, 2, 3, \dots, I$ where I is a number of iterations.

ix) Check whether $R^l(t)$ attains a predefined SIP performance, or the algorithm execution reaches an iterative limit I_{th} and then terminate the selection process. Note that, at iteration I_{th} the model cannot find anymore variation in R^g . Otherwise continue the FS and save the evaluation results of all local best subsets, $M(R^l(t))$ for further use.

x) The values of τ and η for all the features will be updated according to the rules of pheromone update and heuristic value measurement, respectively.

xi) A new set of artificial A ants will be generated and proceed the scheme similarly.

Clearly, the idea behind ACOSIP is straightforward, which used ACOFS algorithm to guide the ants in bounded region (between 3 to 6) of search space and providing mixed technique to the ant's search. To select the distinct features which is helpful for effective learning of NN, a correlation information measurement procedure has been integrated, so performed only once throughout the FS process.

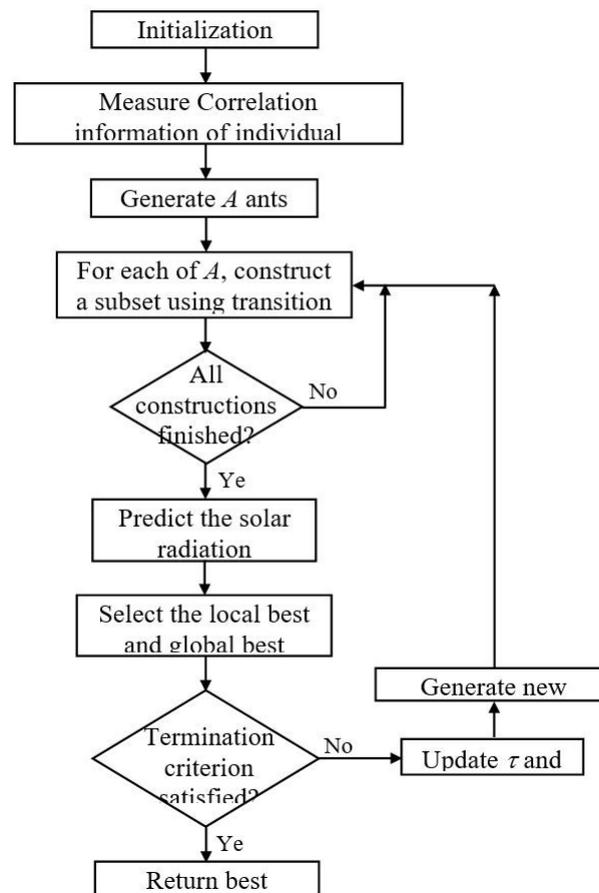


Fig. 1. Flowchart of ACOSIP

3 Hybrid Search process

In ACOSIP, a hybrid search technique has been introduced that is composed by wrapper and feature approaches. These approaches are embedded in the newly designed two sets of rules, which are pheromone update and heuristic value. Note that, these both rules

are involved with correlation information measurement of individual features, which are further described as follows:

3.1 Pheromone Update Rule

On basis of random and probabilistic phenomena, a set of two pheromone update rules are introduced from the original one mentioned in [10] as follows:

Random case: The rule presenting in Eq. 2 is selected from [10], in which case second term is divided by f_i . The features are selected to construct subsets randomly according to their experiences. This enhances the exploration capability of ants. The rule has been accepted as follows:

$$\tau_i(t+1) = (1 - \rho)\tau_i(t) + \frac{1}{f_i} \sum_{A=1}^n \Delta\tau_i^A(t) + e\Delta\tau_i^g(t)$$

$$\Delta\tau_i^A(t) = \begin{cases} M(R^A(t)) & \text{if } i \in R^A(t) \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

$$\Delta\tau_i^g(t) = \begin{cases} M(R^l(t)) & \text{if } i \in R^l(t) \\ 0 & \text{otherwise} \end{cases}$$

Here, i refers to the number of feature ($i = 1, 2, \dots, n$), and f_i is the count for the specific selected feature i in the current iteration. $\Delta\tau_i^A(t)$ is the value of pheromone taken by the local update for feature i , which is included in $R^k(t)$ at iteration t . Similarly, the global update, $\Delta\tau_i^g(t)$, is the value of pheromone for feature i that is included in $R^l(t)$. Finally, ρ and e are the pheromone decay value and elitist parameter, respectively.

Probabilistic case: Eq. (3) shows the updated pheromone rule (probabilistic case).

$$\tau_i(t+1) = (1 - \rho)\tau_i(t) + \sum_{A=1}^n \Delta\tau_i^A(t) + e\Delta\tau_i^g(t)$$

$$\Delta\tau_i^A(t) = \begin{cases} M(R^k(t)) \times C_i & \text{if } i \in R^A(t) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\Delta\tau_i^g(t) = \begin{cases} M(R^l(t)) \times C_i & \text{if } i \in R^l(t) \\ 0 & \text{otherwise} \end{cases}$$

Here, feature l is renown by the global update. The $\Delta\tau_i^g$ is provided only to those features that are efficient, because, global update has an important part in selecting the effective features in ACOSIP.

3.2 Heuristic Value Measurement

In this paper, we propose a set of new heuristic value estimation rule depending on two basic phenomena in SC. To describe the attractiveness for each feature heuristic value, η , is important. So, to overcome the greediness, it is therefore necessary [12], and finally an effective solution may be appeared.

Random case: In the initial stage, the ants are involved to construct the feature subsets randomly, and the heuristic value of all features i can be estimated as follows:

$$\eta_i = \frac{1}{f_i} \sum_{A=1}^n M(R^A(t)) (1 + \phi e^{-\frac{|R^A(t)|}{n}}) \quad (4)$$

if $i \in R^A(t)$

Probabilistic case: In the following iterations, after constructing subsets on the basis of the probabilistic behavior, the following formula is used to estimate η for all features i :

$$\eta_i = f_i \phi_i \sum_{A=1}^n M_a(S^k(t)) C_i (1 + \phi e^{-\frac{|S^k(t)|}{n}}) \quad (5)$$

if $i \in S^k(t)$

In these two rules, ϕ refer to the count of a selected feature I from previous iterations to the current iteration, except for the initial iteration. C_i refers to the correlation value for feature i . The aim of including C_i is based on the following two factors: (a) to select the more distinct features during SCs, and (b) enhancing the robustness of learning model. Thus, best features may be selected in the SC for different iterations, thus of course enhancing the prediction accuracy of the ACOSIP.

3.3 Correlation Value Measurement

To measure relationship between the climatological features which are used as the input pattern in ANN and to choose distinct features correlation information (CI) is important and useful statistics. To reduce the effect of features to each other and to increase the learning power of ANN we need more uncorrelated features set in training purpose. This information is providing to heuristic information measurement η of all individual features. In this work, Pearson product-moment correlation coefficient is used. The correlation coefficient, r_{ij} between two features, i and j is,

$$r_{ij} = \frac{\sum_p (x_i - \bar{x}_i)(y_j - \bar{y}_j)}{\sqrt{\sum_p (x_i - \bar{x}_i)^2} \sqrt{\sum_p (y_j - \bar{y}_j)^2}} \quad (6)$$

The correlation coefficient is computed for all possible combinations of features then ACOSIP computes the correlation of each feature. The correlation, Cor_i , of any feature i is

$$Cor_i = \frac{\sum_{j=1}^n |r_{ij}|}{n-1} \quad i \neq j \quad (7)$$

where n is the number of features used. Then to find the uncorrelated features we use C_i , which is

$$C_i = 1 - Cor_i \quad (8)$$

4 Data Collection

The significance of ACOSIP has been evaluated in this section on twelve well known climatological features as longitude (Ln), latitude (Lt), elevation (ev), year (y), month (m), station constant (sc), day light hour (DL), relative humidity (Rh), maximum temperature (Mat), minimum temperature (Mit), rainfall (Rf) and wind speed (Ws), and the solar radiation data is used as the target. These features have been used in many of solar radiation prediction models using ANN. In this study we use 16 years (2000 - 2016) monthly averaged data, for Ahmedabad, which are collected from Meteorological Centre, Ahmedabad.

5 Methodology

To verify the effectiveness of ACOSIP for FS and RP, a number of experiments have been carried out on ACOSIP. To justify the FS and RP task suitability in ACOSIP, one basic step need to be considered, namely, giving values for user-specified parameters such as $\eta, \tau, \alpha, \beta, \rho$ and ϕ , which are common for the all datasets. These parameters have been chosen after some runs. To achieve a level of balance between exploitation and exploration proper selection of the values of parameters is helpful of ants to select feasible features.

6 Experimental Results

The effectiveness of ACOSIP model is verified by a succession of experiments. To prove the strength of the proposed model ACOSIP models have been designed for Ahmedabad and then results are being averaged to show the global strength of the model – shown in Table 1. As Table 1, the avg. results of PA and MAPE for Ahmedabad is 99.72 % (training), 99.56% (testing) and 0.44% (training), 0.51 (testing), respectively validated the ACOSIP structures for predicting solar irradiance.

Table 1. Averaged performance results of ACOSIP

Local Investigated Data					
DATA	Avg. PA (%)	Avg. MAPE (%)	Avg. RMSE (%)	Avg. MSE (%)	Avg. R (%)
Training	99.72	0.44	0.244	0.077	99.81
Testing	99.56	0.51	0.289	0.104	99.72

Table 2. Performance results for well performed stations

Station Names	DATA	Avg. PA (%)	Avg. MAPE (%)	Avg. RMSE (%)	Avg. MSE (%)
Ahmedabad	Training	99.93	0.14	0.0366	0.000135
	Testing	99.78	0.21	0.0576	0.000332
Dhaka	Training	99.86	0.14	0.0244	0.000059
	Testing	99.74	0.26	0.0454	0.000206
Khulna	Training	99.76	0.24	0.0419	0.000175
	Testing	99.66	0.34	0.0593	0.000352

Table 2 shows the results for different ACOSIP model of each station. The model of Ahmedabad, India region found to be best among all the models.

The effectiveness of FS algorithm and the prediction ability of ANN have been ascertained using the data for Ahmedabad region. In this respect the 15 independent runs for ACOSIP model for Ahmedabad have been performed and then averaged the results – shown in Table 3. As Table 3, the PA and MAPE for training and testing case are 99.84%, 0.16% and 99.70%, 0.3%, respectively. And as the value of R is higher (99.90%), so it can be easily understood that the ACOSIP model has stronger prediction capability.

Table 3. Performance of ACOSIP model and results were averaged over 15 independent runs.

Local Investigated Data for Ahmedabad Region For 15 Independent Run					
DATA	Avg. PA (%)	Avg. MAPE (%)	Avg. RMSE (%)	Avg. MSE (%)	Avg. R (%)
Training	99.88	0.13	0.027941	0.000781	99.94
Testing	99.78	0.25	0.052889	0.002745	99.88

Table 4 depicts the positive effect of feature selection strategy. PA is 99.74% with selected features. Determination of the best solution in ACOSIP can be seen in Fig. 2, it can be showed that, the PAs varied with the size of those subsets and for the

performances of the local best subsets. The PAs were maximized in various points, but the best solution was selected (indicated by circle) by considering the reduced size subset. How the selection of salient features in different iterations progresses in ACOSIP, Fig. 3 and Fig. 4 shows the scenario of such information for a single run. We can see that, features 1, 2, 12, 7, 8 and 4 received maximum values of τ and η so these features have a higher priority of selection. Ultimately, a successful evaluation function NN model is used to find high- quality solutions for ACOSIP in prediction. As training process progresses and converges to a certain limit up to 40 epochs (Fig. 5(a)). However, the training is terminated, because at this instant validation error increases due to over training (Fig. 5(b)). At the termination epoch we also observed that the number of hidden neuron is 3. Table 5 shows the correlation information that the selected features are highly uncorrelated.

Table 4. Performance (averaged) of ACOSIP model for feature selection over 15 independent runs.

Avg. Result for all features				Avg. Result for selected features			
N	SD	PA (%)	SD	ns	SD	PA (%)	SD
12	0.00	95.5	1.12	6.13	0.82	99.74	0.40

Table 5. Correlation information (CI) of selected features

Features	Ln.	Lt.	DL	Rh	Mat	Mit
Correlation	4.8	5.91	5.83	7.6	5.96	8.96
Less Correlated	95.2	95.09	90.17	92.4	94.04	91.04

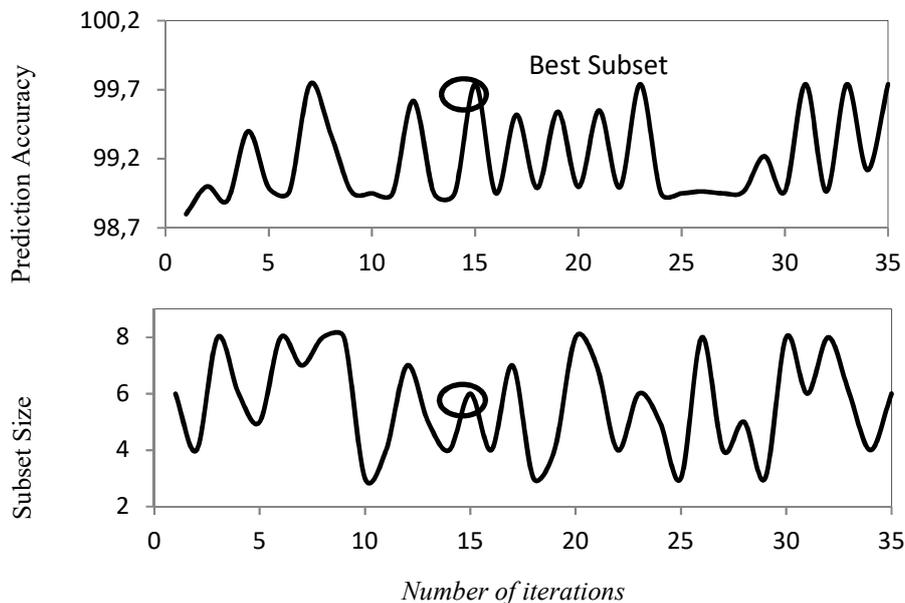


Fig. 2. Finding best subset of the features for a single run.

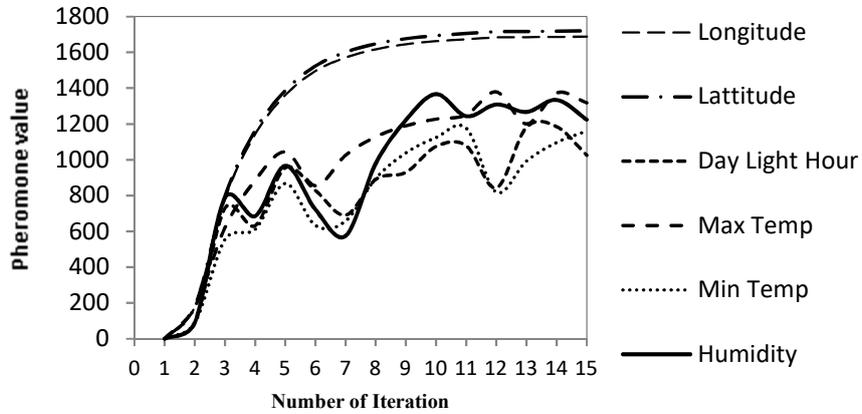


Fig. 3. Distribution of pheromone level in different iterations for a single run.

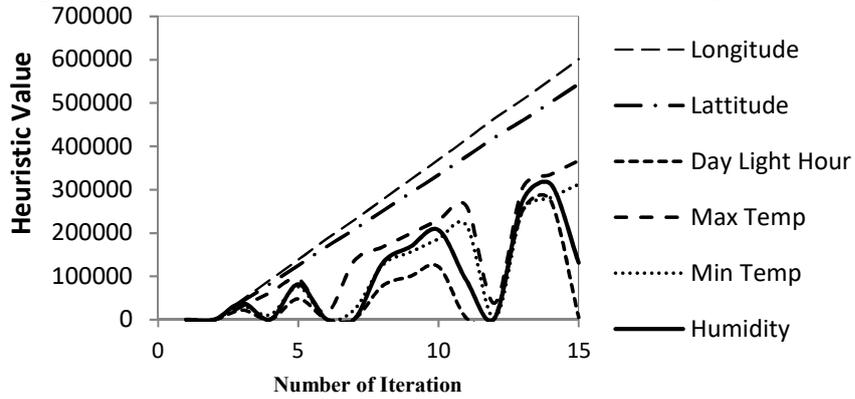


Fig. 4. Distribution of heuristic level in different iterations for a single run.

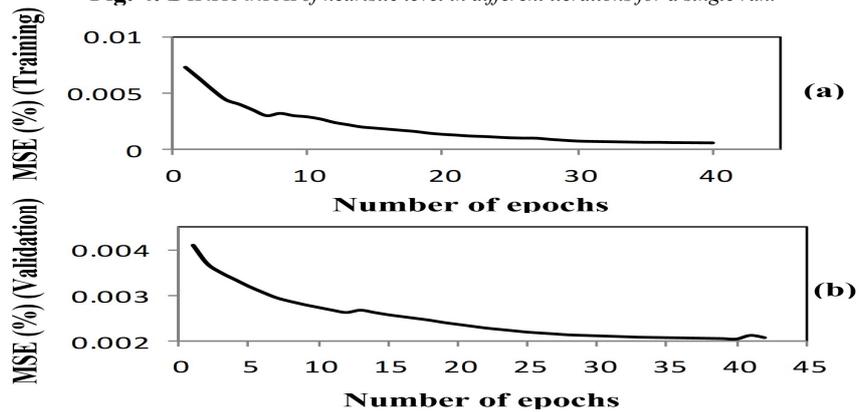


Fig. 5. Training process to evaluate the constructed subsets by ants: (a) training error, (b) validation error.

It can be seen from Fig. 6 that the high correlation is present between actual and predicted data.

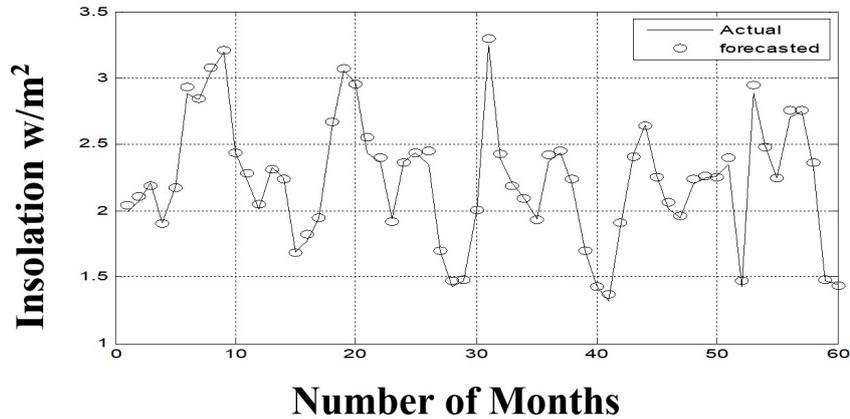


Fig. 6. Actual and predicted insolation curve

7 Comparison with Other Works.

Theoretically, a huge number of attempts have been carried out in designing solar radiation prediction models. The proposed model is quite different from those discussed in [3 – 9], where the prediction models use the features, based on knowledge or availability with no attempts to search the best feature subset. From Table 6 it can be seen that, the proposed ACOSIP perform better with feature selection strategy.

Table 6. Comparison with other works

Name of Authors	MSE (%)	PA (%)	MAPE (%)
Proposed Model for Ahmedabad	0.000459	99.94	0.11
Md. Shahid Iqbal [15]	0.000598	99.86	0.14
Amit Kumar Yadav [6]	0.0024		0.2783
Shafiqur Rahman [7]			4.49
Salman Quaiyum [8]	0.29		

8 Conclusion

This work represents a new model (ACOSIP) for predicting solar irradiance using ACO and NN. In ACOSIP, the most salient climatological features are selected to improve SIP accuracy. To facilitate the activity of ACOSIP, two sets of new rules have been designed for pheromone update and heuristic value measurement, where the correlation information technique and the outcome of NN are embedded. Experiment results show that, ACOSIP performs well in 12 solar radiation data samples collected from Meteorological Centre, Ahmedabad where it selected six most salient features easily with better prediction accuracy. The selected features are longitude, latitude, day light hour, max temp., min temp., and humidity. Furthermore, the averaged prediction accuracy of ACOSIP for Ahmedabad in testing case is 99.78% including the MAPE of 0.21% (Table 2). The proposed ACOSIP also represents high correlation of 99.93% (Table 1) in between the actual and forecasted data. On the other hand, the results of the low standard deviations of prediction accuracies exhibit the robustness of this model. Comparison results presented in Table 6 show that, ACOSIP can perform well comparing to other existing models. Thus, it can be said that, the proposed ACOSIP model may have capability to select the relevant features and to predict solar radiation at any location in the world, provided that the proper data from the locations are available.

References

1. EPIA Renewable Energy House, "European Photovoltaic Industry Association (EPIA): Market Report 2011," Belgium, 2012. Available online: <http://www.epia.org/publications/epia-publications.html> (accessed on 22 November 2011).
2. European Commission Joint Research Centre, "Solar Cell Production and Market Implementation of Photovoltaics: PV Status Report," *Publications Office of the European Union*, Italy, 2011. Available online: <http://ie.jrc.ec.europa.eu> (accessed on 22 November 2011).
3. S. A. Kalogirou, "Artificial neural networks in renewable energy systems applications: a review," *Renewable and Sustainable Energy Review*, vol. 5(4), pp. 373-401, 2001.
4. A. Mellit, S. A. Kalogirou, L. Hontoria, and S. Shaari, "Artificial intelligence techniques for sizing photovoltaic systems: a review," *Renewable and Sustainable Energy Reviews*, vol. 13, pp. 406-19, 2009.
5. M. Mohandas, S. Rehman, and T.O. Halawani, "Estimation of global solar radiation using artificial neural networks," *Renew Energy*, vol. 14, pp. 179-184, 1998.
6. A.K. Yadav, S.S. Chandel, "Artificial Neural Network based Prediction of Solar Radiation for Indian Stations," *International Journal of Computer Applications*, vol. 50, 2012.
7. S. Rahman, M. Mohandes, "Artificial neural network estimation of global solar radiation using air temperature and relative humidity," *Energy Policy*, vol. 36, pp. 571 – 576, 2008.
8. S. Quaiyum, S. Rahman, "Application of Artificial Neural Network in Forecasting Solar Irradiance and Sizing of Photovoltaic Cell for Standalone Systems in Bangladesh," *Int. Journal of Comp. App.*, vol. 32, pp. 51-56, 2011.
9. A. Hasni, A. Sehli, B. Draoui, A. Bassou, and B. Amieur, "Estimating global solar radiation using artificial neural network and climate data in the south- western region of Algeria," *Energy Procedia*, vol. 18, pp. 531-7, 2012.

10. M.M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Systems with Applications*, vol. 39, pp. 3747–3763, 2012.
11. A. Will, J. Bustos, M. Bocco, J. Gotaya, and C. Lamelas, "On the use of niching genetic algorithms for variable selection in solar radiation estimation," *Renewable Energy*, vol. 5, pp. 168-76, 2011.
12. L. Ke, Z. Feng, and Z. Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory," *Pattern Recognition Letters*, vol. 29, pp. 1351-1357, 2008.
13. M. Dorigo, C. Blum, "Ant colony optimization theory: A survey," *Theoretical Computer Science*, pp. 243-278, 2005.
14. M.L. Gambardella, M. Dorigo, "Ant-Q: A reinforcement learning approach to the TSP," *Proceedings of the 12th international conference on machine learning*, 1995, pp. 252-260
15. S. Iqbal, M. Kabir, H. M. I. Hasan, M. J. A. Soeb, A. Mishkat, V. Ray, "A new prediction model for solar irradiance using ant colony optimization and neural network," *2nd International Conference on Electrical Information and Communication Technologies (EICT)*, 2015.

Determination of energy losses in distribution transformers using a compensation algorithm in energy meters.

Paul Cando
Universidad Politécnica Salesiana
pcando@est.ups.edu.ec

Pablo Mendez
Universidad Politécnica Salesiana
pablo.mendez@centrosur.gob.ec

Juan Maldonado
Universidad Politécnica Salesiana
jmalonado@est.ups.edu.ec

Carlos Álvarez
Universidad Politécnica de Valencia
calvarez@die.upv.es

Marco Toledo
Universidad Politécnica de Valencia
Universidad Católica de Cuenca
martoor@doctor.upv.es

Diego Morales
Universidad Católica de Cuenca
dmoralesj@ucacue.edu.ec

Abstract— In the present study, a technical and economic analysis is carried out to determine the losses in the distribution transformer stage, taking into account that this equipment produces the greatest amount of technical losses in the distribution system. The energy dissipated in those machines is not invoiced by the commercial systems since the measurement systems usually used are not capable of registering such losses, representing a technical and economic deficit for the distributors, by means of the construction of the compensation algorithm in measurement systems located on the low voltage side and the feasibility of changing the meters currently used by other whose capabilities allow the compensation of such losses.

Keywords— *Direct and Combined Measurement, Losses, Load Factor, Distribution Utilities, Distribution Transformers.*

I. INTRODUCTION

In the electrical distribution networks power losses and energy is obtained from a calculation process that covers a large number of design parameters and variables of operational status. Most of these losses are located in the main components of the distribution power grids, such as transformers and lines. The power losses in the lines grow as their length increases and the amount of energy they transport, are variable with respect to time and dependent on the configuration of the distribution system and the active and reactive power flows required by consumers. On the other hand, it is known that the losses of power in the transformers depend on the materials with which these are constructed, and that the degradation of these materials can be a great influence in the increase of these losses in the course of time.

The current research focuses its study on Distribution Transformers -DT-, since technically exists two types of losses those of iron and copper, reducing the performance of the transformer causing losses. In this sense the Utilities Distributors -Utility-; among these processes perform technical reviews of each one of the transformers that will be installed in their system in order to contrast the information obtained of the factory protocols; in order to meet the national and international standards. However, the losses that were measured in the laboratory are not constant in time, and their behavior depends on various variables and operating conditions.

In order to contribute to the reduction of the technical losses, the measuring systems have been evolving continuously; in such a way that currently there are electronic meters, whose capacities allow to measure the power and energy that is lost in the transport and transformation stages, through loss-compensation algorithms.

The decision for using compensation in the measurement may be influenced by local rules and practices, the disposition of the substation and the ability to obtain contractual agreements with customers, the compensation of losses through the systems of measurement provides a technical-economical means to store the energy and power of customers measured in low voltage, dynamically compensating the losses of the transformer in the energy meter.

Finally, the study after determine technical losses that produce the -DT, quantifies the investment that should be made by the -Utility to count with a correct measurement, contributing to the improve incomings from the sale of energy.

A. Losses in Distribution Transformers.

The distribution transformers are electric machines that reduce the power and voltage to levels suitable for final use, essential in the distribution stage of a commercial electric power company, so the research will use the data of the measurements and test protocols of the Utility CENTROSUR, located in the city of Cuenca – Ecuador.

With an update till October 2017, the total losses were in the order of 6.93%, 5.94% correspond to the technical losses and 0.99% to the non-technical of a total of energy available of 94,798.95 MWh [1]. The percentages of losses per functional stage in the Distribution Utility are presented in the figure 1. It can be observed that the highest percentage of losses is in the stage of transformers with a 2.13%, so, any action aimed at mitigating these losses and improve the measurement in this stage is mandatory for the Utility [1].

However, the losses in the supply chain specifically in the stages of energy distribution are considerable, so their values are indicated below, in subtransmission lines the value corresponds to 0.56%, substations 0.41%, primary feeders 0.91%, low voltage lines 1.13%, public lighting 0.16%, connections 0.08%, meters 0.29% and no technical losses 0.98%, as shown in Figure 1, also in green color are the representative technical losses produced by the -DT with 2.13%.

This work was supported by Universidad Politécnica de Valencia, Smart Grid Research Group of the Universidad Católica de Cuenca and Universidad Politécnica Salesiana of Cuenca - Ecuador.

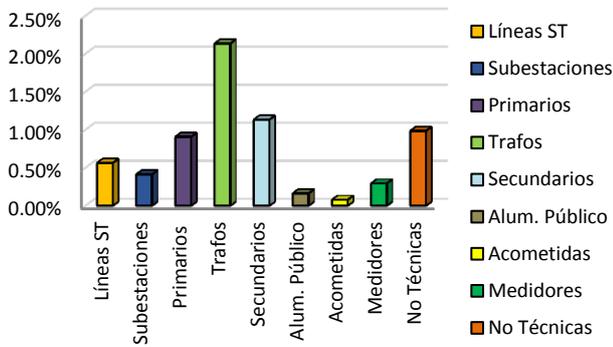


Fig. 1. Percentage of losses per functional stage

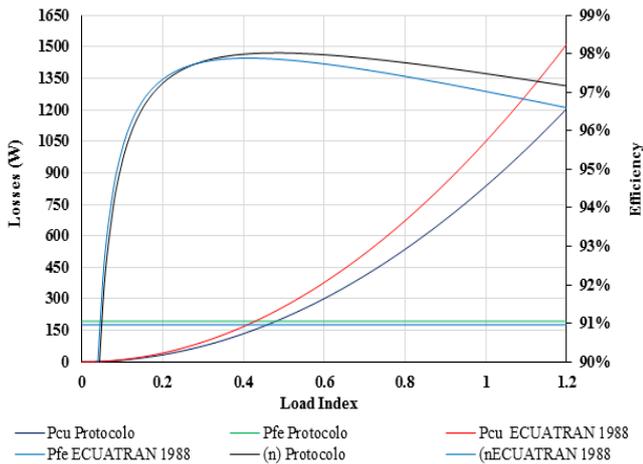


Fig. 2. Transformer 50 KVA - year 1988.

Figure 2 shows the losses in copper, iron and performance presented by a – DT, after 25 years of be operating in an energy distribution system.

B. Existing Measurement Systems.

The measurement systems are classified in i) Direct Measurement (without measurement transformers), ii) Indirect Measurement (with current and voltage transformers) and iii) Combined Measurement (only with current Transformers). Table I shows the types of measurement systems according to the nominal power of the most commonly used in transformation stations in the distribution companies.

TABLE I. TYPE OF MEASUREMENT ACCORDING TO THE NOMINAL POWER OF THE TRANSFORMER.

Power (kVA)	System type	Measuring System	Type energy meter
0-50	SinglePhase	Direct	FM 2S-CL 200 A
50-100	SinglePhase	Combined (BV)	FM 4S-CL 20 A
>100	SinglePhase	Indirect (MV)	FM 4S-CL 20 A
0-30	Three phase	Direct	FM 16A-CL 100 A
30-75	Three phase	Direct	FM 16S-CL 200 A
75-192,5	Three phase	Combined (BV)	FM 10A-CL 20 A
>192,5	Three phase	Indirecto (MV)	FM 10A-CL 20 A

This research is very important in the direct measurement and Combined (low voltage with the use only of transformers of current in the measurement system). This doesn't register the energy consumed by the – DT, and these

losses are transferred to the utility, increasing the technical losses in the supply chain. Table II shows the number of measuring equipment installed in the Utility.

TABLE II. NUMBER OF MEASUREMENT SYSTEMS IN THE –DU.

Measurement Systems	Power of the Transformer (kVA)	Amount of Measurement Systems
Direct	3 up to 75	3.492
Combined	75 up to 192,5	2.419
Indirect	above 200	118

II. DATA ANALYSIS

A. Calculation of Sample Size for the Case Study.

Within the research to determine the sample size, we studied the population of transformers, making use of the database that provides the Geographic Information System of the -DU. Table III shows the number of transformers in the population stratified by powers.

TABLE III. THREE-PHASE DISTRIBUTION TRANSFORMERS WITH DIRECT AND COMBINED MEASUREMENT SYSTEMS (30 -192,5 KVA).

Nominal Power (KVA)	Amount of Transformers		Type of Measurement System
	Utility	Private	
30	354	304	Direct
45	219	76	Direct
50	600	229	Direct
60	126	45	Direct
75	423	175	Combined
100	188	126	Combined
112,5	17	16	Combined
125	4	30	Combined
150	9	33	Combined
160	6	28	Combined
175	0	7	Combined
190	1	7	Combined
192	0	8	Combined
192,5	0	15	Combined
Total	1.947	1.099	

Subsequently, defines the size of the sample and performs a proportional allocation on the basis of the participation of the selected groups. From equation (1) it is calculated for a finite population and known according to the data obtained, it will be considered a confidence level of 96% ($k = 2.04$) with an error of 6%, the values of P and Q selected are 50%, because it represents a larger sample size [3].

$$n = \frac{k^2 * N * p * q}{e^2(N - 1) + k^2 * p * q} \quad (1)$$

Where:

K = constant depends on the level of trust assigned

N = population size

p = probability of success

q = probability of failure

e = sample error (can be assumed from 1% to 10%).

Table IV shows the results obtained from applying the calculation of the sample in each group of – DT of the studied universe of the-DU.

TABLE IV. SUMMARY OF THE SAMPLING AND INFORMATION GATHERING.

Stratum	Population	Sample size	Samples Obtained
Utility	1.947	171	173
Private	1.099	96	125
Total	3.046	267	298

B. Load Analysis at Distribution Transformer.

The electrical charge of the transformer will be the variable of influence in the study, since regardless of the rate of the consumer or the power of the transformer, both the load factor and the use factor, indicate the percentage of growth or decrease of the electrical losses on these machines. Of the 298 load profiles obtained, the load factor and the use factor were calculated according to the nominal power of the transformer, from these factors the sample was grouped within six strata. Table V shows the segmentation performed and the size of the population for each stratum.

TABLE V. THE SAMPLE LAYER DEPENDING ON THE LOAD.

Estratum	Load Factor	Utilization Factor	Amount of Transformer – DU	Amount of Transformer Private
I	0 - 0,55	0 - 0,30	10	36
II	0 - 0,55	0,30 - 0,6	9	25
III	0 - 0,55	0,6 - 1	1	2
IV	0,55 – 1	0 - 0,30	26	45
V	0,55 - 1	0,30 - 0,6	79	13
VI	0,55 – 1	0,6 - 1	48	4

The stratified sample according to the load, allows us to visualize widely the spectrum of the operation of the transformers belonging to the -DU and the private transformers. The following figure shows the dispersion of transformers depending on the use and variability of the load.

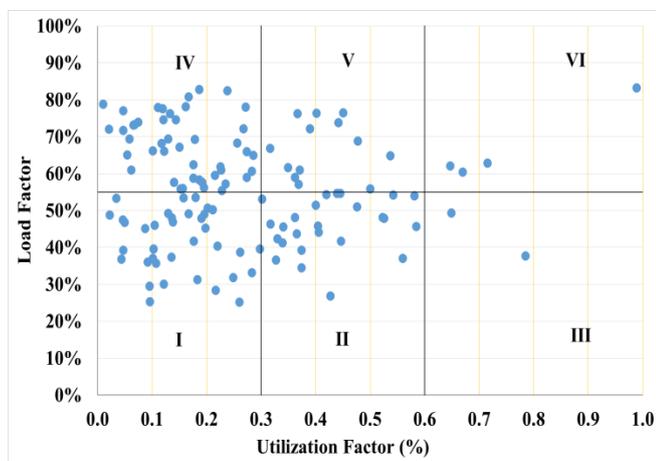


Fig. 3. Dispersion of private transformers according to the load.

-According to the use factor, 64.8% of the transformers studied are oversized (stratum I and IV), 30.4% use the load in an acceptable manner (stratum II and V) and only 4.8% are correctly dimensioned (stratum III and IV).

-Depending on the load factor, considering as slightly constant loads which are higher than 55% and as variable loads the lowest or equal to the same value, it is observed that 52% of the transformers are operating with highly variable loads (Stratum I, II, III) and 58, 5% with loads close to the unit (stratum IV, V, VI).

-The stratum that possibly represents higher percentages of losses with respect to the measured energy, is stratum I; this deduction originates from the low percentages of utilization of the installed capacity, which generates increments in the percentages of iron losses due to equipment oversizing.

For the Distribution Utility owned transformers, the same analysis is performed, from which transformers are dispersed as shown in Figure 3.

-The 20,81 % of transformers analyzed are oversized (stratum I and IV) the 50,86 % have a charging factor acceptable (stratum II and V) and 28,33% are sized correctly (stratum III and VI).

-The 88.44% of transformers have a load factor exceeding 55% that is with a slightly variable load.

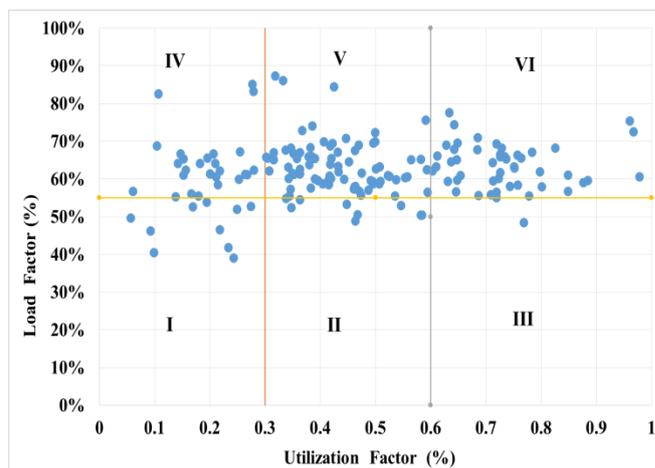


Fig. 4. Dispersion of the – DU transformers according to the load.

From the analysis, it is obtained that the transformers of the-DU are with a use factor on the 55%, this indicates that the dimensioning of the load in the designs and construction had less uncertainty in the projection of the demand. In addition it is worth indicate that private transformers supplies energy to undifferentiated loads (Similar energy uses), while the -DU transformers. It shares different types of loads, such as residential, industrial, commercial, public lighting, among others.

III. ALGORITHM FOR LOSS COMPENSATION IN SOLID-STATE METERS

A. Mathematical Modeling

The few existing literature, published and related to the algorithms for the compensation of losses in energy meters, indicates that there are two methods of calculation of losses in the transformers.

1. Direct modeling based on the loss of the test protocol.

2. Modeling depending on percentage constants.

The deficiency of information reflects that manufacturers of power meters zealously save the information of its design and algorithm in order to avoid being cloned by competition, being nature of construction firms of technology as well as traders. However, this research does not refer to proprietary marks and uses owned accountants from the Distribution utility.

Mathematical Model 1-direct modeling based on the losses determined by the test protocol.

a) *Parameters for the calculation of the losses in the Distribution Transformers.*

This mathematical model used in solid-state meters, occupies the parameters of the factory test protocol and/or the one performed in the Distribution Utility laboratory are included in the energy meter such parameters of the transformer and the subtransmission line connecting the load if necessary [2], [3]:

- Nominal voltage of -DT (V_{TXtest} on the measured side)
- Nominal power of the transformer
- Relation of transformation of the transformer of power
- Losses in the iron or in gap
- Losses in the copper or to full load
- Percentage of current of excitement
- Percentage of impedance
- Length of the line in the side of the load and of the line of supply
- Resistance and reactance for length unit for both lines
- Relation of the transformation of the CT's and PT's (VTR,CTR).
- Information about the location of the meter with respect to -DT, the supply line and the load line.

The manufacturer usually provides this information in the transformer test protocol [4]:

- Nominal power value in VA (VA_{TXtest}).
- Losses in the iron at the nominal voltage ($LWFe_{TXtest}$) in watts.
- Losses in copper at 75°C or 85°C at full load ($LWCu_{TXtest}$) in watts.
- Nominal voltage in the primary and secondary (V_{n1}), (V_{n2}) in volts.

If power factor tests are performed on the low voltage side or if the reactive power losses need compensation, the following additional information is required [2]:

- Percentage of excitation current at rated voltage ($\%I_{exc}$).
- Percentage of impedance at 75°C or 85°C ($\%Z_{cc}$).

The losses reactive without load ($LVFe_{TXtest}$) and at full load ($LVCu_{TXtest}$) may not be provided but are calculated from the above data [1] [6] [7].

$$LVFe_{TXtest} = \sqrt{\left(VA_{TXtest} * \frac{\%I_{exc}}{100} \right)^2 - (LWFe_{TXtest})^2} \quad (2)$$

$$LVCu_{TXtest} = \sqrt{\left(VA_{TXtest} * \frac{\%Z_{cc}}{100} \right)^2 - (LWFe_{TXtest})^2} \quad (3)$$

In order to determine the real losses of the transformer, the test losses must be scaled for use at the current operating voltage and current. [6][7]

$$LWFe = LWFe_{TXtest} * \left(\frac{V_{actual}}{V_{TXtest}} \right)^2 \quad (4)$$

$$LVFe = LVFe_{TXtest} * \left(\frac{V_{actual}}{V_{TXtest}} \right)^4 \quad (5)$$

$$LWCu = LWCu_{TXtest} * \left(\frac{I_{actual}}{I_{TXtest}} \right)^2 \quad (6)$$

$$LVCu = LVCu_{TXtest} * \left(\frac{I_{actual}}{I_{TXtest}} \right)^2 \quad (7)$$

Where:

V_{TXtest} = Voltage line to line of the-DT in the side of the measurement, often the nominal Voltage and the Voltage of test are the same.

I_{TXtest} = Nominal current of the - DT on the measurement side.

V_{actual} = Load Voltage.

I_{actual} = Current in the load.

If the actual voltages and currents are the voltages and the secondary currents of the measuring transformers, as seen on the meter, then these readings need to be scaled by the current transformer ratio (CTR) or the ratio of Voltage Transformer (VTR) to give the values of the primary circuit [7]:

$$LWFe = LWFe_{TXtest} * \left(\frac{V_{tm2} * VTR}{V_{TXtest}} \right)^2 \quad (8)$$

$$LWCu = LWCu_{TXtest} * \left(\frac{I_{tm2} * CTR}{I_{TXtest}} \right)^2 \quad (9)$$

$$LVFe = LVFe_{TXtest} * \left(\frac{V_{tm2} * VTR}{V_{TXtest}} \right)^4 \quad (10)$$

$$LVCu = LVCu_{TXtest} * \left(\frac{I_{tm2} * CTR}{I_{TXtest}} \right)^2 \quad (11)$$

Where:

V_{tm2} = Secondary voltage of potential transformer.

I_{tm2} = Secondary current of the measuring transformer.

Mathematical Model 2.

B. Modeling based on constant percentages.

In this method the calculation parameters of the losses in the line and in the -DT are calculated manually or in a third party program; the result of analysis are four constants. These values are entered into the energy meter algorithm. The meter uses these constants to calculate the losses and to perform the compensation, for which the following data must be programmed in the measuring equipment [7]:

- Percentage of losses in iron in Watts (% $LWFe$)^{*}.
- Percentage of losses in copper in Watts (% $LWCu$)^{*}.
- Percentage of losses in the iron in VAR (% $LVFe$)^{*}.
- Percentage of losses in copper in VAR (% $LVCu$)^{*}.
- Relation of the transformation of the transformers instruments (VTR) and (CTR).
- Nominal input voltage of the meter (V_{Mrated}).
- Nominal input current of the meter (I_{Mrated}).
- Number of meter elements (2 for connections Delta, 3 for WYE).

An important fact for the programmer of the meter of energy is, if it is desired that the losses should be reduced of the delivered energy, negative values must be introduced for the constants of percentage loss [7].

- Calculations of the constants of losses in percentage

$$\%LWFE = \frac{LWFe_{TXtest}}{\text{Potencia en el primario (VA)}} * 100\% \quad (12)$$

$$\%LVFE = \frac{LVFe_{TXtest}}{\text{Potencia en el primario (VA)}} * 100\% \quad (13)$$

$$\%LWCu = \frac{LWCu_{TXtest}}{\text{Potencia en el primario (VA)}} * 100\% \quad (14)$$

$$\%LVCu = \frac{LVCu_{TXtest}}{\text{Potencia en el primario (VA)}} * 100\% \quad (15)$$

- Loss calculations in the transformer using constant percentage

$$LWFe = \frac{\%LWFe}{100} * \text{Pot. prima (VA)} * \left[\frac{V_{actual}}{V_{Mrated} * VTR} \right]^2 \quad (16)$$

$$LVFe = \frac{\%LVFe}{100} * \text{Pot. prima (VA)} * \left[\frac{V_{actual}}{V_{Mrated} * VTR} \right]^2 \quad (17)$$

$$LWFe = \frac{\%LWCu}{100} * \text{Pot. prima (VA)} * \left[\frac{I_{actual}}{I_{Mrated} * CTR} \right]^2 \quad (18)$$

$$LWFe = \frac{\%LVCu}{100} * \text{Pot. prima (VA)} * \left[\frac{I_{actual}}{I_{Mrated} * CTR} \right]^2 \quad (19)$$

C. Simulation in Gui-Matlab

The algorithm requires input arguments and a specific configuration to generate results correctly, which makes their use restricted only to users who are aware of their operation, for this reason they are done using the graphical interface in language MATLAB facilitating handling the same.

The application must import the Excel load profile with reactive and active demand data at 10-and 15-minute intervals, these values will be treated internally allowing the following results to be obtained:

- Load curve without applying the compensation
- Load curve applying a 2% surcharge for losses in Transformers, as indicated in the current tariff sheet (relative to Ecuador).
- Load curve obtained with the implementation of the compensation algorithm.
- Energy compensated in [kWh].
- Percentages of energy not invoiced.
- Power factor before and after compensation.

It should be noted that in the statement tariff Ecuador there is a fixed increment of 2% in the total consumption of energy that is called compensation for losses in the transformer, which is the reason why it is necessary that this parameter is considered in the simulation, for what was included in the test mode with the purpose of comparing the results obtained with the tests performed in the laboratory. The input data requested by the program are the test protocol values submitted by the manufacturer.

TABLE VI. SCRIPT FOR THE MATLAB SOFTWARE FOR TESTING THE METER

```

nombreDeARCHIVO = 'Ingreso de datos.xlsx';
Hoja = 1;
xlRange13 = 'H7:H102'; xlRange14 = 'I7:I102';
xlRange15 = 'J7:J102'

%=== Extracción de valores cada 15 minutos

p= Xlsread ('Ingreso de datos.xlsx',
Iact1=(P1*1000)/((cos(atan2(Q1,P1)))*(Vtest*sq
rt(3)))

LwcupP1=((LWCu*((Iact1/Itest)^2))/1000)+P1+((L
WFe*((Vactual/(Vtest/sqrt(3)))^2))/1000);

LvcuP1=((LVCu*((Iact1/Itest)^2))/1000)+Q1+((L
vFe*((Vactual/(Vtest/sqrt(3)))^4))/1000);

hoja, xlRange13);
T= Xlsread ('Ingreso de datos.xlsx', hoja,
xlRange15);
Q= Xlsread ('Ingreso de datos.xlsx', hoja,
xlRange14);

```

IV. RESULTS

A. Measurements with laboratory equipment and Matlab Algorithm.

The tests in the laboratory and in the energy meters with compensation of losses already installed in consumers with different uses of energy allow to contrast the results of the algorithm of the static meter with the algorithm elaborated in Gui-Matlab. When choosing the test mode in the Gui-Matlab model we will be able to enter the same additional data that is entered in the laboratory calibrators to emulate the load, the parameters introduced and the resulting contrast between the meter with loss compensation and the developed model are shown tables VI and VII respectively.

TABLE VII. PARAMETERS FOR TESTING IN THE METER LABORATORY.

Test Number	Transformer data						TC'S		Data Bench "WECO"		
	Pn (kVA)	Pfe (W)	Pcu (W)	Iexc(%)	Zcc(%)	Vll (V)	Ip (A)	Is (A)	V. Test	I. Tests	P.F.
6	50	286	1.0	1.9	3.9	220	200	5	120	0	1
2	50	286	1.0	1.9	3.9	220	200	5	120	3	1
3	50	286	1.0	1.9	3.9	220	200	5	127	3,2	1
4	100	314	1.6	1,4	3,9	220	300	5	125	4	0,5
5	100	314	1.6	1,4	3,9	220	300	5	125	2,5	0,8
6	200	632	2.5	2,6	3,2	220	300	5	127	0,5	0,5
7	200	632	2.5	2,6	3,2	220	300	5	127	0,5	1

TABLE VIII. CONTRASTING THE RESULTS OF THE TEST IN THE UTILITY LABORATORY.

N° Test	Measured Values		Calculated Values		Error (%)	
	P(W)	Q(Var)	P(W)	Q(Var)	P(W)	Q(Var)
1	256	756	255,27	755,072	0,284%	0,123%
2	44.320	2.144	44.329,26	2.141,9	0,020%	0,098%
3	51.308	2.588	51.317,87	2.605,06	0,019%	0,659%
4	46.653	82.280	46.663,18	82.302,39	0,021%	0,027%
5	45.823	36.246	45.835	36.264,05	0,026%	0,049%
6	11.950	1.265	11.960,74	1.273,47	0,089%	0,669%
7	6.282	11.080	6.246	11.172,14	0,573%	0,832%

As can be seen in table VII the errors obtained between the calculated and measured are below 1%, these results allow validate the correct functioning of the developed model.

B. Field testing and algorithm in Matlab.

The field tests confirmed the operation of the compensation algorithm in a real environment. Outstanding to customers who have systems of measurement with the algorithm for loss compensation, within this group was randomly selected to 3 of them, considering the last peak demand and depending on the use of energy.

Among the consumers chosen to perform the field tests were an Educational Unit, whose operation is only 8 hours a day, from 7am to 14pm, the remainder of the 24-hour period has no consumption, so the losses in the transformer are increased. Another consumer who joined the analysis was a "CASSTORPROCT CIA. LTDA. ". The use of energy is purely industrial, as it has mills to crush Rock and develop material for improvement of roads, working time is approximately 12 hours, from 6am to 6pm.

1) Test 1: Milenio School

According to the information obtained from the commercial system the last registered maximum demand of the transformer is in the range of 9 to 14 KW, with respect to the nominal power of the transformer it has a use factor below

10%. The school has a 200 KVA three-phase transformer; the measurement system consists of a A1830RLC type energy meter class 200, 3 current transformers with -RTC of 300/5 precision class 0.2. In order to test the model and contrast the information, in this section was installed an energy meter in series with the gauge of the – DU, so that both meters store the same currents and voltages as shown in Figure 5.

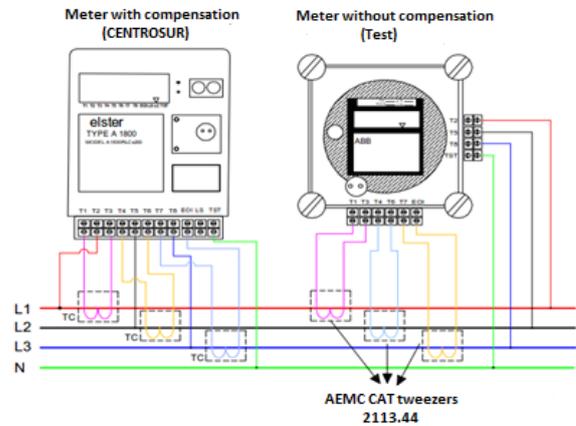


Fig. 5. Diagram of the serial connection with proof-meter.

The test was performed for 6 days, obtaining a register of 576 measurements, to compare with those obtained from the -DU meter with loss compensation during the same period.

TABLE IX. RESULT OF THE TEST 1 IN (UEM SAYAUSI)

Energy registered by the meter without compensation		Energy registered by meter with compensation		Energy calculated by the Algorithm in Gui-Matlab		Error % between compensation and the calculated	
Active (KWh)	Reactive (KVARh)	Active (KWh)	Reactive (KVARh)	Active (KWh)	Reactive (KVARh)	Active	Reactive
709	123,15	783,8	188,8	785,8	195,0	0,25%	0,24%

Subsequently, entered the profile loading without compensation the parameters of the transformer and the meter. The load profile calculated will be exported to Excel and made the comparison with the load profile obtained from the meter with compensation of the consumer (Educational Unit). The results are shown in table VIII.

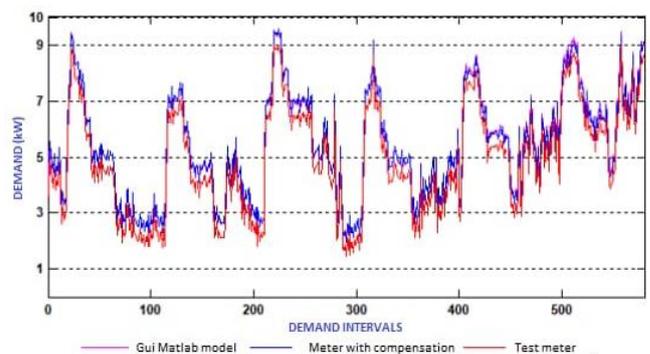


Fig. 6. Load profile of the School of "Milenio"

The errors obtained are relatively low, despite the fluctuations in load the loss-compensation meter installed in the educational unit is working within the acceptable margin in field tests allowed by the $-DU$ (-2%). The error is negligible and attributable to the current equipment used and the losses in each of the 3 elements of the meter.

2) Test 2: Industry CASSTORPROCT CIA. LTDA.

This industrial consumer operates with several $-DT$ within its work area was chosen due to its relatively low load in the 160 KVA transformers. The system of measurement installed is of type combined, has an energy meter and 3 transformers of current 300/5 with class of precision 0.5. The methodology used in this test is the same as the one discussed in the previous consumer. The test was carried out for 6 days, with the information of the load profile of the two measurement systems the same procedure was carried out in the previous test. The results are shown in table IV.

It can be observed that the energy measured by the test meter is practically the same as the industrial consumer equipment, which clearly indicates that this meter is not compensated for the transformer losses, the following figure shows the point-to-point measurements obtained.

TABLE X. RESULTS OF THE TEST 2 IN THE INDUSTRY CASSTORPROCT

Registered power test meter (without compensation)		Stored Energy by the meter compensation		Energy calculated with Gui-Matlab		Error % compensation and calculated	
Active (kWh)	Reactive (kVARh)	Active (kWh)	Reactive (kVARh)	Active (kWh)	Reactive (kVARh)	Active	Reactive
317,1	254,5	316,3	254,0	384,4	301	21,5	18

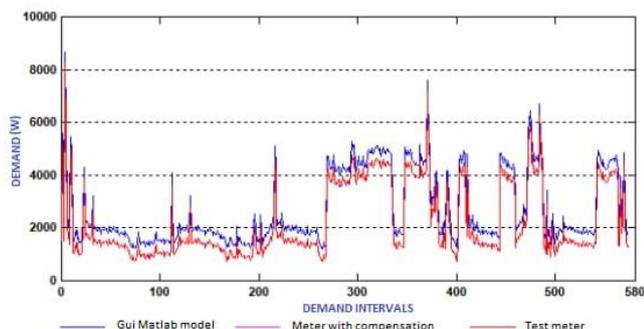


Fig. 7. Load Profile of industry CASSTORPROCT CIA. LTDA.

The load curves of the two meters are equal in each demand interval, so that the meter curve with loss compensation cannot be appreciated. Carrying out the revision of this novelty in the software of the meter indicated that the compensation module is active, so we realized that the physical compensation module presented a malfunction in the integrated circuit in charge of calculating the losses.

V. RESULTS IN DISTRIBUTION UTILITY

A. Quantification of losses in transformers

Once validated the operation of the model developed in Gui-Matlab, we proceeded to realize the quantification of the losses of energy in the private-DT studied, then shows the results obtained for 6 transformers before and After the loss compensation of transformer-CPT.

TABLE XI. RESULTS OF THE COMPENSATION IN 6 PRIVATE CONSUMERS

A	B	C	D	E	F	G
50	29,18	33,77	0,75	0,44	15,71%	I
150	429,43	439,31	0,84	0,76	2,29%	II
192,5	1.236,5	1.253,9	0,83	0,79	1,40%	III
150	146,01	157,02	0,95	0,75	7,54%	IV
50	337,21	342,39	1	0,99	1,54%	V
175	1.490	1.504	0,87	0,84	0,94%	VI

Where:

- A.- Nominal Power (kVA)
- B.- Energy without CPT (kWh)
- C.- Energy with CPT (kWh)
- D.- Power factor before CPT
- E.- Power factor after CPT
- F.- Energy not registered (%)
- G.- Estratum (I-VI)

In table X the 2%-compensation according to the tariff statement issued by the Ecuadorian regulatory body for the Utility-, in no case reflects the actual losses of the transformers studied, you can notice a similar trend, the losses increase and depending on how the use and load factor does. If an average percentage of losses are obtained according to the energy measured daily for each one of the strata analyzed, the following results are obtained.

TABLE XII. AVERAGE PERCENTAGE OF UNREGISTERED ENERGY PER STRATUM

Estrato	Percentage of unregistered losses	Power factor after compensation
I	9,36%	0,67
II	2,19%	0,86
III	1,31%	0,87
IV	5,52%	0,85
V	1,51%	0,89
VI	1,09%	0,91

The results shown in table XI corroborate what was previously analyzed, the average loss percentage shows a pronounced difference between the strata that have a use factor below 30%, in the same way it could be noticed a decrease in power factor in these groups.

The percentages of losses with respect to the registered energy increase in the strata of lower factor of use and greater variation of the load this is due to the oversizing of the installed transformers, whose constant losses in the iron can overcome the load of the transformer. Figure 9 shows the load curve of a 50 kVA transformer, how losses in iron and copper evolve during the day.

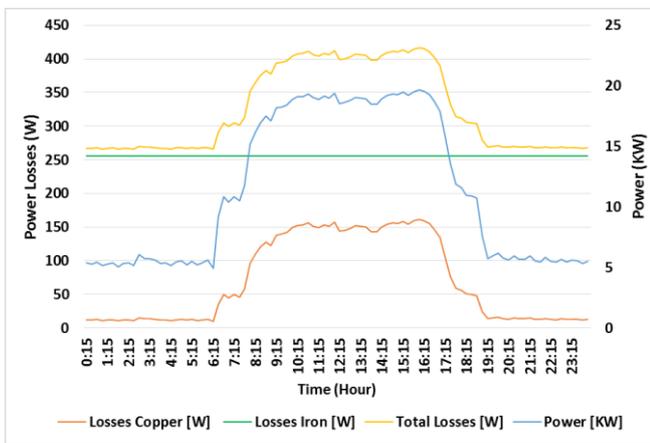


Fig. 9. Losses in Distribution Transformer and Load Curve.

The behavior of the losses in iron and copper during the operation of the distribution transformer, whose maximum use factor is close to 40%, can be concluded:

- There is no variation in the voltage, so the losses in the iron remain constant during 24 hours a day.
- Losses in copper vary according to the fluctuation of the load, observing that during the course of the day it changes from minimums to maximums, being proportional to consumption.
- Losses in iron are the dominant ones, due to the low power use factor, that is, the over-dimensioning of the transformer.
- While the transformer is maintained with utilization factors below 30%, the percentages of losses in relation to the registered demand will be much higher.

VI. ECONOMIC PROFITABILITY STUDY

The quantification of losses allowed to obtain the energy not registered by the meters installed on the side of low voltage of 125 consumers who have -DT private. The analysis will allow -DU, to take the decision to replace the existing meters in order to mitigate the energy losses in the transformation stage.

TABLE XIII. THE ECONOMIC STUDY FOR THE SAMPLE

Total Expenses	Total initial investment	125 electronic meters, 4 wires, class 20, form 10 A with active energy, reactive energy, demand, load profile and energy quality, with loss compensation (745,48 USD)	93.185
		Labor installation (120 USD)	15
		Initial Investment	108.185
Total Revenue	Energy price not registered	Battery change every 8 years (30 USD)	3.75
		Daily (USD)	102,92
		Monthly (USD)	3.087,59
		Annual (USD)	37.051,03

The economic study is projected for 15 years, which is the useful life of the solid state meters with compensation of losses, with a interest of 12% that is the current rate. In table

XII, the income and expenses presented in the economic study are shown, for the calculation of the income the average energy cost is taken at a value of 0.10 cUSD/kWh. From the results of the cash flow you will find the values corresponding to the NPV, IRR, C/B and recovery time, the results are shown in the table XV.

While, in contrast to the transformer's primary measurement, the low voltage measurement with loss compensation yields great advantages from a technical and economic point of view, this due to the ease of installation and the low Costs that are generated in comparison to a measurement with the use of compact transformers of measure on the side of medium voltage, in table XIII the two systems of measurement are compared from the economic point of view.

TABLE XIV. ECONOMIC COMPARISON BETWEEN HIGH VOLTAGE MEASUREMENT AND LOW VOLTAGE MEASUREMENT

Medium Voltage Measurement System		Low Voltage Measurement System with Loss Compensation	
Compact Transformers for Measurement in HV	USD 7.000	Transformers of Measurement in LV	USD 210
Installation	USD 200	Installation	USD 120
Meter without Compensation	USD 350	Meter with Compensation	USD 745,48
Total	USD 7.550	Total	USD 1.075,48

The measurement in low voltage is cheaper than a measurement in the transformer's primary, the two methods would register correctly, which would benefit both the consumer and the Distributor, because at all times it would be measuring properly, However the initial investment for a medium-voltage side measurement system would be approximately 7 times more expensive than installing counters with loss-compensation algorithms in the transformer's secondary.

According to the regulation on measurement systems in Ecuador, the consumer is responsible for acquiring the necessary equipment for the measuring system (such as CT's and TP's), except for the energy meter; Therefore, the best decision is to use low voltage measurement through counters that compensate for energy losses in the processing station.

TABLE XV. RESULTS OF PROFITABILITY INDICATORS.

Profitability Indicator	Result	Decision to take
Net present value (VAN)	USD 142.649,24	The project can be accepted by obtaining a VAN greater than zero
Internal rate of return (TIR)	35%	IRR > discount rate so you should invest
Benefit Cost (B/C)	2,3	B/C >1 therefore the investment is profitable
Recovery Time	2,92 años	The recovery period is short in relation to the lifespan of the meters, so it is profitable to invest

The results can be approximated for the entire population studied, taking into account that this is confirmed by private customers with three-phase transformation station measured in low voltage (30 kVA a 192,5 kVA), the population is 1,099 -DT and the total energy not stored for losses in these electric machines would be 271.46 MWh/month.

VII. REDUCTION OF INDEX OF TECHNICAL LOSSES IN THE UTILITY

With cut to October of the 2017 the-DU presents the following percentage of losses with respect to the available energy.

The additional monthly energy that could be measured using meters with low voltage loss compensation would be approximately 271.46 MWh, which would imply a reduction in the technical loss index of 0.27%, as indicated in the following boards.

TABLE XVI. TOTAL SYSTEM LOSSES -DU.

Available energy (MWh)	Technical losses in -DT		Total Technical Losses		Total Losses	
	%	MWh	%	MWh	%	MWh
97.411,07	2,13	2.074,02	5,67	5.523,31	6,65	6.480,33

TABLE XVII. EXPECTED LOSS RATE - AFTER ANALYSIS.

Energy available (MWh)	Losses in -DT		Losses Total technical		Total losses	
	%	MWh	%	MWh	%	MWh
97.411,07	1,85	1.802,55	5,39	5.251,84	6,37	6.208,86

VIII. CONCLUSIONS

Technical losses in the – DU show a percentage of more than 6%, with the phase of – DT which contribute most to the total losses within the distribution system. According to the loss compensation analysis made to consumers with three-phase DT with low voltage measurement (30 a 192.5 kVA). This index could be reduced by 0.27% if the currently used meters are replaced by meters with a loss compensation module that allows to store and invoice the energy dissipated in the transformer

The study of electrical losses in the distribution transformers according to the operational ageing of the transformer requires a series of operational parameters; environmental conditions and constructive characteristics of each machine, so it is difficult to know or find a factor that represents the losses of electricity, since the magnitudes are dynamic in the time.

Use of the power use factor for stratification makes it possible to obtain a better idea on the use of electrical installations of both the-DU and consumer, especially in the face of the massive use of loads that absorb current only for a small part of the time period of the daily load curve. The

application of these loads leads to the distribution systems suffering from high levels of voltage variations, as well as have a decrease in the power capacity of the electrical system, resulting in a condition to the quality of the technical product and its dynamic performance.

According to the economic study, the investment for the acquisition of energy meters with loss compensation could be recovered over a period of 3 years, and since the useful life of these meters is forecast by the manufacturers for a period of 15 years, this investment would have a high efficiency and low risk of damage to the equipment

Compensated meters have an equal accuracy and even greater than the measurement in the primary of the-DT, discarding the use of CT's and PT's in measuring systems in medium voltage (22kV).

IX. RECOMENDATIONS

Due to the high efficiency of the gauges with loss compensation and the high costs that would be avoided when placing a measurement in the primary of the distribution transformers with the use of compact transformers of measurement, which can be up to 7 Times more expensive than loss-compensation meters, the – DU is recommended to perform the replacement of all the meters used to measure customers with private processing station measured in low voltage by meters with compensation, and in this way to reduce the current index of technical losses in a more affordable and equally efficient way.

X. REFERENCES

- [1] Empresa Eléctrica Regional CENTROSUR, informes mensuales, Dirección de Comercialización, 2018.
- [2] Edison Electric Intitute, Handbook for Electricity Metering, Washinton D.C: Edinson Electric Institute, 2002.
- [3] Schneider Electric, «Transformer/Line Loss Calculations for Power Logic ION,» 2011. [En línea].Available: <https://www.schneider-electric.us/en/faqs/FA221594/>.
- [4] Empresa Eléctrica Regional CENTROSUR, informes de Laboratorio de Transformadores, Protocolos de Prueba de Fabrica, Dirección de Distribución, 2000 - 2018.
- [5] I. E. Salvado, «Tipos de Muestreo,» 2016. [En línea]. Available: <http://www.bvs.hn/Honduras/Embarazo/Tipos.de.Muestreo>.
- [6] JM.Toledo Orozco y B. Trelles Chitacápa , «Metodología para la determinación de pérdidas en instalaciones de transformación particulares en sistema de medición instalados en el lado secundario,» ECUACIER, 2016.
- [7] G. B. SCHLEICHER, «Compensating Metering in Theory and Practice,» IEEE Xplore, vol. 1, n° 1, pp. 1-2, 1933
- [8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [9] R. C. Ávila y P. Quituisaca, «Informe de evaluación de pérdidas de energía en la UTILITY,» Cuenca, 2017.
- [10] D. S. Takach y R. L. Boggavarapu, «Distribution Transformer No-Load Losses,» IEEE, Vols. %1 de %2PAS-104, n° 1, pp. 1-4, 1985.
- [11] J. C. Olivares Galván, R. Escalera Perez, P. S. Georgilakis y E. Campero, «Separation of No-Load Losses for Distribution Transformers Using Experimental Methods:Two Frecuencias and Two Temperatures,» IEEE, pp. 1-5, 2010
- [12] Centro de Excelencia Técnica, «CNS -NT -06: Sistemas de Medición de Energía,» 2015. [En línea].
- [13] G. B. SCHLEICHER, «Improvements in Transformer-Loss Compensators for Watt-hour and Var-hour Meters,» IEEE, vol. 70, pp. 1-5, 1951.
- [14] A. Hannah, «Algorithms for Computing and Programming Transformer Loss Constants in Solid-State Meters,» IEEE, pp. 1-17, 1998.

- [15] ARCONEL, «Pliego Tarifario para las Empresas Eléctricas.» ARCONEL, 2017. [En línea]. Available: <http://www.regulacionelectrica.gob.ec/tarifas-del-sector-electrico/>.
- [16] S. Krishnamoorthy y D. Jayabal, «Evaluation of Transformer Loading and Energy Loss For Increasing Energy Efficiency in Distribution System.» IEEE, pp. 1-3, 2015.
- [17] Instituto Ecuatoriano de Normalización, «NTE INEN 2116:98, Transformadores. Impedancia y pérdidas con carga.» 1998. [En línea]. Available: <https://archive.org/stream/ec.nte.2116.1998#page/n0/mode/2up>.
- [18] A. J. PETZINGER y B. E. Lenehan, «Metering Power on the Low-Voltage Side.» IEEE, vol. 62, n° 3, pp. 1-3, 1943.
- [19] S. Hasan, S. Taib, S. Hardi, A. Shukri y A. Rahim, «Core Loss Characteristics Analysis of Power Transformer Under Different Frequencies Excitation.» IEEE, vol. 1, n° 8, pp. 1-4, 2013.
- [20] S. J. Chapman, Máquinas Eléctricas, Sídney: MCGRAW-HILL, 2012.
- [21] W. Flores, E. Mombello, G. Rattá y J. A. Jardini, «Vida de transformadores de potencia sumergidos en aceite.» IEEE LATIN AMERICA TRANSACTIONS, vol. 5, n° 1, pp. 1-3, 2007.
- [22] I. Margallo Gasco, «Diagnóstico del consumo de vida de un transformador a través del análisis de compuestos furánicos.» 2012. [En línea]. Available: https://e-archivo.uc3m.es/bitstream/handle/10016/.../TFG_Isabel_Margallo_Gasco.pdf.
- [23] ELSTER, A1800 ALPHA meter, Technical manual TM42-2410E, 2009.

XI. BIBLIOGRAPHY

Paul Marcelo Cando Naula, was born in 1993 in the city of Cuenca-Ecuador. He received the degree of technical bachelor in industry in facilities, equipment and electrical machines in the College "Tecnico Salesiano". Currently, Astudying in the university studies at the Universidad Politécnica Salesiana in the city of Cuenca, the degree in Electrical Engineering.

Juan Carlos Maldonado, was born in 1992 in the city of Zarúma, grew up in the canton Atahualpa where he finished his secondary studies at the school "Ángel Tinoco Ruíz" in the specialty of mathematical physics, currently studying at the Universidad.

Marco Antonio Toledo Orozco, Electrical Engineer (2010) of the Universidad Politécnica Salesiana of Cuenca-Ecuador. He is currently a PhD student at the Polytechnic University of Valencia – Spain; He works in the Ecuadorian Electricity Sector since 2002, in the areas of Regulation and Control, Distribution and Commercialization of electric energy (Laboratory of Meters, Losses of Energy, Big Consumers, and Electrical Markets). His areas of interest are Distribution Planning, Intelligent Networks, Demand Management and Electrical Markets.

Pablo Alejandro Méndez S. He was born in Cuenca in 1979. He received his Electrical Engineering degree from the Salesian Polytechnic University in 2004 and since 2013 he is a Master in Electrical Power Systems from the University of Cuenca. He currently works as a teacher in the faculty of Engineering from the Salesian Polytechnic University and is an official of the Commercialization Directorate of the Electric Company CENTROSUR.

Carlos Álvarez Bel, Professor, Universidad Politécnica de Valencia (UPV), received his MSc and PhD in Electrical Engineering in 1976 and 1979 from the Universidad Politécnica de Valencia, where he is Professor since 1989. His professional activity has been performed in the electric energy systems field in the framework of utilities, research centers and Universities. He has been involved in many projects and consulting work with utilities both in Spain and abroad, in the fields of state estimation, load modeling, standard markets, micro-grids, etc. His current interest areas are

basically related to the distribution and commercialization of the electric energy in the free electricity market, such as voltage quality, reliability of distribution systems, demand side management, integrated distribution management systems, and new micro-grid structures.

Diego Morales, He received his degree in Electrical Engineering from the Universidad Politécnica Salesiana in 2009; Master degree in Geographic Information Systems from the University of Salzburg - Austria in 2013; Master in Electrical Engineering specialty Smart Grids and Buildings from Grenoble Institute of Technology-France (Grenoble INP) in 2014. He is Ph.D. in E Electrical Engineering at Université Grenoble Alpes. Eng. Morales is part of Smart Grids Team and Professor of Electrical Engineering at the Universidad Católica de Cuenca. In addition, Eng. Morales works in the Ministry of Electricity and Renewable Energy as Distribution Specialist. His research interests are Smart Grid Application Development and Geographical Applications for calculation of electricity demand.

Oil Flow Rate Forecasting For Wells Drilled in Unconventional Reservoirs

Umer Farooq¹, Randy Hazlett² and Krishna Babu³

^{1,2} University of Tulsa, Tulsa, OK 74104, USA

³ Potential Research Solutions, Houston, USA

Abstract

A new analytical model has been employed forecast inflow performance in unconventional reservoirs producing under bottomhole pressure constraint. The traditional decline curve analysis done in the industry for the flow rate time series forecasting, is mere regression and extrapolation of production data and suffers from lack of theoretical basis, and there is no physical significance to the fitting parameters. A new forward model solution to the diffusivity equation has been developed using convolution in Laplace space to capture constant bottomhole pressure constrained production from the fundamental solution for constant rate production. Analytic Laplace inversion is performed to yield a result in time with decay constants directly related to geometry.

While curve fitting with a sum of exponentials is often an ill-posed problem, our physically-based method for exponent relations has geometrical interpretation. The new model contains a sum of exponential decay terms relating production to the geometry of the complex fracture system. The geometrical parameters relate to the medium permeability and proximity to various boundaries, whether they be system boundaries or those developed due to fracture interference. The most ground breaking feature of the method is that it adds physical significance the mathematical terms as opposed to a mere curve fit to the data. This analytical solution to the Diffusivity Equation under constant pressure constraint gives a new mathematical structure for analyzing decline in unconventional wells. Since the hyperbolic curve fit is typically used in decline curve analysis of unconventional reservoirs. It is demonstrated that any hyperbolic fit can be reinterpreted as a sum of terms, each with exponential damping, only now analytically with mathematical rigor and physical significance.

This methodology can change the way unconventional wells are routinely modeled, giving physical significance to the various contributing terms. It may also allow unconventional wells already modeled with piecewise hyperbolic and exponential functions to be easily reinterpreted. The new approach eliminates the need to patch together two separate models at an ad-hoc determined time. Thus, rate time analysis can be used to characterize not simply the well, but the efficacy and extent of the stimulation, designed to create a network of extensive contact with the well.

Keywords: Petroleum, Rate Time Analysis, Decline Curve Analysis.

1 Introduction

The petroleum industry is evolving faster than ever and is responsible for meeting a major portion of the world's energy demands. After producing for decades from conventional formations with high permeabilities, the focus shifted to unconventional resources. Where a conventional well produces at a constant flow rate until a bottom-hole pressure constraint is reached and the well goes into rate decline, unconventional wells typically enter immediately to rate declining mode.

In unconventional formations, due to extremely low permeabilities, the wells are drilled horizontally to increase the contact area with the formation. The wells are then fractured to further increase the degree of contact to produce at an economically viable rate. While theory advocates exponential decline in conventional wells, the decline in an unconventional formation cannot be easily modeled with a single exponential function, and is a topic of great interest in the oil and gas industry.

The most commonly used technique for such flow rate forecasting is decline curve analysis, which is a mere curve fit of the production data. Decline curve analysis is an industrywide accepted method but suffers from a major weakness. The curve fit parameters have no physical significance. Hyperbolic decline models the initial decline period well, but the recovery estimates remain inaccurate.

In this study a form of analytical solution is used to give physical meaning to the fitting parameters that link to the geometry of the reservoir. A literature review is presented followed by a demonstration of hyperbolic decline. The new form of the equation is then used to demonstrate how the fitting parameters reveal extremely useful information about the geometry of the system.

If the rate decline can be studied in detail and accurate methods are developed to characterize it, one can make forecasts on the future performance of the well. Predicting the future performance of a well and future behavior of the reservoir is of utmost importance, not only for technical reasons, but also for analyzing the economics involved and, ultimately, estimating profit and the economic life of a well. Other important decisions, such as when to employ improved and enhanced recovery methods to increase the production, are also directly affected by productivity and well performance forecasts.

2 Background

Attempts to model rate decline behavior began in the 1940s. Arps [1] proposed some of the fundamental mathematical relationships for decline curve analysis. Such analyses are very important for the appraisal of wells, in which future performance is extrapolated from historical performance. Different classifications of decline curves were presented based on the shape of the flow rate versus time plot.

Exponential decline is characterized by a drop in production flow rate per unit time proportional to the production rate. Exponential decline is mathematically represented as

$$Q = Q_o e^{\frac{-t}{a_o}}, \quad (1)$$

where a_o is a constant and Q represents the production rate, $Q = Q_o$ at $t = 0$. If the exponential decline curve is plotted on a semi-log paper, the slope is observed to be constant and equal to $-1/a_o$. The decline percentage can be represented as:

$$D = -100 \frac{1}{Q} \frac{dQ}{dt}. \quad (2)$$

A future production calculator was presented by Arps [1] based on the exponential production decline.

In hyperbolic decline, the decline percentage is proportional to the production rate raised to some power b_o . The mathematical expression can be written as

$$Q = Q_o \left(1 + \frac{b_o t}{a_o}\right)^{-\frac{1}{b_o}}, \quad (3)$$

and the percentage decline can be written as

$$D = \frac{100}{a_o Q_o^{b_o}} Q^{b_o}. \quad (4)$$

For special cases where $b_o = 1$, the decline is called harmonic.

Fetkovich [6] introduced type curves in decline curve analysis. Dimensionless log-log type curves were generated for exponential, hyperbolic and harmonic decline curves developed in Arps [1]. In another publication Fetkovich [7] discussed strategies for rate forecasting and reserve estimation.

Ilk *et al.* [10] has done significant effort with tight gas reservoirs, to get some diagnostic understanding of the system from the curve fit parameters of the hyperbolic decline form from Arps [1]. In another publication Ilk *et al.* [11] investigated more applications of decline curve analysis and attempted to identify the reservoir flow regimes based on the hyperbolic decline parameters. Freeborn & Russell [8] demonstrated the use of stretched exponentials in reservoir evaluation. Some of the most recent work includes Ogunyomi *et al.* [13] where model based analysis is performed to get physical interpretation about the system. A detailed analysis of improved decline curve models can be found in a recent paper published by Paryani *et al.* [14].

In his research thesis, Gonzalez [9] studied probabilistic decline curve analysis. Multiple decline curve analysis models were coupled to different probabilistic methods were performed to analyze the reliability of the models. Logistic growth model is used to reduce the uncertainty in volumetric analysis and reserve estimation. Brito *et al.* [2] also worked with probabilistic production forecasts. More work can be found in McNulty & Knapp [12], where some statistical methods have been employed for decline curve analysis.

3 Reinterpretation of Hyperbolic Rate Decline

The most common method used in decline curve analysis for unconventional wells is the hyperbolic decline from Arps [1]. Eq. 3 represents the mathematical form of the model. We observed that it is possible to decompose the hyperbolic decline data as a sum of exponentials. The hyperbolic data was mapped into only one exponential first

to see the trend, then two exponentials and then five. Figs. 1a, 1b, and 1c show the hyperbolic curve with a normalized rate and $b_o = 0.6$, along with the exponential approximations. It can be seen in Fig. 1 that as the number of exponentials is increased, the match with data becomes better. The respective exponential forms are shown in Eqs. (5-7)

$$q(t) = 0.78e^{-0.09t} \quad (5)$$

$$q(t) = 0.74e^{-0.19t} + 0.21e^{-0.03t} \quad (6)$$

$$q(t) = 0.1e^{-0.56t} + 0.34e^{-0.27t} + 0.034e^{-0.12t} + 0.17e^{-0.05t} + 0.04e^{-0.01t} \quad (7)$$

This is a great motivation to study the rate declines as composite exponentials, but that alone is not enough. The goal is to relate the exponential parameters to the physics of the system to be able to diagnose the reservoir rather than just being able to forecast the flow rate.

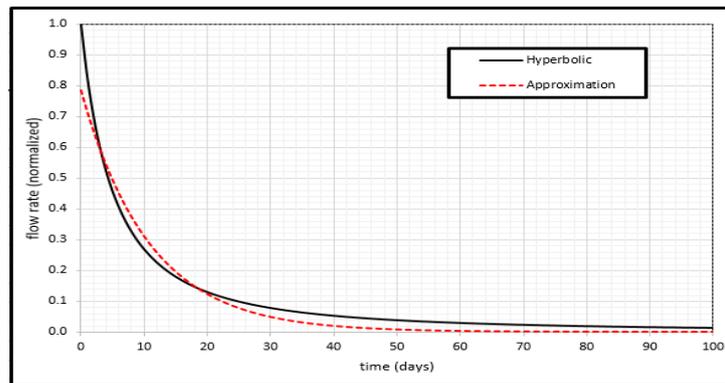


Fig. 1a. Approximation of hyperbolic decline with one exponential

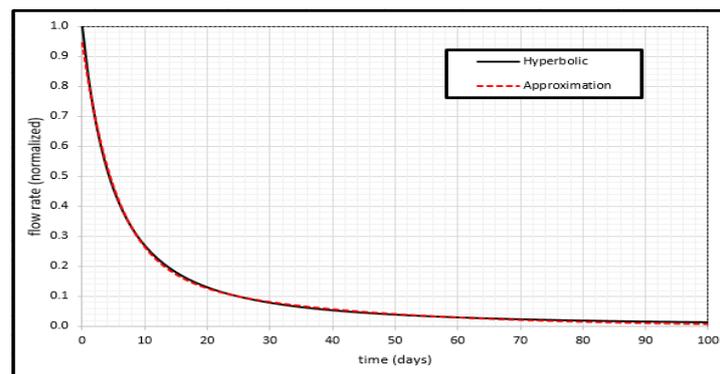


Fig. 1b. Approximation of hyperbolic decline with two exponentials

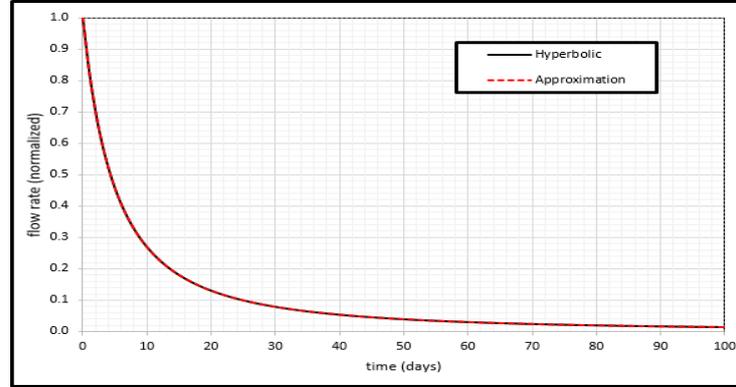


Fig. 1c. Approximation of hyperbolic decline with five exponentials

4 Rate Decline Expressed as Exponentials

A state of the art three-dimensional model is employed to study the rate decline data in an attempt to attain some diagnostic understanding of the reservoir and well geometry. As the developed analytical model is very general in nature, it can be used for cases of lower dimensionality. The 3D solution contains contributions from the various projections of the well onto planes and axes. The model can be used simulate fully penetrating fractures in the reservoir by dropping one dimension. Superposition can then be used to model flow through matrix to a fracture gathering system. The analytical expression for the flow rate of well with a bottomhole pressure constraint, in a three dimensional anisotropic medium is shown in Eq. 8. The model was created for a sealed outer boundary. The development and validation of the model is beyond the intended scope of this text and can be found in Farooq [5].

$$q_D(t_D) = \frac{e^{-t_D}}{D} - \frac{e^{-t_D}}{D^2} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{\rho_{lmn}(2-\rho_{lmn})}{(1-\rho_{lmn})^2} \right] + \frac{e^{-t_D} t_D}{D^3} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{\rho_{lmn}}{(1-\rho_{lmn})} \right] + \frac{1}{D^2} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{e^{-\alpha_{lmn} t_D}}{(1-\rho_{lmn})^2} \right], \quad (8)$$

where

$$t_D = \frac{t}{\phi \mu c_t}, \quad (9)$$

$$q_D(t_D) = \frac{\mu q_v(t_D)}{V \Delta P_{wf}} \quad (10)$$

and

$$\alpha_{lmn} = \pi^2 \left(\frac{k_x l^2}{a^2} + \frac{k_y m^2}{b^2} + \frac{k_z n^2}{h^2} \right). \quad (11)$$

The parameters D and ρ_{lmn} are directly related to α_{lmn} . This form of the equation is extremely powerful as the exponential terms encapsulate the effects of reservoir geometry ($\mathbf{a}, \mathbf{b}, \mathbf{h}$), in relation to well position.

The value of D has physical meanings and represents the stabilized response of the system. Value of D is independent of time, so it can be seen from Eq. (8) that the only exponential decay term with the combined effect of space and time (in the exponent) is the last summation. Eq. 8 can be rewritten as:

$$q_D(t_D) = e^{-\frac{t_D}{D}} \left\{ \frac{1}{D} - \frac{1}{D^2} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{\rho_{lmn}(2-\rho_{lmn})}{(1-\rho_{lmn})^2} \right] \right\} + e^{-\frac{t_D}{D}} t_D \left\{ \frac{1}{D^3} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{\rho_{lmn}}{(1-\rho_{lmn})} \right] \right\} + \frac{1}{D^2} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{e^{-\alpha_{lmn} t_D}}{(1-\rho_{lmn})^2} \right] \quad (12)$$

In Eq. (12), the coefficients within the first two flower brackets are time independent and can be replaced by notations α and β respectively.

$$q_D(t_D) = e^{-\frac{t_D}{D}} \alpha + e^{-\frac{t_D}{D}} t_D \beta + \frac{1}{D^2} \sum_{l,m,n \neq 0}^{\infty} \left[D_{lmn} \frac{e^{-\alpha_{lmn} t_D}}{(1-\rho_{lmn})^2} \right] \quad (13)$$

After running simulations it was concluded that three exponentials from the last series provide good fit for the data produced. Based on the form of the solution, the following form of equation is proposed to fit the rate decline data for geometrical interpretation. For simplicity, we will drop the subscript D with the flow rate and time, but the rate and time are meant to be in dimensionless form.

$$q(t) = e^{-\frac{t}{D}} \alpha + e^{-\frac{t}{D}} t \beta + A e^{-a_1 t} + B e^{-b_1 t} + C e^{-c_1 t} \quad (14)$$

Eq. (14) provides the final form of the decline equation with exponential parameters directly linked to the system geometry.

4.1 Effect of Reservoir Geometry

The parameters $\alpha, \beta, A, a_1, B, b_1, C,$ and c_1 in accordance with Eq. (8) carry information about the reservoir geometry. We begin to demonstrate this with three different well geometries in a cube, where all dimensions are normalized with the largest dimension a , so a is always equal to 1.

Figs. 2a, 2b, and 2c represent the three cases considered. Fig. 2a shows the decline behavior of a fully penetrating vertical well in a cubic reservoir. A very general case of a partially penetrating diagonal well is presented in Fig. 2b. The well starting and ending coordinates are chosen to be (0.25, 0.25, 0.25) and (0.75, 0.75, 0.75) respectively. In Fig. 2c, a partially penetrating horizontal well a cubic reservoir is considered where well start and end coordinates are chosen to be (0.25, 0.5, 0.5) and (0.75, 0.5, 0.5) respectively. The rate needs to be plotted on a log-log scale to make the changes observable.

The data is fit into the form of Eq. (14) and then analyzed. The first term in Eq. (14) with α comes from two separate terms. We use the terms separately while fitting to make the identification of α by spotting repeating delay constants.

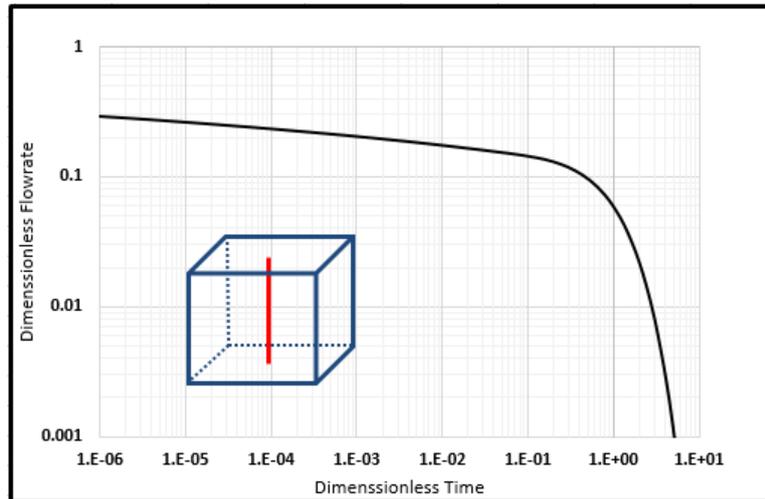


Fig. 2a. Rate decline of a fully penetrating vertical well in a cubic reservoir

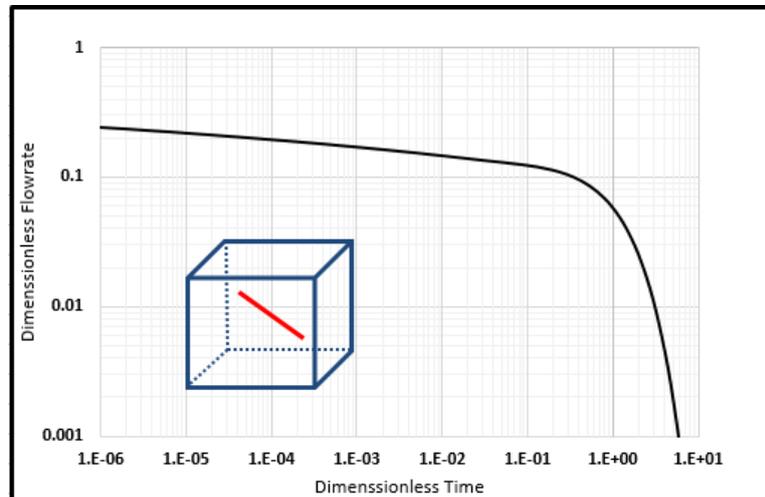


Fig. 2b. Rate decline of a partially penetrating diagonal well in a cubic reservoir

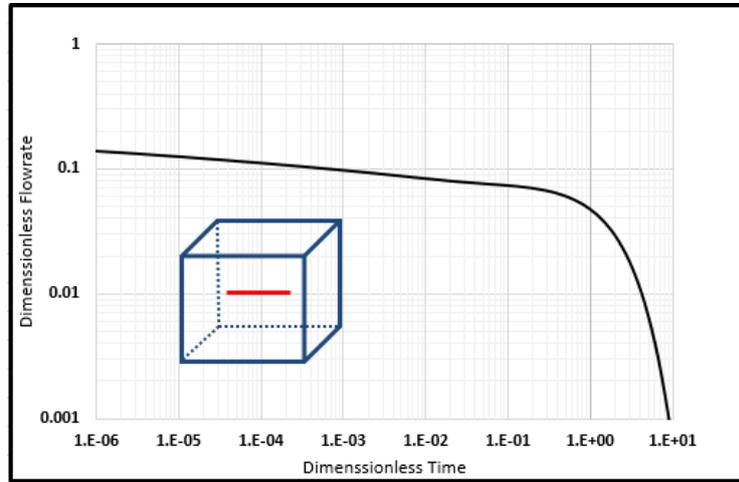


Fig. 2c. Rate decline of a partially penetrating horizontal well in a cubic reservoir

After fitting the data, parameters for cases 2a, 2b, and 2c are shown in Eqs. (15), (16), and (17) respectively. They written so as to match the shape of Eq. (14)

$$q(t) = 6.41e^{-1.03t} - 20.49te^{-259t} + 18.96e^{-0.37t} + 6.31e^{-1.09t} + 0.10e^{-1800t} \quad (15)$$

$$q(t) = 6.38e^{-0.79t} - 15.71te^{-244t} + 18.97e^{-0.80t} + 6.34e^{-0.82t} + 0.09e^{-1690t} \quad (16)$$

$$q(t) = 7.09e^{-0.37t} - 6.95te^{-205t} + 21.02e^{-0.37t} + 6.9e^{-0.38t} + 0.05e^{-1515t} \quad (17)$$

We clearly observe a signature for the cubic reservoir geometry. The second term with t multiplied is always negative irrespective of the well geometry and the decay constants tend to decrease going from Eq. (15) to (17). This signature will be compared with a slab geometry in the next section.

Reservoir Diagnostics. A major advantage of this semi-analytical approach is the inverse problem solving capability. The factor D in the equation is represents the psuedo-steady state contribution of the solution even if the regime is never observed in a practical time scale.

The values of D factor can be pre-calculated for specific geometries and made into tables available to a petroleum engineer. Once the data have been fit the form of the equation, it can be directly compared with Eq. (14) to reveal valuable information about well geometry relative to various boundaries.

We will use cases 2a and 2b to demonstrate this methodology. The D factor for vertical well has been calculated, using computation from pseudo steady state formu-

lation, to be 0.99 and for diagonal geometry to be 1.18. Now if we compare the first terms of Eqs. (14) and (15), we can relate the exponents together to diagnose the geometry of the system. Table 1 represents both the computed values of D factor and the estimated values from data, for the cases of vertical and diagonal wells.

Table 1. Computed and Estimated Values of D Factor for a Cubic Reservoir

	Vertical Well	Diagonal Well
Computed Value	0.99	1.18
Estimated Value	0.97	1.2

The value of D factor in both the cases turns out to be very close to the calculated value. This is a very useful tool which can be applied to characterize the well geometry in the case of pressure constrained production.

4.2 Effect of Height in Slab- like Reservoirs:

In unconventional formations the well are drilled laterally for miles in a slab-like formation. In this section, two scenarios are considered where the height of the reservoir is considerably smaller than the lateral dimension. Fig. 3 shows the cases, with a $1000 \times 1000 \text{ ft}^2$ area and a partially penetrating horizontal well, as in case 2c, but the reservoir geometries are different. In first case, the reservoir height is 50 ft. and in second its 100 ft.

The distance of the boundaries from the well carries significant information about the flow regimes that can be expected at a certain point in time in the reservoir. The knowledge of the flow regime can be used to run tests on the well and characterize the reservoir properties. We will observe the two cases 3a and 3b to see how the signature compares to a cubic geometry.

The approximation for $h = 50 \text{ ft}$ case is shown in Eq. (16), and for $h = 100 \text{ ft}$ case by Eq. (17).

$$q(t) = 0.96e^{-20.74t} + 5.04te^{-0.32t} \quad (18)$$

$$q(t) = 0.68e^{-17.32t} + 4.67te^{-0.42t} \quad (19)$$

Just as expected, a signature is seen for a partially penetrating well in a slab reservoir as compared to the signature from a cubic reservoir. In a slab geometry all decay terms other than the second term overlap reducing the form to two exponentials only. It can also be noticed that all coefficients remain positive compared to a negative second term for a cubic geometry.

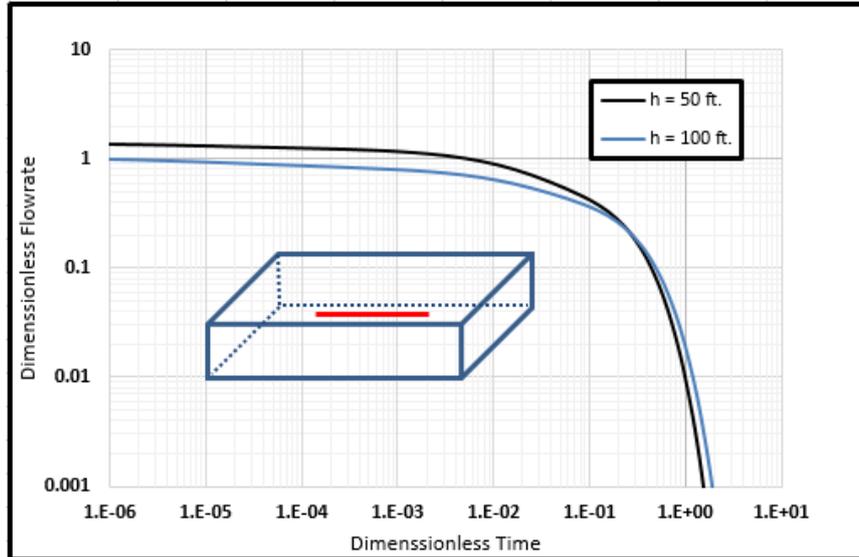


Fig. 3. Rate decline of a partially penetrating horizontal well in a slab reservoir

As the height of the reservoir is doubled from 50 to 100 ft., a similar consistent signature is observed. The differences between the parameters between Eqs. (18) and (19) were as expected. When the reservoir height is doubled the onset of stabilized flow is delayed, so correspondingly, we see the reduction in the decay constant.

4.3 Effect of Fracture Penetration

Unconventional wells are orthogonally fractured in stages to stimulate the formation. Attempts to model the rate decline behavior of such fractures can be abundantly found in the literature. Chaudhary *et al.* [3] has studied the oil production from stimulated shale reservoirs. Doung [4] has presented a method for rate decline analysis in fracture-dominated shale reservoirs. Decline curve analysis has been carried out using type curves in a publication by Pratikno *et al.* [15].

While stimulating treatment, the goal is to make a good quality fracture that penetrates the whole length distance between adjacent horizontal wells. The penetration ratio is given by l/b (where l is length of fracture and b is width of drainage area). The geometry of the fracture significantly affects the production rates. When fractures are present in stages they tend to form no flow boundaries at the interfaces of the drainage areas of individual fractures.

A general scenario that is widely studied in petroleum engineering is the case of a single vertically fully penetrating fracture orthogonal to a horizontal well. The fracture drains from a specified area with no flow outer boundaries. Such a scenario can be described in a two dimensional domain. The schematic of this general case is shown in Fig. 4. The black line represents the no flow boundary, dotted red line

shows a fracture that is fully penetrating in the vertical direction penetrating the fracture and the blue arrows represent fluid flux.

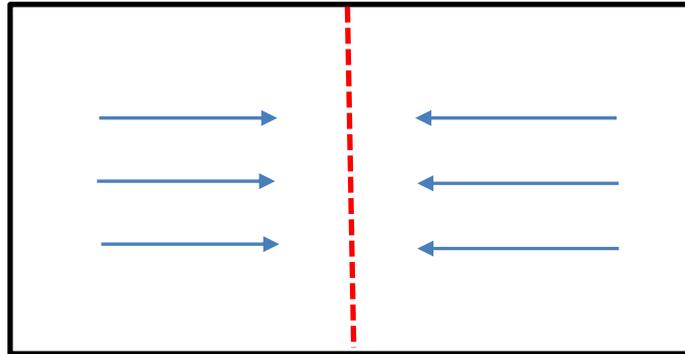


Fig. 4. Schematic of a fully penetrating fracture in 2D, with $l/b=1$

The state of the art analytical model used can be reduced to 2D domain to simulate fractures. The model also has the capability to simulated partially penetrating or arbitrarily oriented fractures.

The investigation of the effects of penetration of a fracture into the formation is of particular industrial interest. This is demonstrated in a $1000 \times 500 \text{ ft}^2$ two dimensional drainage area. Simulations are performed for the cases of $l/b = 0.1, 0.5$ and 1 , where l is the fracture length and b is the minor axis length. The results from the scenarios are shown in Fig. 5.

The least square regressions for the $l/b = 0.1, 0.5$ and 1 cases are shown in Eqs. (20), (21), and (22) respectively.

$$q(t) = 0.21e^{-1037t} - 2373te^{-12998t} + 0.29e^{-260t} + 0.18e^{-47t} + 0.59e^{-3.871t} \quad (20)$$

$$q(t) = 0.2e^{-638t} - 1286te^{9136t} + 0.73e^{-224t} + 0.77e^{-44t} + 0.93e^{-6.77t} \quad (21)$$

$$q(t) = 0.02e^{-1507t} - 1124te^{-10241t} + 0.42e^{-309t} + 2.2e^{-46t} + 0.95e^{-8.62t} \quad (22)$$

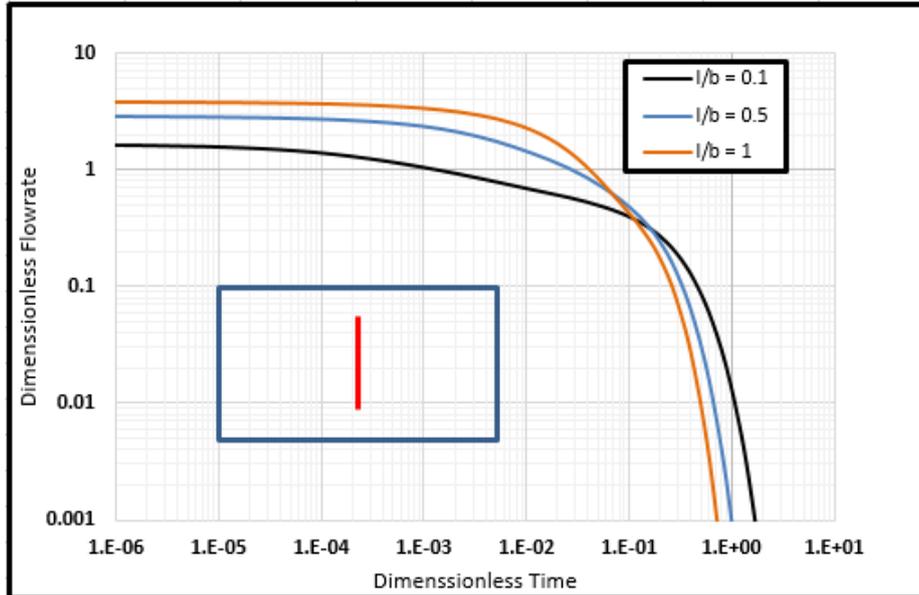


Fig. 5. Rate decline of a partially penetrating fracture in a 2D domain.

After inspecting the results, the signature for a fracture can be clearly seen, where the second term is always negative with a large coefficient and an extremely high time constant is observed. The most interesting thing to note is the first term that clearly distinguishes the partially penetrating fractures in the first two cases from the fully penetrating fracture. The term becomes considerably smaller as fractures approach full penetration.

4.4 Effect of Flow Boundaries on a Fractures:

No flow boundaries are formed in between different stages of fractures between adjacent unconventional wells. Here we consider two different scenarios to demonstrate this behavior. A single partially penetrating fracture is considered enclosed by sealed boundaries. The start and end points of fractures in both the cases are (0.5, 0.25) and (0.5, 0.75) respectively. In first scenario an aspect ratio of 1:2 is considered for the boundaries, and then compared with a second scenario of aspect ratio 1:1. The results from both the cases are shown in Fig. 6.

After fitting the data for both cases, the exponential approximations for 1:2 and 1:1 cases reduce to in Eqs. (23), and (24), respectively.

$$q(t) = 2.596e^{-24.83t} + 1.792te^{-1.406t} \quad (23)$$

$$q(t) = 1.674e^{-18.62t} + 4.268te^{-0.792t} \quad (24)$$

Exponentials with same decay constants are observed and combined together to get an approximation using two exponentials. It can be observed that as the aspect ratio moves from rectangular to square, both the coefficient and decay constant for the first terms decrease as seen in Eqs. (23) and (24). The coefficient of the second term increases but the decay constant decreases. The effects of boundaries can be more very efficiently analyzed using derivative analysis.

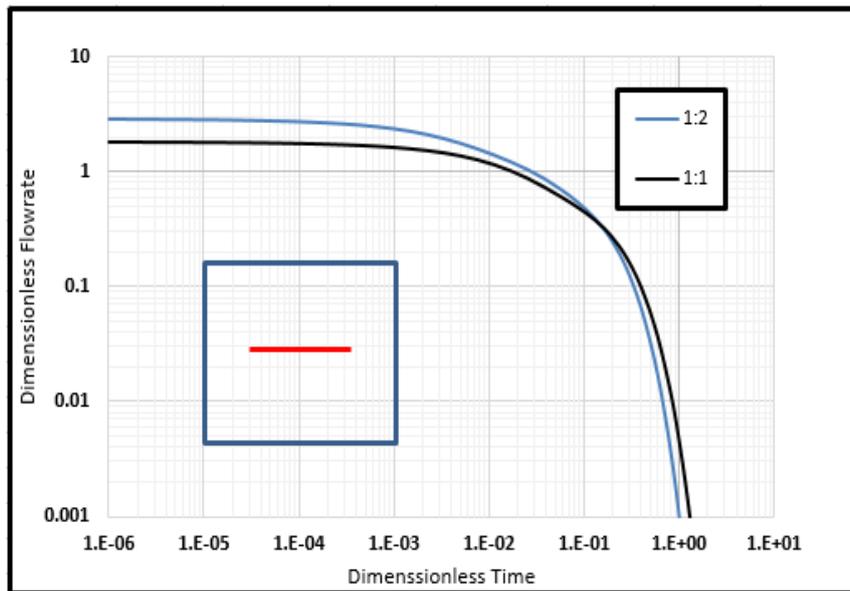


Fig. 6. Rate decline of a partially penetrating fracture, with drainage area aspect ratios of 1:2 and 1:1

4.5 Derivative Analysis:

The rate decline information is crucial for forecasting the production of a well, but for reservoir characterization purposes the rate of change of the oil rate with time is of particular interest. Shahamat *et al.* [16] used beta derivative to do the analysis of decline curves. Interpretation techniques were discussed in the paper for various reservoir and well parameters. In a different study rate derivative curves have been plotted for various 2D geometries by Zhang & Grader [17].

The form of Eq. (8) allows for convenient analytical computations of the rate derivative. A particular type of derivative used in well testing is the Bourdet derivative. The Bourdet derivative is described as follows:

$$\frac{dq_D(t_D)}{d \ln t_D} = t_D \frac{dq_D(t_D)}{dt_D} \quad (25)$$

Here the dimensions of time cancel out and we are left with the dimensions of rate only. The plots of the Bourdet derivatives exhibit signature slopes for the flow regimes happening in the reservoir. To demonstrate such diagnostics, we consider a

partially penetrating fracture in three different cases. The length of the drainage area is kept at 1000ft and the width is varied. Three separate widths of 200ft, 500ft, and 1000ft are considered. The log-log derivative plots for the three cases are shown in Fig. 7.

Most important to note in the derivative plots are the linear portions, specific slopes, and durations. Each of these portions signify a flow regime for the fluid streamlines that encapsulates the effects of geometry. When production from a fracture begins, the fluids flow linearly into the fracture and hence the flow regime is called linear. The signature of linear flow on a log-log derivative plot is known to be a slope of 0.5. The absolute value of the derivative is plotted here.

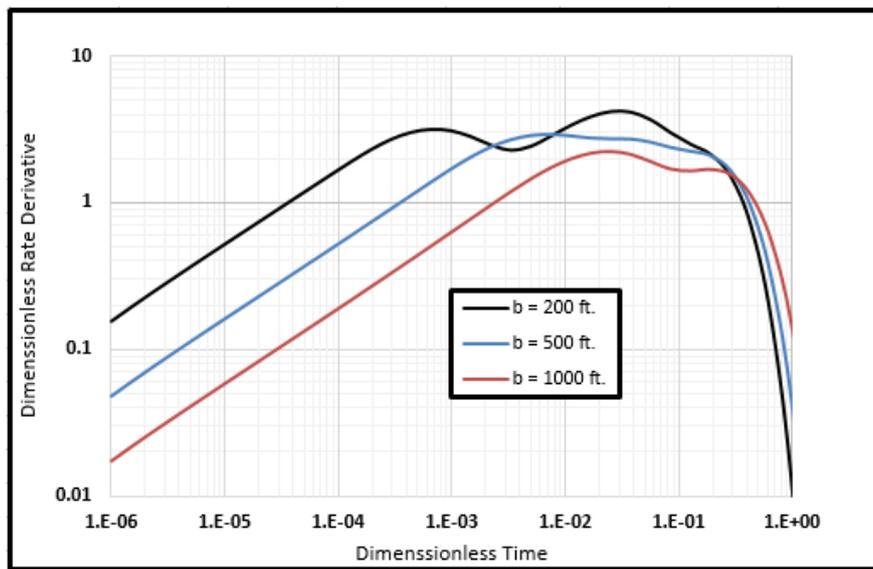


Fig. 7. Log-log derivative plots for a fracture in a rectangular domain of different aspect ratios

It can be seen in Fig. that in all the three cases we begin with a linear flow but we see a deviation from 0.5 slope by the 200ft case first, followed by 500ft and 1000ft cases respectively. This is diagnostic of the fact that the boundaries are felt by the fracture earliest in the 200ft case demonstrated by a lower value of t_D for deviation from linear flow. A square drainage area can be characterized by the fact that end of linear flow happens at an approximated t_D value of 0.01.

Another scenario that can be used to demonstrate the effectiveness of derivative analysis is fracture rotation. The orientation of a fracture can reveal crucial information about the preferential fracture growth direction in the formation. Let's consider a fracture in a 1000 x 500 ft² of drainage area that is located in the center and is parallel to the y-axis. The fracture is also fully penetrating along the y-direction. We consider two more scenarios keeping the same fracture length and drainage area. In second case, the fracture makes a 45° angle with the y- axis and in the third case its

parallel to the x-axis. It should be noted that when the fracture is rotated, it becomes partially penetrating in the plain, with its tips lying inside the drainage area.

If the rates are plotted on a log-log scale for the above three cases, the behavior is nearly indistinguishable due to very similar production rates. Much more insight can be gained into the fracture orientation using a derivative plot for the scenario. The derivative plot for the three scenarios is shown in Fig. 7. The angle in degrees shown in the figure shows the orientation of the fracture w.r.t the minor axis (y-axis).

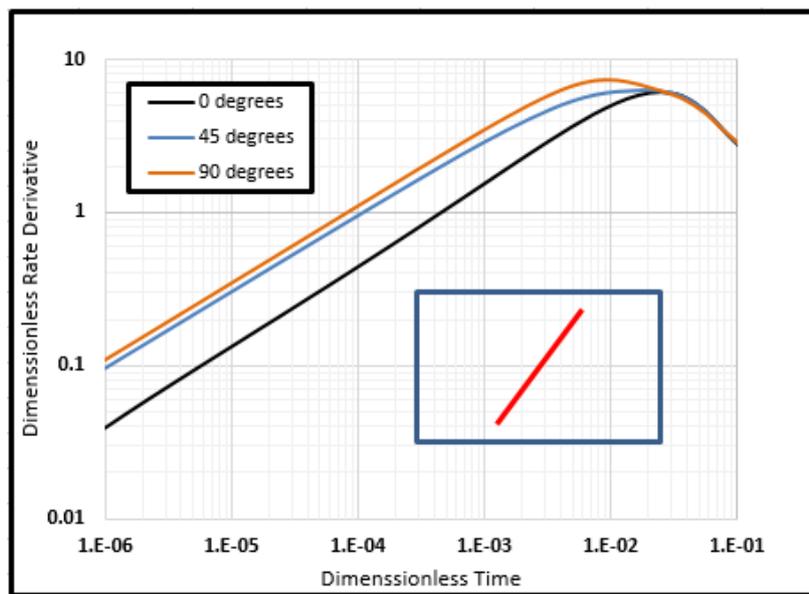


Fig. 7. Log-log derivative plots for a fracture with different orientations in a rectangular domain

As it can be observed, a clear distinction is seen for different fracture orientations in a rectangular domain. It should also be noticed that the fractures that are parallel to one of the axes show a clear hump at the end of linear flow in comparison to a much gradual transition in the case of diagonal fracture. These derivative plots can be generated for a much wider variety of fracture orientations to be able to precisely characterize the system. Such derivative analysis can similarly be applied to any reservoir and well geometry of interest.

5 Conclusions

A semi analytical approach was developed as an alternative to the traditional decline curve analysis for flow rate forecasting in unconventional formations. The target was to achieve rate transient forecasting with geometrical interpretation of fitting parameters. All the cases simulated showed very encouraging results and displayed distinct signatures which are crucial for reservoir diagnostics. A remarkable feature of this

interpretive model is the influence of the predicted pseudosteady state behavior throughout time, even if the flow regime is never attained within the window of observation.

The approach can be used not only to diagnose reservoir geometries but also simulate the effects of fracture penetration and orientation in unconventional wells. From the fracture results, we could observe the behavior of the parameters as fracture length was increased. Moreover, we were able to distinguish a fully penetrating fracture very distinctly from all the partially penetrating cases. The challenge is to learn about the reservoir characteristics from the rate decline behavior. The model used allows directly for such interpretation.

Nomenclature

a = Reservoir dimension in x direction
 a_1 = Curve fit parameter
 A = Curve fit parameter
 b = Reservoir dimension in y direction
 b_1 = Curve fit parameter
 B = Curve fit parameter
 c_t = Total compressibility
 c_1 = Curve fit parameter
 C = Curve fit parameter
 $1/a_o$ = Monthly percentage decline rate
 h = Height of reservoir
 k_x = Permeability in x direction
 k_y = Permeability in y direction
 k_z = Permeability in z direction
 l = Length of well/ fracture
 q = Flow rate
 q_o = Constant flow rate
 q_v = Variable flow rate
 q_D = Dimensionless flow rate
 t = Time
 t_D = Dimensionless time
 α = Curve fit parameter
 β = Curve fit parameter
 φ = Porosity
 μ = Viscosity

References

1. Arps, J. J.: Analysis of Decline Curves. Trans. AIME, 160, 228-247 (1945).
2. Brito, L. E., Paz, F., & Belisario, D. R.: Probabilistic Production Forecasts Using Decline Envelopes. Society of Petroleum Engineers (2012).
3. Chaudhary, A. S., Ehlig-Economides, C. A., & Wattenbarger, R. A.: Shale Oil Production Performance from a Stimulated Reservoir Volume. Society of Petroleum Engineers (2011).
4. Duong, A. N.: Rate-Decline Analysis for Fracture-Dominated Shale Reservoirs. Society of Petroleum Engineers (2011).
5. Farooq U.: Pressure Constrained Production from an Arbitrarily Oriented Line Source in a 3D Anisotropic Reservoir, University of Tulsa (2015).
6. Fetkovich, M. J.: Decline Curve Analysis Using Type Curves. 4629-PA SPE J. Paper. Phillips Petroleum Co. (1980).
7. Fetkovich, M. J., Fetkovich, E. J., & Fetkovich, M. D.: Useful Concepts for Decline Curve Forecasting, Reserve Estimation, and Analysis. Society of Petroleum Engineers (1996).
8. Freeborn, R., & Russell, B.: How To Apply Stretched Exponential Equations to Reserve Evaluation. Society of Petroleum Engineers (2012).
9. Gonzalez, R. A.: Using Decline Curve Analysis, Volumetric Analysis and Bayesian Methodology to Quantify Uncertainty in Shale Gas Reserve Estimates, Texas A&M (2012).
10. Ilk, D., Rushing, J. A., Perego, A. D., & Blasingame, T. A.: Exponential vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves. Society of Petroleum Engineers (2008).
11. Ilk, D., Rushing, J. A., & Blasingame, T. A.: Decline-Curve Analysis for HP/HT Gas Wells: Theory and Applications. Society of Petroleum Engineers (2009).
12. McNulty, R. R., & Knapp, R. M.: Statistical Decline Curve Analysis. Society of Petroleum Engineers (1981).
13. Ogunyomi, B. A., Dong, S., La, N., Lake, L. W., & Kabir, C. S.: A New Approach to Modeling Production Decline in Unconventional Formations. Society of Petroleum Engineers (2014).
14. Paryani, M., Ahmadi, M., Awoleke, O., & Hanks, C.: Using Improved Decline Curve Models for Production Forecasts in Unconventional Reservoirs. Society of Petroleum Engineers (2016).
15. Pratikno, H., Rushing, J. A., & Blasingame, T. A.: Decline Curve Analysis Using Type Curves - Fractured Wells. Society of Petroleum Engineers (2003).
16. Shahamat, M. S., Mattar, L., & Aguilera, R.: Analysis of Decline Curves on the Basis of Beta-Derivative. Society of Petroleum Engineers (2015).
17. Zhang, W., & Grader, A. S.: Analysis of Rate Decline Derivatives. Society of Petroleum Engineers (1994).

Predictive model of the techno-environmental performance of novel multi-function window combined ventilation system and solar photovoltaic blind using finite element method

Taehoon Hong¹, Jongbaek An¹, Jeongyoon Oh¹, Woojin Jung¹, and Minhyun Lee¹

¹Department of Architecture and Architectural Engineering, Yonsei University,
Seoul, 03722, Republic of Korea

{hong7, ajb2577, omk1500, andybrian, mignon}@yonsei.ac.kr

Abstract. This study proposed innovative multi-function window integrating solar photovoltaic blind and ventilation system that can be applied to building facade as building-integrated photovoltaic system and improve indoor environmental quality. In order to effectively apply the multi-function window, this study sought to develop a model for predicting the techno-environmental performance of the multi-function window based on finite element method by the following four stages: (i) selection of design elements affecting the multi-function window; (ii) development of database by using energy simulation programs; (iii) estimation of the techno-environmental performance of the multi-function window; and (iv) systemization. The developed model enables the designer or PMr to easily predict the techno-environmental performance of the multi-function window by inputting simple design information (e.g., orientation, size of ventilation system in the multi-function window and visible transmittance of glazing). In addition, the methodology suggested in this study can be extended to other projects and industries.

Keywords: Multi-function window, finite element method, techno-environmental performance, energy simulation

1 Introduction

As part of efforts to save energy in building sector, many buildings have been designed and constructed with high thermal performance in terms of insulation and airtightness. However, buildings with high levels of insulation and airtightness caused the deterioration of indoor environmental quality. Also, as people spend a lot of times indoors in modern society (e.g., Koreans stay in the room for about 87% of the day), indoor environmental quality has become a major factor affecting the occupants' health and productivity [1-3]. Therefore, this study considered a novel multi-function window combined ventilation system and solar photovoltaic blind. This system not only improves indoor environmental quality by introducing fresh air through ventilation, but

also can be utilized as a distributed photovoltaic system via electricity generated from solar photovoltaic blind.

In this study, a predictive model based on finite element method was proposed to analyze the techno-environmental performance of the multi-function window by design variables for effective introduction of the multi-function window. In addition, the process for predicting the techno-environmental performance of the multi-function window was systemized via *Microsoft Excel-based visual basic for application*, so that users such as designer or PMr can conveniently obtain the analysis results of the techno-environmental performance of the multi-function window.

2 Materials and methods

This study aimed to develop a model for estimating the techno-environmental performance (i.e., the amount of solar power generation, heating and cooling loads in terms of technical performance, indoor carbon dioxide (CO₂) concentration and predicted percentage of dissatisfied (PPD) in terms of environmental performance) of the multi-function window based on finite element method. This study consists of the following four stages: (i) selection of design elements affecting the multi-function window; (ii) development of database by using energy simulation tools; (iii) estimation of the techno-environmental performance of the multi-function window by using finite element method; and (iv) systemization.

- *Stage 1. Selection of design elements affecting the multi-function window*: This study defined design elements of multi-function window in terms of architectural, (i.e., region and orientation), window (i.e., glazing type, size of exterior window, size of ventilation system in the multi-function window and visible transmittance of glazing), and, solar photovoltaic blind (i.e., efficiency of photovoltaic panel) design elements [4-6].
- *Stage 2. Development of database by using energy simulation tools*: First, this study defined the standard classroom of South Korea for establishing database. Then, based on design elements defined in stage 1, database of technical (i.e., the amount of solar power generation, heating and cooling loads) and environmental (i.e., indoor CO₂ concentration and PPD) performance of the multi-function window was constructed through energy simulation tools. For analyzing the amount of solar power generation, *Autodesk Ecotect Analysis* was utilized, and *EnergyPlus v8.5* was used for deriving heating and cooling loads, indoor CO₂ concentration and PPD [7,8].
- *Stage 3. Estimation of the techno-environmental performance of the multi-function window by using finite element method*: The finite element method is a numerical method for solving problems, such as nonlinearity, of engineering and mathematical physics [9]. In order to solve the nonlinearity of the techno-environmental performance of the multi-function window by design elements (e.g., orientation, size of ventilation system in the multi-function window and visible transmittance of glazing), the finite element method was used according to the following processes: (i) selection of design elements to apply the finite element method (i.e., orientation

(ξ) and visible transmittance of glazing (η) for estimating the amount of solar power generation, orientation (ξ) and size of ventilation system (η) in the multi-function window for analyzing heating and cooling loads, indoor CO₂ concentration and PPD); (ii) discretization of continuous design elements; and (iii) deduction of shape and interpolation function [6,7].

- *Stage 4. Systemization:* The complex processes required to analyze the techno-environmental performance of the multi-function window according to design elements (e.g., orientation, size of ventilation system in the multi-function window, and visible transmittance of glazing) was automated based on *Microsoft Excel-based visual basic for application*.

3 Conclusion

This study sought to develop a predictive model based on finite element method of the techno-environmental performance of the multi-function window in educational facilities. By using the developed model, the users (e.g., designer or CMr) can conveniently analyze the techno-environmental effect by applying the multi-function window to building facade according to design elements (e.g., orientation, size of ventilation system in the multi-function window, and visible transmittance of glazing).

References

1. Heinzerling, D., Schiavon, S., Webster, T., Arens, E.: Indoor environmental quality assessment models: A literature review and a proposed weighting and classification scheme. *Build. Environ.* 70, 210–222 (2013).
2. Lan, L., Wargocki, P., Wyon, D.P., Lian, Z.: Effects of thermal discomfort in an office on perceived air quality, SBS symptoms, physiological responses, and human performance. *Indoor Air.* 21, 376–390 (2011).
3. Xue, P., Mak, C.M., Ai, Z.T.: A structured approach to overall environmental satisfaction in high-rise residential buildings. *Energy Build.* 116, 181–189 (2016).
4. Lee, M., Koo, C., Hong, T., Park, H.S.: Framework for the mapping of the monthly average daily solar radiation using an advanced case-based reasoning and a geostatistical technique. *Environ. Sci. Technol.* 48, 4604–4612 (2014).
5. Koo, C., Park, S., Hong, T., Park, H.S.: An estimation model for the heating and cooling demand of a residential building with a different envelope design using the finite element method. *Appl. Energy.* 115, 205–215 (2014).
6. Oh, J., Koo, C., Hong, T., Jeong, K., Lee, M.: An economic impact analysis of residential progressive electricity tariffs in implementing the building-integrated photovoltaic blind using an advanced finite element model. *Appl. Energy.* 202, 259–274 (2017).

7. Hong, T., Koo, C., Oh, J., Jeong, K.: Nonlinearity analysis of the shading effect of the technical-environmental performance of the building-integrated photovoltaic blind, *Appl. Energy*. 197, 467-480 (2017).
8. EnergyPlus, <https://energyplus.net/>
9. Finite element method, https://en.wikipedia.org/wiki/Finite_element_method

Acknowledgements

This research was supported by a grant (18CTAP-C117226-03) from Technology Advancement Research Program (TARP) funded by Ministry of Land, Infrastructure and Transport of Korean government.

Establishment of operational strategy of the ventilation system in a building by considering the indoor and outdoor concentration of fine dust

Taehoon Hong¹, Jeongyoon Oh¹, Woojin Jung¹, Jongbaek An¹, and Hakpyeong Kim¹

¹Department of Architecture and Architectural Engineering, Yonsei University, Seoul, 03722, Republic of Korea
{hong7, omk1500, andybrian, ajb2577, ibk1930}@yonsei.ac.kr

Abstract. Under the background of increased need for ventilation and deterioration of air pollution, this study aims to establish the operational strategy of the ventilation system in a building considering both indoor and outdoor concentration of fine dust (i.e., PM10 and PM2.5) based on regression analysis and management standard of indoor environmental quality. This study is composed of the following three steps: (i) data collection by using the real-time monitoring sensor; (ii) regression analysis for estimating the concentration of indoor fine dust; and (iii) establishment of operational strategy of ventilation system. Through the proposed operational strategy of ventilation system in a building, it is expected to create a pleasant indoor environment and improve the occupants' health.

Keywords: Ventilation system, fine dust, regression analysis, operational strategy

1 Introduction

In today's modern society, the importance of ventilation emerged as people live about 80~90% of the day in an enclosed indoor environment [1]. However, as the air pollutant is getting worse to such an extent that more than six million people die each year, there is a need to operate the ventilation system in a building by considering the outdoor environmental quality [2]. Especially, in South Korea, which has been suffering from high concentration of fine dust in recent year, it is necessary to control the ventilation system in a building considering both indoor and outdoor fine dust concentration. Table 1 shows the results of analyzing the concentration of indoor fine dust (i.e., particulate matter less than $10\mu m$ (PM10) and particulate matter less than $2.5\mu m$ (PM2.5)) according to the occupant's ventilation behavior based on collected data in office building located in Seoul, South Korea. From the analysis results, it can be seen that since the concentration of outdoor fine dust is higher than that of the room, the concentration of indoor fine dust increases after the end of ventilation. Therefore, this

study sought to establish the effective operational strategy of the ventilation system in a building considering both indoor and outdoor concentration of fine dust.

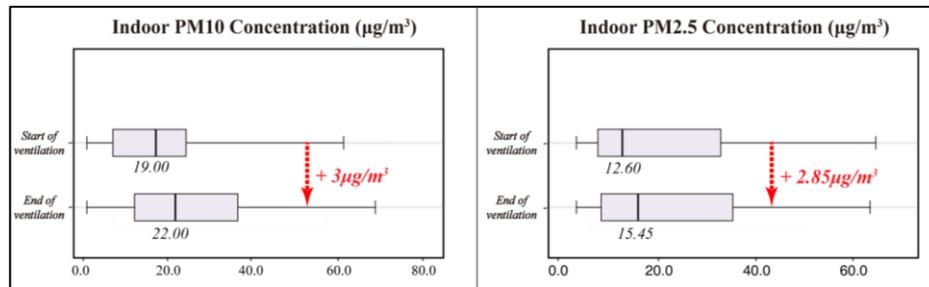


Figure 1. Box plot of fine dust (i.e., indoor PM10 concentration and PM2.5 concentration) according to ventilation

2 Materials and methods

In order to establish the operational strategy of the ventilation system in a building, this study conducted the following three-step process: (i) data collection by using the real-time monitoring sensor; (ii) regression analysis for estimating the concentration of indoor fine dust (i.e., PM10 and PM2.5); and (iii) establishment of operational strategy of ventilation system.

- *Stage 1. Data collection by using the real-time monitoring sensor:* This study collected data in terms of the following three perspectives through real-time monitoring sensor: (i) concentration of indoor fine dust; (ii) concentration of outdoor fine dust; and (iii) start and end of ventilation. Namely, in this study, indoor and outdoor fine dust (i.e., PM10 and PM2.5) concentration at the beginning and end of ventilation was collected in real-time.
- *Stage 2. Regression analysis for estimating the concentration of indoor fine dust:* In this study, regression analysis was performed to predict the concentration of indoor fine dust (i.e., PM10 and PM2.5) that can change due to outdoor fine dust as ventilation progresses. The explanatory variable for the regression analysis was defined as outdoor fine dust (i.e., PM10 and PM2.5) concentration, and the dependent variable was defined as the indoor fine dust (i.e., PM10 and PM2.5) concentration.
- *Stage 3. Establishment of operational strategy of ventilation system:* In this study, the operational strategy of the ventilation system in a building was established based on the regression model derived from stage 2 and management standard of indoor environmental quality in South Korea (refer to Table 1). In other words, if the indoor fine dust (i.e., PM10 and PM2.5) concentration predicted by the regression model is higher than the indoor environmental quality standard, operating only the exhaust ventilation system can be effective in terms of indoor environmental quality. In addition, it is desirable to operate both the supply and exhaust ventilation systems when the indoor fine dust (i.e., PM10 and PM2.5) concentration through the

regression model is lower than the management standard of indoor environmental quality.

Table 1. Management standard of indoor environmental quality in South Korea

Facility type	Classification						
	Temperature (°C)		Humidity (%)	PM10 ^a (μg/m ³)	PM2.5 ^b (μg/m ³)	CO ₂ ^c (ppm)	TVOC ^d (μg/m ³)
	Heating	Cooling					
Educational	18~20	26~28	30~80	100	35	1,000	400
Office	-	-	-	150	-	1,000	500

Note: ^a *PM10* stands for the particulate matter less than 10μm; ^b *PM2.5* stands for the particulate matter less than 2.5μm; ^c *CO₂* stands for the carbon dioxide; and ^d *TVOC* stands for the total volatile organic compounds.

3 Conclusion

In this study, the operational strategy of the ventilation system in a building was established based on the regression analysis and management standard of indoor environmental quality. It is expected that applying the proposed methodology to the ventilation system of a building will improve not only indoor environmental quality but also occupant's health. In addition, since this methodology can be extended, it will be possible to establish an integrated operational strategy of ventilation system in a building considering various indoor and outdoor pollutants such as total volatile organic compounds in the future work.

References

1. Indoor air quality management in multiple-use facilities, <https://seoulsolution.kr/ko/content/2097>
2. Air pollution, <http://www.who.int/airpollution/en/>
3. Indoor air quality management standard in educational facilities, <http://easylaw.go.kr/CSP/CnpClsMain.laf?popMenu=ov&csmSeq=544&ccfNo=4&cciNo=3&cnpClsNo=2>
4. Indoor air quality management standard in office building, <http://easylaw.go.kr/CSP/CnpClsMain.laf?popMenu=ov&csmSeq=544&ccfNo=4&cciNo=3&cnpClsNo=3>

Acknowledgements

This research was supported by a grant (18CTAP-C117226-03) from Technology Advancement Research Program (TARP) funded by Ministry of Land, Infrastructure and Transport of Korean government.

Analysis of interchannel phase connectivity for EEG event-related potentials using auditory oddball paradigm in attention tasks

J.V. Hurtado-Rincón ^{*}, F. Restrepo ^{**}, J.I. Padilla- Buriticá ^{***}, J.D. Martínez-Vargas [†], J.M. Ferrández^{***}, and G. Castellanos-Domínguez^{*}

jvhurtador@unal.edu.co
Manizales, Colombia

Keywords: Brain connectivity, Phase synchronization, Electroencephalography, Oddball paradigm.

Abstract. Nowadays, cognitive stimulus processing using Electroencephalographic (EEG) recordings is accomplished by analyzing individually the time-frequency information belonging to each EEG channel. Nevertheless, several studies have characterized cognitive functions as synchronized brain networks depending on the underlying neural interactions. As a result, connectivity analysis provides essential information for improving both the interpretation and interpretability of brain functionality under specific tasks. In this research, we perform functional connectivity analysis by measuring the stability of the phase difference between EEG channels, aiming to include synchronization patterns for studying the brain reaction to cognitive stimulus. Experiments are carried out in subjects responding to an oddball paradigm. Results show statistical differences between target and non-target labels, making the proposed methodology a suitable alternative to support cognitive neurophysiological applications.

1 Introduction

Studies of routine activities (like attention, perception, or decision-making) have been widely used in cognitive and clinical research of neurological diseases. Because of the provided high temporal resolution and low implementation cost, electroencephalography (EEG) is a very suitable tool to investigate the brain activity patterns, mainly, in applications regarding the oddball paradigm experiments, in which two different testing sensory stimuli (i.e., *rare* and *frequent*) are randomly displayed to a subject under examination. The rare stimuli, labeled

^{*} Signal Processing and Recognition Group, Universidad Nacional de Colombia

^{**} Universidad Autónoma de Manizales, Manizales, Colombia

^{***} Diseño Electrónico y Técnicas de Tratamiento de Señal, Universidad Politécnica de Cartagena, Cartagena, Spain

[†] Instituto Tecnológico Metropolitano, Medellin, Colombia

as a target, are produced with a higher recurrence than the frequent stimuli, or non-target. In the oddball task, an evaluated subject must distinguish either stimulus by pushing a button or mentally counting each target stimulus, while ignoring the non-targets occurrence.

In the stimulus processing, the event-related potentials (ERPs) are measured by averaging all EEG signals collected from a single subject, enabling to interpret his neural response of a brain reaction to a given stimulus. An example of ERP is P300 waveform that is an extended latency component, appearing at 300 - 500 ms as a positive wave after each target stimulus is triggered. Detailed analysis of long latency components is frequently performed by time-frequency methods [10], which are prone to be biased by the powerful individual channels. In fact, the neuronal activity involved in cognitive functions is not evoked by the isolated brain regions, meaning that P300 waveforms are elicited by integrating structurally and functionally different brain areas [6].

Analysis of all neural interactions, termed brain connectivity, is assumed to provide valuable knowledge for better understanding of cognitive and clinical research [2]. In particular, EEG synchronization assessments are often used, which are intended to measure the phase synchronization intensity from neural models. Since the volume conduction effect alters the electrical brain activity measured through scalp EEG, the synchronization values might not reflect brain activities accurately, producing spurious detection of connectivity and resulting in a high dimensional connectivity matrix containing real information and noise.

In this work, we characterize the evoked responses through the functional connectivity quantified by the phase locking value (PLV), computing the pairwise relationships between all possible EEG channels. To investigate confidence of task-induced changes, the extracted PLV features are statistically tested to choose the most relevant ones, tending to identify the most significant EEG channel-to-channel connections; the statistical analysis is performed with the aim of identifying the channels and connections involved during attentional processes and memory work. The results performed on real-world EEG data, relying on the oddball experimental paradigm of cognitive evoked potentials, show that there are substantial differences in connectivity response to the discrimination of target and non-target.

2 Methods

Since the transient linking of different cognitive tasks is associated with the integration and interaction of different brain structures, long-range oscillatory phase synchronization is computed with the aim of searching spatially distributed patterns of coherent neural oscillations at specific frequencies, even if their amplitudes are uncorrelated [9]. To this end, the validation methodology appraises the following stages: *i*) Computation of grand average ERP waveforms, *ii*) Extraction of inter-channel connectivity features using PLV, and *iii*) Estimation of a significant connectivity feature set.

2.1 Experimental paradigm of cognitive evoked potentials

ERP data description: The used EEG data were acquired from 25 children, aging from 5 to 16 years and having different education levels¹. Children had been selected to fulfill the following exclusion criteria: intellectual disability, neurological antecedents, psychiatric hospitalization's history, autism, and related. The child's parents were also requested to sign written permission for authorizing their participation.

The experimental design is implemented using the oddball paradigm that measures the perceptual discrimination between evoked ERPs, in which the cognitive functions such as attention and work memory are necessary. In the concrete case, each subject is instructed to listen to a series of tones and to pay attention to the targets and to count their incidence while ignoring the existence of non-targets. Either investigated ERP elicitation (target or non-target) lasts 130 ms, holding a silent time between two consecutive stimuli of 1 s. The number of target-occurrence is fixed at 20 % from the total trials (namely, 200), while the non-target stimulus – the remaining 80 %. The experiment contains approximately 200 trials.

Grand average calculation: All employed EEG recordings were collected symmetrically using the International 10-20 system for 19-electrode placement. Further, the EEG data were sub-sampled at 250 Hz and segmented in N trials within the time interval, starting from -0.2 s to 1 s of the stimulus onset instant (0 s). Note that the interval measured between -0.2 s to 0 is termed baseline. For each child, a representative ERP data are extracted from the segmented trial set, $\{\mathbf{Z}_n \in \mathbb{R}^{C \times T} : n \in [1, N]\}$, where T is the number of time samples and $C=19$ is the number of channels. Then, the grand average ERP waveform is estimated by averaging the data of the 25 subjects in this experiment, holding a sequence of positive and negative components. With the aim of studying the active brain areas related to the stimuli processing, the spatial distributions of grand average energy are extracted from six-time intervals, starting from -0.2 to 1 s of the stimulus onset instant.

2.2 Extraction of inter-channel connectivity features

Phase Locking Value (PLV): This synchrony measure represents the relations between phases at specific frequencies of coupled oscillating systems in which phases are adjusted as a function of their phase difference even if their amplitudes remain uncorrelated [9]. By using PLV, we quantify the instantaneous phase consistency between two brain activity signals recorded at different locations, relying on the response to a repeated stimulus. Specifically, PLV is employed that evaluates whether the phase difference of a couple of channels c and c' remains constant across the trial set, calculating the stable timing between scalp locations as [8]:

¹ Child population includes students of private and public schools of the city Manizales, Colombia.

$$\gamma_{c,c'}(t) = \mathbb{E} \left\{ \left| \exp(j(\Phi_n^{c'}(f,t) - \Phi_n^c(f,t))) \right| : n \in N, f \in F \right\}, \quad (1)$$

where notation $\mathbb{E} \{ \cdot \}$ stands for the common expectation operator, and $\Phi_n^c(f,t)$ is the instantaneous phase of n -trial, measured at c -electrode over $T \in \mathbb{N}$ time samples and $F \in \mathbb{N}$ frequency bins. Here, the instantaneous phase is computed using the continuous wavelet transform with a Morlet mother wavelet at frequency f as given in [1].

As suggested in [3], we remove the record of ongoing synchronization unrelated to task demands, applying baseline normalization of the computed PLV values by subtracting the mean (calculated for the baseline interval) from every data point. As a result, the lower the normalized PLV value $\bar{\gamma}_{c,c'}(t) \in \mathbb{R}[0, 1]$, the worse the synchronization between a tested couple of neural recordings at different scalp locations. That is, a zero value means a purely random rise and fall, while a unitary value means that one signal perfectly follows another.

In cognitive tasks, longer latency components, such as the P300, is related to the speed of stimulus evaluation since it appears when a cognitive function is performed during discrimination between the elicited stimuli. Nonetheless, the ERP waveform has low-frequency dynamics mostly under 30 Hz. Consequently, the connectivity analysis is conducted for the following four narrow-frequency bandwidths: $\delta \in [2, 5]$ Hz, $\theta \in [5, 8]$ Hz, $\alpha \in [8, 12]$ Hz, and Low- $\beta \in [12, 17]$ Hz.

At each frequency bandwidth, the functional connectivity feature set is pairwise computed over all available channels, yielding the channel-to-channel connection matrix for either testing condition: target $\mathbf{Y}^1 \in \mathbb{R}^{C \times C \times t}$ or non-target $\mathbf{Y}^0 \in \mathbb{R}^{C \times C \times t}$, each one holding elements $\bar{\gamma}_{c,c'}(t)$, respectively. Accordingly, 171 available connectivity features is estimated for each labeled stimulus and frequency bandwidth.

2.3 Estimation of significant connections

The extraction of all possible inter-channel interactions leads to connectivity matrices with a high amount of information, including redundant or worthless features for a specific task. Therefore, it becomes necessary to extract a set of connections to identify the discriminating properties of the estimated connectivity feature set [5]. Consequently, we conduct the statistical comparison of the obtained data for both stimuli. For each stimulus, a paired sample t -test is accomplished on the feature set split into six intervals of time: $-200 - 0, 0 - 200, \dots, 800 - 1000$ ms.

For assessing the connectivity difference with a confidence level across subjects, we use the null hypothesis to state that there are significant differences between the *target* and *Non-target* conditions, validating one by one each extracted feature. Otherwise, the alternative hypothesis asserts that the mean of one stimulus connection among subjects is higher than the another. Nonetheless, the amount of comparisons to be achieved is huge, diverging greatly from one trial to another. To deal with this issue, the set of most significant connectiv-

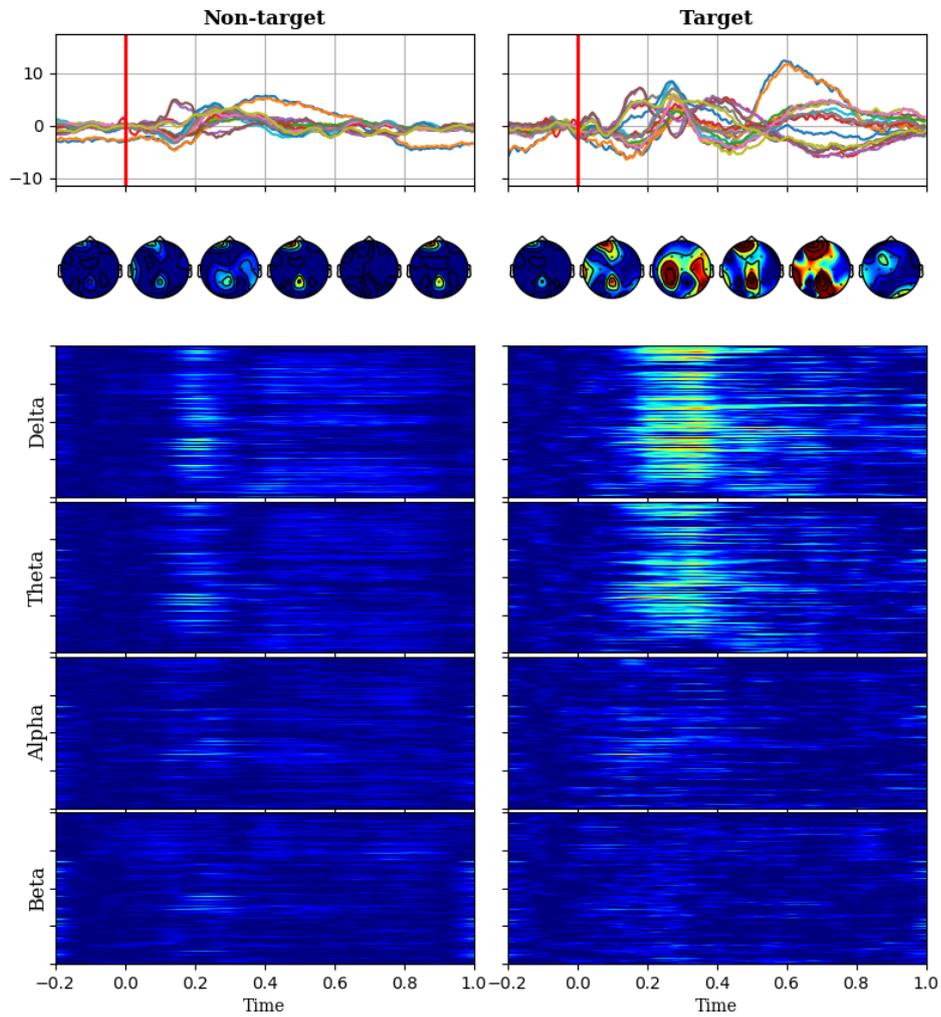


Fig. 1: Average of the estimated ERP waveforms for each one of tested children and PLV results. Top row, ERP Time evolution. Second row, Topograms of the energy spatial patterns. Bottom rows, Averaged PLV time evolution for the studied frequency bands.

ity features is estimated by the Bonferroni correction, for which a connection becomes relevant if the null hypothesis is rejected with a corrected $p < 0.01$.

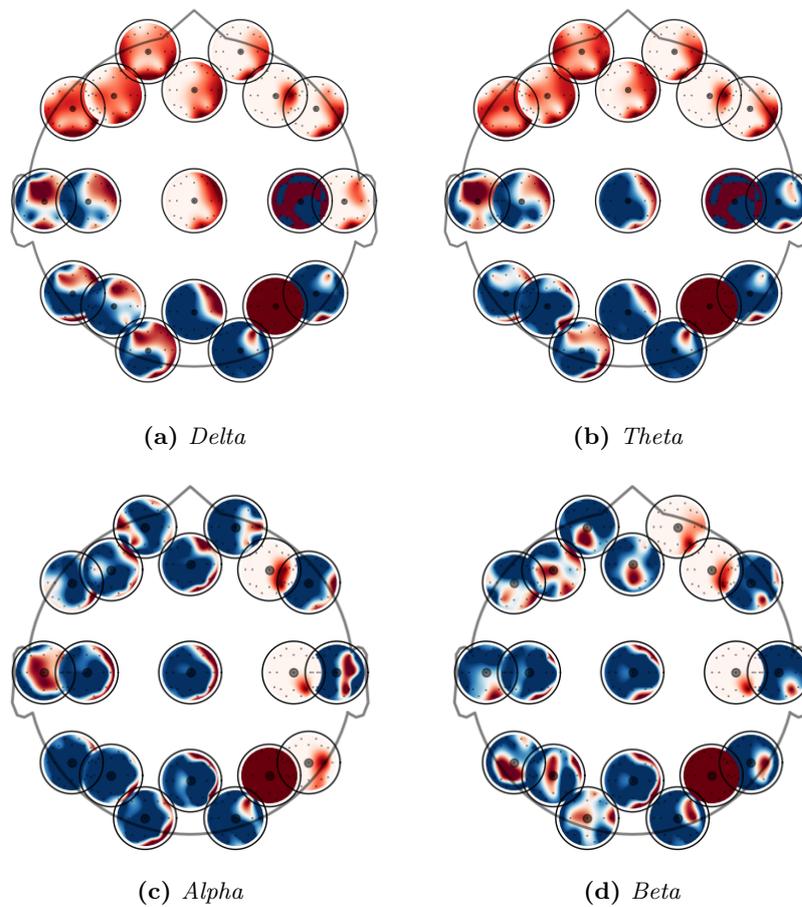


Fig. 2: Plots are showing all the PLV values between 200 and 400 ms averaged by subjects. The figures represent the strength of the relationships between each channel (small circles) against the remaining channels, where the color represents the intensity of the connection between the specific channel and the scalp region (the red color states high intensity and the blue color states low activity), while the outside circle represents the whole scalp.

3 Results and discussion

3.1 Visualization and Analysis of Connectivity

Fig. 1 displays in the upper rows, the grand-average evolution over the time domain together with the topograms that are estimated for each considered time interval, revealing the spatially distributed patterns of neural activity. Thus, regardless the analyzed stimulus, each topogram points out on the almost non-response baseline interval. At the next time intervals, either response increases,

resulting in a higher amplitude of target stimuli because of P300 component and making evident the brain regions that are mostly related to the P300 generation. As seen, the right frontal, temporal and parietal lobes are highly activated. In this regard, the parietal and temporal lobes handle all processing stimulus functions, which reach the central nervous system like the auditory stimulus processing and the discrimination between target and non-target stimuli. In turn, the frontal lobe activations (regarding the attention and working memory [12]) can be detected starting at 200 ms after the target stimulus. This situation holds until the last split interval when neural activity patterns notoriously vanish. As a result, the presence of target ERP waveform can be correlated with the performing cognitive functions: attention and memory work.

As seen in the bottom rows of Fig. 1, the phase connectivity values rise as soon as the stimulus is triggered, making clear an important neural activity as a response to the elicitation. Note that the target responses prompt a higher amplitude in low-frequency bands, improving the discrimination between the target and non-target conditions. This behavior can be promoted by the low-frequency dynamics in ERP responses, reinforcing that cognitive brain processes are related to the low-frequency components of EEG activity [7]. On the opposite, PLV measurements do not seem discriminating between the labels in high-frequency bandwidths.

As proposed in [11], the color graphs of Fig. 2 display the assessed phase interactions in target condition of each electrode with other brain areas in the interval between 200 and 400, ms when the P300 latency takes place. This time interval is selected because of the evident changes in the connectivity patterns seen in Fig. 1. So, the outside circle pictures the whole scalp, while the rings (at each electrode position) embrace the connectivity relationships between a specific electrode and remaining electrodes. As seen, the patterns of high connectivity of the target stimulus, involving the frontal brain areas, are distinguished in the analyzed interval. Moreover, Fig. 2 reveals that delta and theta bandwidths perform similar connectivity, showing high values at the frontal lobes, where the cognitive functions, as attention and memory, are generated to perform the discrimination process between conditions. Nevertheless, the previous behavior does not appear in the high-frequency bands: alpha and beta, where connectivity values are lower.

3.2 Significant connections

Lastly, Fig. 3 displays the connections estimated as significant for the studied cognitive task. The blue lines represent the results of the t-test showing the statistical differences between the groups: target and non-target. The connections were selected if rejected after the FDR correction ($p < 0.01$). The proposed analysis shows that the most informative neural, according to the connectivity patterns, are within 200 and 600 ms. Consequently, the connectivity patterns mostly appear within the time interval when it is expected that the subject is performing the attentional process and memory work. Moreover, the proposed

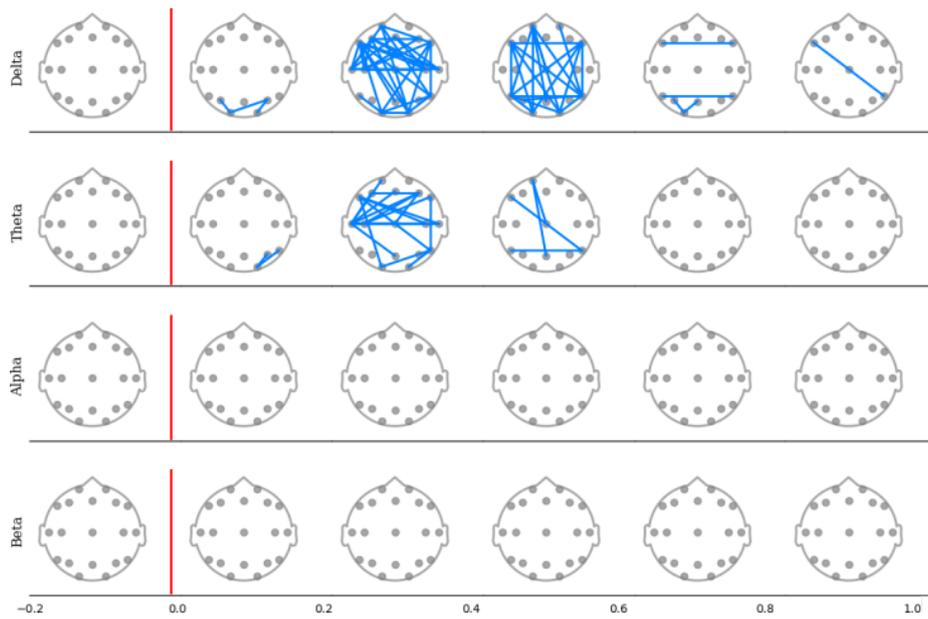


Fig. 3: *Interchannel connections selected that are significantly different between the conditions target and non-target; these connections are presented as blue lines at each time interval and are the rejected PLVs in the performed t-test having higher p-value than the defined significant level (0.01). No significant connections are found in Alpha and Beta bands.*

analysis indicates that the main difference between target and non-target conditions concerning brain connectivity can be found in low-frequency bands, being associated with changes in theta and delta sub-frequency bands and involving connections between the frontal, temporal and parietal channels. These results are congruent with the findings in [4]. Finally, no statistically significant differences are found in the studied high-frequency bands.

4 CONCLUSIONS

In this work, we characterize the functional connectivity for the P300 component using the oddball paradigm. The synchronicity assessment is performed by the phase locking value, computing the pairwise relationships between the measured scalp EEG channels. Because there is a meager amount of significant differences in the phase-locking patterns between the triggered stimuli, we investigate the confidence of task-induced changes, relying on the p -value that is implemented as a statistical analysis. This methodology is used as a measure to identify and select the most significant EEG channel-to-channel connections. With the proposed framework we have studied the statistical differences in the connectivity patterns, and we have analyzed the behavior of phase networks in attention tasks.

The results performed on real-world EEG data show that the channel-to-channel connectivity present distinct temporal patterns for each of the four studied frequency bands. Also, the performed analysis indicates that there are substantial differences in connectivity response in the discrimination of target and non-target stimulus during the attention tasks. Moreover, the findings show that it is possible to differentiate the stimulus with a small set of low-frequency connections, and these connections are associated with the active channels in the time-frequency study of the oddball paradigm (frontal, temporal and parietal).

In general, with the proposed study, we found changes in the functional connectivity patterns that are related to the changes in the well studied ERP behavior. Therefore, we describe, in the sense of connectivity, the structure of the target and non-target auditory stimuli responses. From cognitive neuroscience, this technique contributes to the study of cognitive processes with very accurate temporal resolution.

5 Future Work

As a future work, the proposed methodology should be tested in the source space representation based on the EEG signals, facilitating the physiological interpretation of the active connections and regions during the task. Additionally, brain connectivity networks, using the selected connections should be created to implement graph theory measures, providing further features of brain organization.

ACKNOWLEDGEMENTS

This work is supported by:

- Project 111077757982 funded by *COLCIENCIAS*.
- Programa Nacional de Becas de Doctorado convocatoria 647(2014).

References

1. Bob, P., Palus, M., Susta, M., Glaslova, K.: EEG phase synchronization in patients with paranoid schizophrenia. *Neuroscience letters* 447(1), 73–77 (2008)
2. Brázdil, M., Mikl, M., Mareček, R., Krupa, P., Rektor, I.: Effective connectivity in target stimulus processing: a dynamic causal modeling study of visual oddball task. *Neuroimage* 35(2), 827–835 (2007)
3. Handy, T.C.: *Brain Signal Analysis: Advances in Neuroelectric and Neuromagnetic Methods*. MIT Press (2009)
4. Harper, J., Malone, S.M., Iacono, W.G.: Theta-and delta-band eeg network dynamics during a novelty oddball task. *Psychophysiology* 54(11), 1590–1605 (2017)
5. Hurtado-Rincón, J.V., Martínez-Vargas, J.D., Rojas-Jaramillo, S., Giraldo, E., Castellanos-Dominguez, G.: Identification of Relevant Inter-channel EEG Connectivity Patterns: A Kernel-Based Supervised Approach. In: *International Conference on Brain and Health Informatics*. pp. 14–23. Springer (2016)
6. Ingber, L., Nunez, P.L.: Neocortical dynamics at multiple scales: EEG standing waves, statistical mechanics, and physical analogs. *Mathematical Biosciences* 229(2), 160–173 (2011)
7. Inouye, T., Shinosaki, K., Iyama, A., Matsumoto, Y., Toi, S.: Moving potential field of frontal midline theta activity during a mental task. *Cognitive brain research* 2(2), 87–92 (1994)
8. Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., et al.: Measuring phase synchrony in brain signals. *Human brain mapping* 8(4), 194–208 (1999)
9. Lowet, E., Roberts, M.J., Bonizzi, P., Karel, J., De Weerd, P.: Quantifying neural oscillatory synchronization: a comparison between spectral coherence and phase-locking value approaches. *PloS one* 11(1), 20 (2016)
10. Luck, S.J.: *An introduction to the event-related potential technique*. MIT press (2014)
11. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clinical neurophysiology* 115(10), 2292–2307 (2004)
12. Redolar Ripoll, D.: *Cognitive neuroscience*. Editorial Panamericana, Madrid p. 5 (2014)

Big-Learn 2+: Integrating Apache Spark with Solr Framework to improve the online search in Big Data environment

K. AOULAD ABDELOUARIT¹, B. SBIHI² and N. AKNIN³

^{1, 2, 3} TIMS Research Unit, LIROSA Laboratory
Abdelmalek Essaadi University
Tetuan, Morocco

¹ abdelouarit.karim@gmail.com

² bsbihi@hotmail.com

³ aknin_noura@yahoo.fr

Abstract. In this paper, we study the possibility of integrating the Apache Spark solution with the Solr Framework to improve the process of online search in the Big Data environment. And this, as part of the refinement of our new Big-Learn tool for online search, used by a learner in an e-learning context. The purpose of this study is to evaluate the Spark system in terms of fast data processing, large-scale complex analysis, and also in terms of ease of use, execution and integration of this solution with other layers related to the data search process and especially with the Solr Framework. Apache Spark is considered a better tool in terms of fast processing large amounts of data, as well as in real time analysis. It is in this context that we propose to study the use of the Spark solution with the Solr Framework in order to offer a technique that processes better the massive data and which thus makes it possible to improve and organize the results of the online search. Our solution is based on the combination of two open source technologies from the Apache series (Spark and Solr) for the processing and searching of massive data, in addition to the use of the Lucene engine for data indexing.

Keywords: Big Data; Online Search; Spark; Solr, Lucene, MapReduce.

1 Introduction

The explosion of data on the Internet and especially on the Web has made it difficult for traditional search engines to find interesting relationships between objects and to analyze and extract knowledge from the raw data generated by the Big Data phenomenon [10]. This article builds on our previous work on designing and implementing an architecture model for online search in the Big Data environment used by a learner in a training or distance learning context [4]. As we presented in our previous work, it is a question of designing a complete system which relies on the collection of data corresponding to the user request of the online search, then these data are processed by the MapReduce technique and stored on the Hadoop Distributed File System (HDFS).

Then, this data will be indexed by the Lucene engine so that it can be easily queried and returned to the user using the Solr Framework [2]. However, our solution architecture model as presented can be improved to cover a better process both in terms of speed of data processing and in terms of complex and large-scale analysis [13]. The following section describes the use of Apache Spark in the Big Data environment by introducing the Spark system architecture specializing in massive data processing; Paragraph 3 introduces the concepts and approaches of the Spark technique for large data processing with an example of use. We then present in paragraph 4 the Big-Learn solution, based on the Apache Spark tool and integrated into the Solr Framework, to build a complete online search system, which aims at the fast processing of raw data generated by the Big Data layer and which allows a better organization and customization of the search results to thus offer an optimal presentation of the information for the online search user. The last paragraph presents a general conclusion outlining a series of perspectives.

2 The use of Apache Spark for online search in the Big Data environment

2.1 Using a Spark-based model to improve the online search process

Nowadays, traditional search engines are no longer able to process the vast amount of data circulating on the Internet because of the emergence of the Big Data phenomenon [10]. The processing of this amount of massive data requires sophisticated algorithms capable to handle a large number of queries [2]. Among the Big Data processing technologies, we find the Apache Spark tool, that it is an open source project designed to perform sophisticated analysis with ease of use and fast processing of massive data in real time [11].

With its data processing technology, Spark is able to cache data in memory using its Resilient Distributed Dataset (RDD) concept. These properties allow Spark to be faster for iterative algorithms because it avoids the use of the disk. The RDD is defined as a collection of records that can only be created from a stable storage or through transformation from other RDD's. The other features of an RDD are that it's a read-only after its creation and partitioned.

In Apache Spark, the programmer can access and create RDD's through the in-built language API either in Scala or Java. Indeed, the programmer can choose to either create the RDD from storage, a file or several files that is stored in any of the supported storage options like the HDFS. The other option is to create an RDD through transformations such as map and filter, their primary use is often to create new RDD's from already existing RDD's [1].

Therefore, in our architecture model for the online search process, the Spark system will be able to replace the conventional data collection and data mining layer with its distributed and real-time data processing technique and thus improve the operation of processing massive data [2]. Figure 1 shows the architecture model of the online search solution in the Big Data environment by integrating the Apache Spark system.

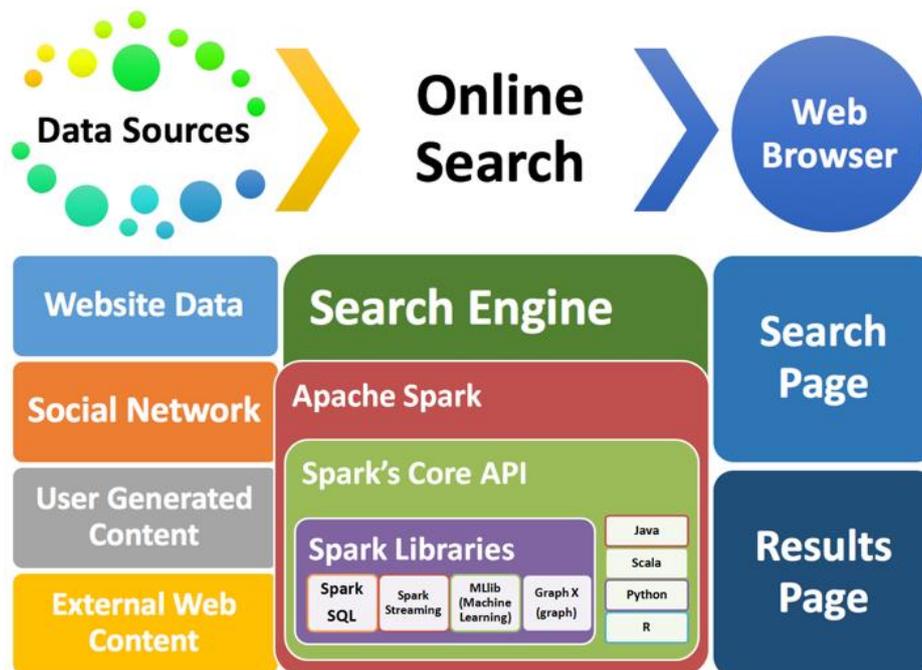


Fig. 1. Online search model using Spark in Big Data environment.

As shown in this figure, the user accesses the online search page via the Web browser to initiate his query. The search engine intercepts the user's request and starts searching the data on the Internet based on the entered keywords. At this level, the Spark system will intercept the collected data and process it at the same time it is stored in memory, without transmitting it to another engine as required by the previous Big Data processing systems like Hadoop's MapReduce technique. The collected data is referenced and indexed before presenting it to the user on the results page. The analyzed data includes all raw data from the Internet such as: websites, social networks, user-generated content and other external data sources than web data.

2.2 The architecture of Apache Spark

Apache Spark is today the most popular and used Big Data framework. Basic programming abstraction is a type of immutable, implicitly distributed collection data called Resilient Distributed Dataset (RDD). RDDs serve as high-level programming interfaces and transparently manage fault tolerance [12].

Spark libraries.

Spark allows to quickly developing applications in Java, Scala or Python since it uses a unified API. In addition, it can be used interactively to query data from a Shell.

Finally, Moreover, Apache Spark supports additional libraries to work with Big Data and Machine Learning [13]. These libraries are:

- **Spark SQL:** is designed to execute SQL queries;
- **Spark Streaming:** for real-time processing of flow data;
- **Spark MLlib:** is a Learning machine library that contains all classical learning algorithms and utilities;
- **Spark Graph X:** It allows processing graphs.

Although Apache Spark is a new platform, it is constantly evolving with new additions and it has already been adopted as a real-time processing framework in many big companies, such as: Amazon, Yahoo, Samsung, Nokia, IBM, eBay, etc [13].

Resilient Distributed Datasets.

Spark is designed to cooperate with Hadoop, especially through the use of its HDFS storage system. To resolve these disadvantages, Spark introduces an abstraction called Resilient Distributed Datasets (RDD) that represents a read-only collection of partitioned objects across a set of machines. It allows loading a data set in memory once and reading multiple times of the executor process without having to load it in each iteration as happens with Hadoop [8]. Not only introduces Spark an in-memory storage but also it provides a complete set of programming primitives allowing users to implement much more complex distributed applications than those implemented in the classic MapReduce. It allows implementing several distributed models like MapReduce, SQL like or even Pregel [10].

Resource Management.

Spark enables the development of complex, multi-step data processing pipelines and data sharing in memory so that multiple jobs can work on the same dataset. Spark runs on HDFS infrastructures and offers additional features. Instead of seeing Spark as a replacement for Hadoop, it is more correct to say that it is an alternative to the MapReduce. Spark was not intended to replace Hadoop but to provide a complete and unified solution for Big Data processing. Figure 2 below shows these components of Spark architecture model [13].

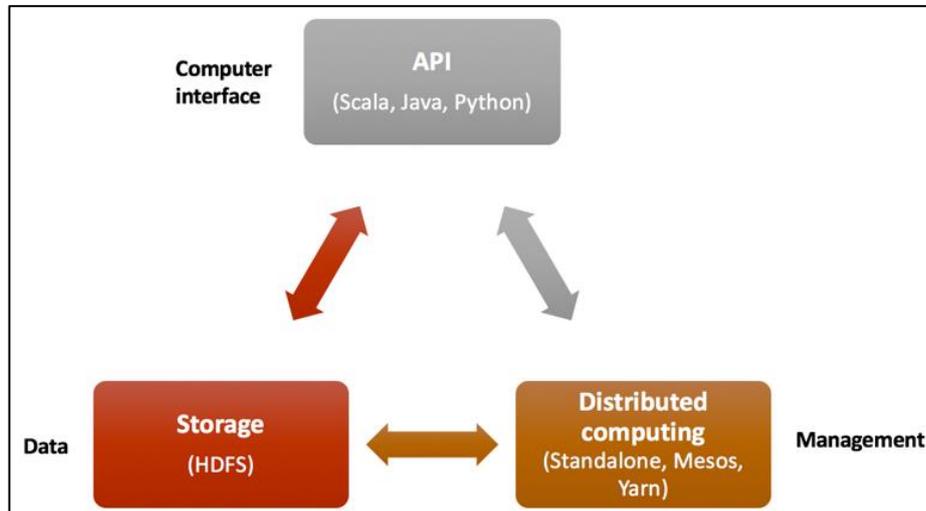


Fig. 2. The technical architecture used by Apache Solr System.

As shown in this figure, the architecture of Spark is based on three essential components:

- **Computer interface:** Spark allows to quickly developing applications in Java, Scala or Python since it uses a unified API. In addition, it can be used interactively to query data from a Shell.
- **Data storage:** Spark uses HDFS file system to store data like Hadoop.
- **Management:** Spark can be deployed as a standalone server or it can be on a distributed computing framework like Mesos or YARN.

3 Processing massive data with Apache Spark

3.1 How Spark processes the massive data

Apache Spark is another increasingly popular alternative to replacing MapReduce with a more efficient execution engine, but still uses Hadoop HDFS as a storage system for large datasets. Spark brings improvements to MapReduce through less expensive Shuffle steps. Also, using memory storage and real-time processing, performance can be faster than other Big Data technologies [9]. Spark maintains the intermediate results in memory rather than on disk, which is very useful especially when it is necessary to work several times on the same dataset. The runtime is designed to work both in memory and on disk. Operators run external operations when the data does not fit in memory; this allows larger datasets to be processed than the aggregated memory of a cluster. Spark tries to store as much memory as possible before switching to disk. Indeed, it allows to work with one part of the data in memory, another one on disk. It is necessary to review its data and use cases to assess its memory requirements, as

Spark can offer significant performance benefits based on the work done in memory [11].

Spark was normally designed to work with static data through its Resilient Distributed Datasets (RDD). Spark uses batch program to deal with streams. This technique divides incoming data and processes small parts one at a time. The main advantage of this approach is that the structure chosen by Spark, called DStream, is a simple queue of RDDs. This technique allows users to switch between streaming and batch as both have the same API. Unlike Hadoop MapReduce, Spark has support for data re-utilization and iterations. Spark stores data in memory through iterations via explicit caching. However, Spark perform its executions as acyclic graph plans, which implies that it needs to schedule and run the same set of instructions in each iteration [5]. The following Figure 3 present the technique used by Apache Spark and to the implementation of the MapReduce programming model.

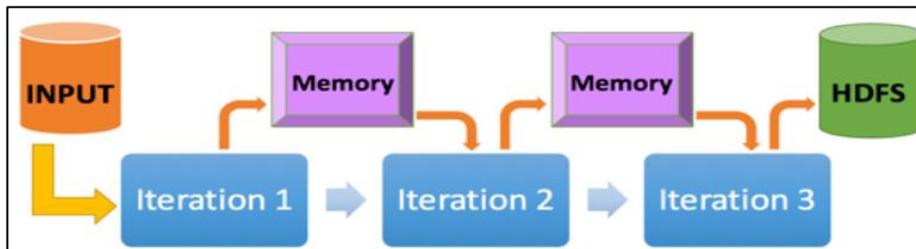


Fig. 3. Processing data by Apache Spark.

As described in the figure, Spark performs operations in memory by copying the data from the distributed storage to RAM which is much faster. Consequently, the Read/Write operation time is reduced. However, at the Hadoop-MapReduce the process tends to be slow because each MapReduce operation stores the data on the disk. As a result, multiple requests on the same dataset read data separately and create a lot of Read/Write operations on the disk [6][7]. Spark allows other features that include:

- More Features other than Map() and Reduce() functions;
- The optimization of graphs of arbitrary operators;
- The fast evaluation of queries, which helps to improve the overall processing workflow;
- A consistent APIs in Scala, Java and Python;
- An interactive Shell for Scala and Python (that is not yet available in Java).

Apache Spark is written in Scala language and runs on the JVM (Java Virtual Machine). The currently supported languages for application development are: Scala, Java, Python, Clojure and R.

3.2 Example of processing data with Apache Spark

Spark introduces an abstraction called Resilient Distributed Datasets (RDD) that represents a read-only collection of partitioned objects across a set of machines. We can imagine the RDD as a table in a database. It can hold any type of data. Spark stores data in RDD in the form of partition. They are also fault tolerance because an RDD know how to recreate and re-compute the datasets. RDDs are immutable. You can modify an RDD with a transformation, but the transformation returns you a new RDD whereas the original RDD remains the same. RDD contains two types of operations:

- **Transformation:** This operation doesn't return a single value, they return a new RDD. Nothing gets evaluated when you call a Transformation function, it just takes an RDD and return a new RDD. Some of the Transformation functions are map, filter, flatMap, groupByKey, reduceByKey, aggregate-ByKey, pipe, and coalesce.
- **Action:** This operation evaluates and returns a new value. When this function is called on a RDD object, all the data processing queries are computed at that time and the result value is returned. Some of the Action operations are reduce, collect, count, first, take, countByKey, and foreach.

3.3 How to Interact with Spark

Once Spark is installed and running, we can connect to it using the Spark shell for interactive data analysis. Spark Shell is available in both Scala and Python languages.

We need to use the commands “spark-shell.cmd” and “pyspark.cmd” to run Spark Shell using Scala and Python respectively.

When Spark is running in any mode, we can view the Spark job results and other statistics by accessing Spark Web Console via the following URL: <http://localhost:4040>.

Figure 4 below show the Spark Web Console with tabs for Stages, Storage, Environment, and Executors.

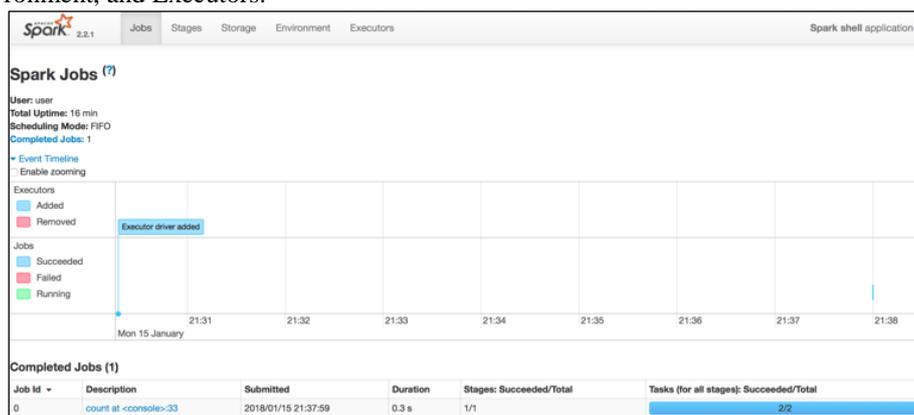


Fig. 4. The Spark web console.

As shown in this figure, the Spark Web console offers five tabs:

- **Jobs:** it is a combination of multiple tasks.
- **Stages:** each job is divided into smaller set of tasks called stages. Each stage is sequential and depend on each other.
- **Storage:** this tab is for the RDD size and memory use.
- **Environment:** for environment information (variables, configuration, ...).
- **Executors:** It is a process launched on the worker node that runs tasks. It uses worker node's resources and memory.

3.4 Sample Spark Application: Text Search

The sample Spark application covered in this article is a simple Text search application that will perform a search through the error messages in a log file and we'll use the Spark Scala Shell.

The text file and the data set in this example are small, but same Spark queries can be used for large size data sets, without any modifications in the code.

First, we open a new Spark Scala Shell and we will run these commands for our example:

```
val txtData = sc.textFile("hdfs://...")
txtData.cache()

//Creates a DataFrame having a single column named "line"
val df = txtData.toDF("line")
val errors = df.filter(col("line").like("%ERROR%"))
```

We call the cache function to store the RDD created in the above step in the cache, so Spark doesn't have to compute it every time we use it for further data queries. That cache() is a lazy operation. Spark doesn't immediately store the data in memory when we call cache. It is performed when an action is called on an RDD.

Now, we can call the count function to see how many lines contains the errors in the log file:

```
// Counts all the errors
errors.count()
```

Finally, we can run the following commands to perform the text search count for any MySQL error. The count shows up next to each found word in the text file.

```
// Counts errors mentioning MySQL
errors.filter(col("line").like("%MySQL%")).count()

// Fetches the MySQL errors as an array of strings
errors.filter(col("line").like("%MySQL%")).collect()
```

We can also combine the Spark processing with Spark SQL, Machine Learning and Spark Streaming. With several integrations and adapters on Spark, we can combine other technologies with Spark like the Solr Framework to build a complete online search system that processes massive data generated from the Big Data layer.

4 Towards a system based on Spark and Solr to improve the online search

4.1 The integration of Apache Spark with the Solr Framework

The Big-Learn system for online search in Big Data environment, as already presented in our previous work [2] intercepts data from the Big Data layer via the Hadoop system which stores them in its HDFS storage system after their processing through its MapReduce technique. Thus, the data is loaded and transformed during the Map() phase, and then combined and saved during the Reduce() phase to write the Lucene index. The Lucene layer reads the data stored in HDFS, and stores it using Lucene Scheme, which in turn records the data as Lucene document in an index. Once all files are indexed to the Lucene layer, their queries are possible via the Solr layer. However, based on our assessment of both Spark and Hadoop solutions with respect to mass data processing, it is now essential to see Spark as a replacement for the Hadoop-MapReduce technique in the processing of raw data generated by the Big Data layer. Figure 5 shows the new architecture of the Big-Learn system after integrating Apache Spark system.

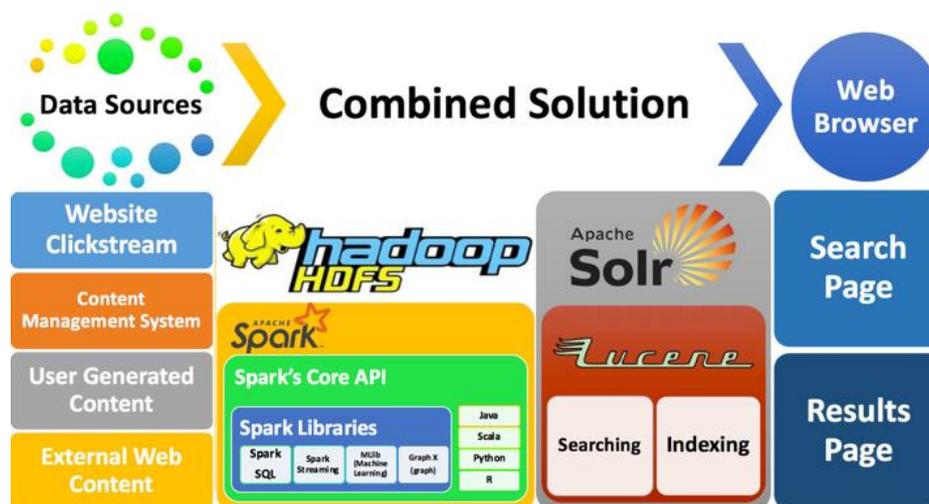


Fig. 5. Integrating Apache Spark in the Big-Learn system.

As shown in this figure of the new architecture, data coming from the Big Data will be intercepted and processed by the Apache Spark system this time and will always

use the HDFS system for storing processed data. These data will be indexed at the Lucene layer so that they can be interrogated via the Solr interface.

4.2 Results and discussion

Our study consists of the performance analysis between Apache Spark and MapReduce and how they can be applied for processing massive data and integrated with our Big-Learn Solution. Indeed, Both Spark and MapReduce are giants in the open-source big-data analytic world. At first glance, we have the old and refined MapReduce framework that has many implementations in different languages, but on the other side we have a new but very promising framework Apache Spark that has one implementation but can be executed almost everywhere and can be seen as more flexible in that approach.

One of Apache Spark's key advantages is its ability and flexibility in working with all types and formats of unstructured data coming from different data sources such as text or CSV to well-structured data such as relational database. On the other hand, Apache Solr is a fast search platform built on top of Apache Lucene.

The combination of Apache Spark and Solr allows easily exploring a lot of data, and then providing the results quickly via a flexible search interface. Using this combination of open-source frameworks in a distributed environment is proven in this paper as a feasible approach to improve the online search process using large data sets.

5 Conclusion and future work

In this article, we have seen how the Apache Spark Framework, with its standard API, helps us in processing and analyzing data. We have seen also how Spark positions itself in relation to traditional MapReduce implementations like Apache Hadoop. Spark relies on the same file storage system as Hadoop, so it is possible to use Spark and Hadoop together in case applications have already been made with Hadoop. Spark can also combine other types of processing via its libraries such as Spark SQL, Spark Machine Learning and Spark Streaming. Thanks to its different modes of integration and adapters, Spark also allows interfacing with other technologies like Java, Scala, Python, etc.

As a perspective of this work, we will implement the scenario of using online search with this new architecture model of the Big-Learn solution. Then, in a second step, we will study the degree of integration and participation of our solution to improve learning and scientific research for students and which will take as case study students from Abdelmalek Essaadi University [3].

References

1. Andersson, L. 2016. Natural Language Processing In A Distributed Environment: A comparative performance analysis of Apache Spark and Hadoop MapReduce.

2. Aoulad Abdelouarit, K., Sbihi B. and Aknin N., 2016. Towards an approach based on hadoop to improve and organize online search results in big data environment. In Communication, Management and Information Technology: Proceedings of the International Conference on Communication, Management and Information Technology (ICCMIT 2016) (Jul. 2016), 543-550. CRC Press.
3. Aoulad Abdelouarit, K., Sbihi B. and Aknin N., 2016. Solr, Lucene and Hadoop: Towards A Complete Solution To Improve Research In Big Data Environment (Case of The UAE) MEDITERRANEAN CONGRESS OF TELECOMMUNICATIONS (CMT'2016), 12-13 May, 2016.
4. Aoulad Abdelouarit, K., Sbihi B. and Aknin N., 2015. Big-Learn: Towards a Tool Based on Big Data to Improve Research in an E-Learning Environment. International Journal of Advanced Computer Science and Applications (IJACSA) 6(10): 59–63.
5. García-Gil, D., Ramírez-Gallego, S., García, S., and Herrera, F., 2017. A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. Big Data Analytics, 2(1), 1.
6. Kolosov, I., Gerasimov, S., and Meshcheryakov, A., 2017. Architecture of processing and analysis system for big astronomical data. arXiv preprint arXiv:1703.10979.
7. Kulkarni, A. P. and Khandewal, M. 2014. Survey on Hadoop and Introduction to YARN. International Journal of Emerging Technology and Advanced Engineering, 4(5), 82-87.
8. Lenka, R. K., Barik, R. K., Gupta, N., Ali, S. M., Rath, A., and Dubey, H. 2016. Comparative Analysis of Spatial Hadoop and GeoSpark for Geospatial Big Data Analytics. arXiv preprint arXiv:1612.07433.
9. Mavridis, I., and Karatza, H. 2017. Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. Journal of Systems and Software, 125, 133-151.
10. Padillo, F., Luna, J. M. and Ventura, S. 2017. Exhaustive search algorithms to mine subgroups on Big Data using Apache Spark. Progress in Artificial Intelligence, 1-14.
11. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... and Ghodsi, A. 2016. Apache Spark: a unified engine for big data processing. Communications of the ACM, 59(11),56-65.
12. Essertel, G. M., Tahboub, R. Y., Decker, J. M., Brown, K. J., Olukotun, K., & Rompf, T. (2017). Flare: Native Compilation for Heterogeneous Workloads in Apache Spark. *arXiv preprint arXiv:1703.08219*.
13. Aoulad Abdelouarit, K., Sbihi B. and Aknin N., 2017. Improving Online Search Process in the Big Data Environment using Apache Spark. Proceeding of The Mediterranean Symposium on Smart City Applications (SCAMS 2017), 25-27 October, 2017.

Keyword Index

A Two Sector DSGE model	907
additive outliers	42, 1011
AHP	559
Air pollution forecast	264
air quality monitoring network	1017
Airport	76
Alexandrium minutum blooms	1108
altimetry	1063
analysis of time series	645
Andalusia	1406
ANN	182
Anomalous signal	1038
anomaly detection	982
Anomaly Detection	1417
ANP	559
Ant colony optimization	1680
anthesis	1026
Application Programming Interface	64
Applied Time Series	239
ARIMA	586, 895, 1565
ARMA	54
ARMA model	1110
Artificial neural network	1680
Artificial Neural Network	1091, 1660
Artificial Neural Networks	385, 1468
artificial neural networks	182
Associative Search Models	783
atmospheric energy budget	1041
Augmented Dickey-Fuller test	933
Autocovariance function	42
automatic classification	1270
automation	1469
Autoregressive Hilbertian Process	665
B-splines approximation	966
backtesting VaR	99
Baltic Dry Index	111
Bankruptcy	1495
Bayes-estimation	907
Bayesian approach	1340
Bayesian dynamic STAR models	966
Bayesian forecasting models	966

Bayesian maximum entropy	260
bifurcation	1089
bifurcation diagrams.	1640
Big Data	158, 1355, 1738
bilinear model	1110
BINAR(1)	1135
Binary Resource Model	1429
Binomial thinning operator	997
blind source separation	469
Border-inspection	76
Brain Computer Interface	499
Brain Computer Interfaces	64, 1281
Brain connectivity	1728
Brain mapping	519
Brain-computer interface	489
Breaks	1193
Breast Cancer diagnosis	331
buoy	1385
business context	1589
Business Cycle	1565
Business cycle	1051
Caatinga	95
capacity planning	302
Carbon dioxide output	130
Cardiovascular Disease	318
Center kernel alignment	1316
Central Europe	1111
Cerebral Asymmetry	1612
Cerebral Metabolism	1612
Change Point Detection	408
changepoints	300
Changes	1013
Channel Selection	1344
chaotic Hénon map	1640
Chinese markets	332
Classification	1495
classification	1166
Climate	1406
Cloud workload	982
Clustering	1091, 1147
co-integration	332
Coherence	1147
Cointegration	1193
Cointegration Breakdown	1193
Cointegration Rank	933
collaboration network	1650
Collaborative judgment	559

Color Perception	1612
Coloured Noise	654
Common Spatial Patterns	1281
complex networks	90
complex system	1010
Component selection	499
Composite likelihood	997
compression algorithms	90
computational topology	1367
computer aided diagnosis	331
Conditional Heteroskedastic	825
Conditional inference trees	20
Connectivity analysis	1308
Constant work-rate test	130
Constraint Programming	982
continuous-time chaos	637
COPD	130
Copula	1519
Copula Model	598
Correlated predictors	20
correlation	1063, 1158, 1265
Count data	1519
covariate selection	1469
Credit indices	1178
crime forecasting	432
crime prediction	1166, 1554
Critical Success Index	1111
cross-correlation	192
Cross-covariance function	1011
cryptanalysis	637
Cyber Security	1147
Data	76
Data Mining	444
Data mining techniques	677
Data Quality	729
Data Stream Processing	408
DAX	1113
DCE-MRI	331
Decision	876
Decision Making	1451
decision rules	864
Decision Tree	373
decision trees	1166
Decision Trees	677
Deep-learning algorithm	1015
demand forecast	170
Density forecasting	1340

Diffusion process	325
Dirac equation	952
Direct	385
Direct and Combined Measurement	1693
Disagreement measures	1543
Discrete Wavelet Transform	64
dissimilarity visualisation	717
Distribution Transformers	1693
Distribution Utilities	1693
Dynamic Diff-in-Diff	1300
Dynamic Nelson-Siegel Model	798
dynamical systems	1089
e-Learning	1738
early warning signal	1089
Echo State Networks	883
Econometric and Forecasting	795
Econometric models	811
Econometrics	239
economic growth	227, 921
Economic uncertainty	1543
Education area	1574
EEG	469, 1344
EEG inverse problem	509
eeg time series	1270
Electricity prices	852
electrochemical impedance spectroscopy	625
Electroencephalograms	64
Electroencephalography	1291, 1728
electroencephalography	1270
EMA	529
Empirical Mode Decomposition	134, 519, 1660
employment	362
encryption algorithm	637
Energy	1703
Energy consumption	350
energy forecast	170
Energy prices	146
Energy score	852
Energy simulation	1720
Energy-Macroeconometric modeling	795
Enrollment and re-enrollment	1574
Ensemble forecasting	1660
ensemble methods	331
Ensembles	895
Entropy	1344
Epidemiology	318
Epilepsy	327, 519

equivalent circuit	625
Error correction model	753
Error measures	559
estimation	42
Estimation algorithm	753
Estimator	825
evapotranspiration	1041
Evapotranspiration	665
Event study	1300
excess mortality	1111
Failure Prediction	729
FARIMA	654
fault modelling	1469
fBm	654
Feature selection	1291
Feature Selection	1680
Feedback Trading	1009
Filtered Historical Simulations	1009
Financial Forecasting	1009
Financial Time Series	1660
Financial time series analysis	1178
Fine dust	1724
Finite element method	1720
Firefly algorithm	325
FIS/ANFIS	1385
flood risk	260, 1374
fMRI	469
forbidden ordinal patterns	457
forecast	1328
Forecast	318
Forecast combinations	583
Forecast comparison	811
Forecast evaluation	811
Forecast performance	811
Forecast scheme	529
Forecasting	32, 119, 385, 586, 1036, 1443, 1468
forecasting	300, 302, 1096, 1385, 1670
Forecasting Electricity Consumption	373
Forecasting evaluation	852
Forecasting Maritime Climate	598
forecasting of yield curves	1483
Forecasting techniques	1574
Fourier Analysis	1612
Fourier series	1110
Fourier Transform	239
FPGA	883
fractal dimension	1113

Fractional Cointegration	753
fractional integration	288
fractional integration.	1
Frequency	1612
Fuel economy	350
Functional Data	665
functional decomposition	1670
Fuzzy time series	420
GDP	300
generalized additive model	1554
Genetic algorithm	396
Geodetic Time Series	654
geographical factors	1125
Geometric	1135
Geometry	1543
geosciences	645
Gibbs Sampling	798
Globalization	192
Glucose Metabolism	1612
GQL	1135
GRACE	1601
Granger causality	332, 614
Groundwater level	1073
growth and yield	1026
growth model	921
Growth Models	1443
Guadalquivir Estuary	1374
Hadoop	1417
HAR models	583
health economics	306
healthcare	302
heat demand	170
heat stress	1111
Heavy Tails	654
Heterogeneous Agents	1009
hindcast	1385
Holt–Winters model	1073
human capital	227
Hurst exponent	1113
Hurst Exponent	111
Hurst Parameter	654
Hybrid forecasting method	1038
Hybrid forecasting models	559
Hybrid model	1073
Hybrid Models	1431
hybrid site index	1125

Hyperbolastic curve	325
identification of nonstationary processes	741
ignorance score.	550
Imagined Speech	64
Imbalanced data	1495
income	306
Independent component analysis	469
Inductive Knowledge Base	783
Inflation	1036
influential statistics	432
Info-Metrics	1355
infrared sensors	182
Instance selection.	1291
institutional quality	227
intracranial Electroencephalography (iEEG)	327
intrinsic excitability measures (IEM)	327
investment	921
IR	182
IT governance	1650
JDemetra+	20
Jump diffusion model	852
Kalman Filter	994, 1417
Kalman filter	1507
Kalman Filtering	1531
Kernel adaptive filters	1316
Kernel Methods	1468
Knowledge Base Prediction	1429
Kupiec Test	99
Label variables	373
Lagged phase synchronization	1308
Large Dataset	1468
Lean Energy Analysis	158
Learning on a memory budget	1253
Learning with memory constraint	1253
Levy Processes	654
Light Stimulation	1612
Limited Data	665
linear regression	1601
linear regression model	614
Linguistic activity	64
Load Factor	1693
Local fractal	645
local optimization	250
Local Signal Energy	134

location-based applications	1471
log determinant	920
log score	920
logistic regression	1166
Long Memory	825
long memory	1, 288
Long Segment	444
long VAR models	288
Long-Short term memory network	1015
long-short term memory recurrent neural networks	1036
Losses	1693
low-carbon energy systems	92
LSTM	264, 895
Lucene	1738
Lyapunov exponents	1640
M-cross-periodogram	1011
machine learning	982, 1483
Machine Learning	571
Machine learning	264
Macroeconomic forecasts	811
Macroeconomics	197
Magnetic field	1010
management	1374
Management	1451
Management consulting	876
Mann-Kendall trend test	1601
Mapreduce	1738
Market Impact	1300
Marketing	1451
Markov Regime Switching	798
Markov Switching model	146
Markov Test	99
MATLAB	994
Maximum Entropy	1355
maximum likelihood	1471
Maximum likelihood optimization	1178
Maximum mean discrepancy	1316
MCDM	559
mean square error	920
MEG Inverse problem	1308
Meta-Learning	895
migration	227
MLE	753
MLP	264
Monte Carlo tests	1342
Motor Imagery	1291, 1344
Multi-equational models	1340

Multi-function window	1720
Multi-Instance Learning	1291
Multi-signal Wavelet Packet	509
Multi-state Model	318
Multi-Step ahead	385
Multichannel data	1316
Multidimensional Hawkes process	1178
Multilayer Perceptron	677
Multiple Kernel Learning	1281
multiple term structures	1483
Multivariate GARCH	852
Multivariate models	997
multivariate models	420
Multivariate Time Series	571
Multivariate time series	1519
naive Bayes classifier network	1108
NDVI	95
Network Monitoring	1147
neural network	432, 1367
Neural Networks	1576
neural networks	1265, 1483
Neuroimaging	509
NIST tests	1640
Nitrogen	1013
noise	457
Non-convex optimization	1178
Non-parametric Density Difference Estimation	408
non-standard analysis	952
non-stationarity	300
Non-Stationary	1135
Non-tradable Share (NTS) Reform	332
Nonlinear AR models	966
nonlinear time series analysis	550
Nonstationarity	1193
nonstationarity	1
normal mixtures	1471
O&M	1328
Oddball paradigm	1728
Offset Printing Machines	729
Oil prices	1051
Oil Rate Forecasting	1703
Oil-Dollar	1028
One-pass Learning	1253
Online Searching	1738
Operational strategy	1724
optimal control	864

optimal embedding dimension	457
optimisation	1469
Option Valuation	1009
Out-of-sample performance	1576
Outlier Prediction	571
Outsourcing	876
Oversampling	1495
Oxygen uptake	130
panel data	306
parametric model	250
parametric spectrum estimation	741
Pareto analysis	158
partial least squares	1270
particulate matter (PM)	1017
Passenger-information	76
pattern detection	982
Pattern similarity	602
Peak period	119
pension reform	921
perfect model scenario	550
Performance	1443
performance evaluation	1589
Performance Index	444
Performance measurement	876
periodical components	1096
permutation entropy	192
Permutation entropy	489
Persistence	385
Phase synchronization	1728
phase synchrony	327
physical and meteorological parameters	1108
Pinus densiflora	1125
Point and counting processes	1178
point of interest	1166
Polynomial forecasting models	966
polysomnography	1270
potential analysis	1089
power law	1113
Power Spectrum	1406
power spectrum	469
Power systems	602
precipitation	1041
Prediction	1495, 1519
prediction algorithms	182
prediction error	614
Predictive modeling	1443
Prevalence	318

Preventive Maintenance	444
price forecasting	1576
Principal Component Analysis	1468
principal component analysis	1270
PRNG	1640
Probabilistic Forecasting	852
probabilistic prediction	550
Probit model	1051
production forecast	170
Profitability	876
prognosis	1041
Prophet	586
PSM Model	1300
Public Health Modelling	318
qPCR	325
R-peak detection	134
R/S analysis	111
random forest	1166
random number generator	637
rapeseed	1026
Rapeseed	1013
Rate Transient in Unconventional Reservoirs	1703
Rational spectrum	54
Re-Commissioning	158
realized variance	1342
reanalysis	92
Recession	1051
Recurrent neural networks	119
Recursive	385
Reduced Rank Regression	753
Regression analysis	1724
remittances	227
renewable energy	1328
renewable energy forecasting	1589
renewable grid integration	92
Renyi Entropy	499
Reservoir inflow forecasting	1015
ReservoirComputing	1253
residual analysis	717
Residue calculus	54
Resonance	1612
Return Volatility	825
risk assessment	1650
river discharge	1096, 1374
Road congestion	119
robust estimation	1011

Root Cause Analysis	729
S&P 500	1113
SARIMA model	95
Sarima Model	665
satellite	1385
scenarios	921
sea level	1096
Seasonality	396
seasonality	1017, 1022
Sector Specific Parameters.	907
secure communication	637
selection of estimation memory	741
selection of model order	741
SETAR	1565
Shale Oil	1028
Short term load forecasting	420, 602
Short-term electricity load forecasting	966
Signal Analysis	519
Simulation	32, 76
simulation	302, 1026
simulation experiment	933
Simulation study	20
singular spectrum analysis	1096
site index model	1125
sleep disorders	1270
SMA	529
Smart cities	717
Smoking	318
social network analy-sis	1650
solar heat production	170
Solar photosphere	1010
Solar radiation forecasting	1680
solar time series	92
Solr	1738
sovereign credit default swap	1507
sovereign default intensities	1507
space weather	1158
Spark	1738
Sparse Bayesian Learning	1531
spatio-spectral relevance	489
spatio-temporal patterns	1471
spatiotemporal random field	260
Spectral Methods	1147
Spectrum	1013
Spectrum Analysis	1612
stability	1367
Stacking	895

standard model	952
state of charge	625
state space model	1507
State Space systems	994
stochastic component	1110
Stochastic processes	146
stochastic volatility	1342
Stochastic volatility	1340
Stock Index Prediction	1660
Storm characterization	260
Storm Evolution	598
Strategy	1451
Structural breaks	583
structural breaks	306
Structural changes	396
Structural Models	994
Subjects identification	64
subspace algorithms	288
substations	182
supervised learning	982
Supervised machine learning	20
Supervised Temporal Patterns	1344
suppliers management	1650
support vector machine	1270
support vector machines	1483
Support Vector Regression	1660
Survey expectations	1543
Surveys	197
Suspended Sediment Load	1091
sustainability	362
SVM	1166
Switching Algorithm	753
synchronization of chaotic systems	637
System Dynamics	864
system log	1460
Tax evaluation	677
Tax evasion	677
Techno-environmental performance	1720
Telemetric Data Analysis	1429
temporal trend	1022
Terrestrial water storage TWS	1601
Test Power	99
Test Size	99
text mining	1460
the Nelson-Siegel model	1507
the random forest analysis method	1125
the Russian economy	921

thermal indices	1111
thermography	182
tide gauge	1063
tides	1374
Time series	32, 158
Time Series	665, 895, 1253
time series	90, 192, 300, 1041, 1063, 1096, 1265
time series analysis	637
Time Series Analysis and Forecasting	795
time series classification	331
Time Series Classification	1431
Time Series Econometrics	1009
Time series forecasting	883
Time Series Forecasting	1431
time series modelling	717
time-delay embedding reconstruction	614
Time-Frequency Betas	274
Time-Frequency Rolling Regression	274
Time-series	1147
time-series	1017, 1022
Time-series analysis	1038
Time-Series data	1460
tipping point	1089
TOPSIS	559
total factor productivity	921
tourism	362
trajectories forecasting	864
Transfer function	130
Transportation sector	350
trend	1017, 1096
tropospheric ozone	1022
turning points	1576
Uncertainty	197
Unconstrained	1135
Unit commitment problem	602
unit roots	306
Unobserved Components Models	994
unsupervised learning	982
Value-at-Risk	99
Variational Bayes	1531
variational mode decomposition	1670
Vector Autoregression	598
Vegetation indices	1013
Ventilation system	1724
virtual leave one out	432
volatility	1670

volatility forecasting	1342
warning system	1111
waterlogging impact factors	1026
wave conditions	1328
Wavelength	1612
Wavelet	1091
wavelet coherence	1158
Wavelet Decomposition	1028
Wavelet Identification of Nonstationary Objects Virtual Predictive Models	783
Wavelet Transform	1073
Wavelets	274, 408
weather prediction	250
Whittle likelihood	469
wind and wave data	1385
Wind speed	385
Wind time series	92
WMA	529
Yield Curve Modeling	798
Zinc air battery	625

Author Index

Abukhurma, Ruba	1495
Acuña, José	1417
Aerts, Jean-Marie	130
Aguiar, Rui	571
Ahmadi-Abhari, Sara	318
Ahsan, Md Nazmul	1342
Aihara, Kazuyuki	550
Aknin, Noura	1738
Al-Dahidi, Sameer	385
Alawadhi, Shafiqah	1110
Albano, Giuseppina	583
Alijani, Farshad	1073
Aljawazneh, Huthaifa	1495
Alomar, Miquel L.	883
Alonso, Julio	1036
Alosaimi, Sarah	350
Alsawalqah, Hamad	1495
Alswiti, Wedyan	1495
Alvarez Bel, Carlos	1693
Alvarez, Ignacio	331
Alvarez-Meza, Andres Marino	1316
Ambrosino, Fabrizio	1038
Amigo, Jose Maria	550
An, Jongbaek	1720
Ando, Koki	1253
Andresen, Gorm B.	92
Antonio, Nelson	876
Antunes, Mário	571
Aoulad Abdelouarit, Karim	1738
Aranda Cotta, Higor Henrique	42, 1011
Araujo, Aracy	95
Arsenal, Roberto	1340
Avendaño, Daly	331
Aznar, Antonio	933
Babu, Krishna	1703
Badulescu, Yvonne	559
Bagnall, Anthony	331
Bahamou, Achraf	1178
Bakhtadze, Natalia	783, 1429
Balashova, Svetlana	921
Bandosz, Piotr	318
Baquerizo, Asunción	260, 598
Barraj, Nouha	1108

Barrera, Antonio	325
Bauer, Dietmar	288
Bebes, András	798
Bebesi, László	798
Bel Hassen, Malika	1108
Beldiceanu, Nicolas	982
Bengherifa, Samir	1091
Benhmad, Francois	1028
Benhmad, François	332
Benqatla, Mohammed Salim	1650
Bermad, Abd El Malek	1091
Birylo, Monika	1041, 1063
Bondarenko, Kseniia	227
Bondon, Pascal	42, 1011
Borowik, Grzegorz	1166, 1554
Boshnakov, Georgi	1519
Bounabat, Bouchaib	1650
Bratčikovienė, Nomedā	1565
Briestenský, Miloš	1038
Brunner, Eric	318
Brust, Peter	1612
Buekers, Joren	130
Bueno López, Maximiliano	519
Burchard, Bernd	952
Caicedo Acosta, Julian Camilo	1291
Canals, Vicente	883
Cando, Paul	1693
Cao, Hong-Xin	1013, 1026
Capewell, Simon	318
Cardenas Peña, David	499
Cardenas, David	1344
Cardenas-Peña, David	1291
Carlini, Federico	753
Carretero-Peña, Selena	1017
Castellanos Dominguez, Cesar German	1308
Castellanos Dominguez, German	489, 499, 509
Castellanos, German	1316, 1344
Castellanos, Germán	1728
Castellanos-Dominguez, German	1281, 1291
Castillo Valdivieso, Pedro	1495
Cerrato Alvarez, María	1022
Chapman, Jamie-Leigh	300
Cheikhrouhou, Naoufel	559
Chen, Wei-Tao	1013
Cheung, Ying Lun	1
Chikhaoui, Dikra	1650
Chojnacki, Damian	741

Christakos, George	260
Chvosteková, Martina	614
Cichosz, Pawel	1166, 1554
Cichosz, Pawel	432, 864
Claveria, Oscar	1543
Cobos, Manuel	260, 598
Cohen, Edward	1147
Correa, Debora	90
Cosovic, Marijana	264
Costa Silva, Guilherme	420, 602
Costa, Maria Da Conceição	1355
Cryns, Hanne	130
Cucina, Domenico	396
Cárdenas Peña, David	489
Cárdenas-Peña, David	1281
Datcu, Octaviana	1640
De Boever, Patrick	130
De Gaetano, Davide	583
de Lucas Santos, Sonia	677
Delgado Rodríguez, Maria Jesús	677
Díaz, Javier G.	1036
Dimoulkas, Ilias	182
Donnat, Philippe	1178
Doumergue, Maud	1178
Duempelmann, Matthias	327
Dufour, Jean-Marie	1342
Díez Minguito, Manuel	1374
Eckley, Idris	300
Elmoaqet, Hisham	385
Elpashev, Denis	1429
Ergun, Salih	637
Faria, Alvaro	966
Faris, Hossam	1495
Farooq, Umer	1703
Fearnhead, Paul	1469
Feki, Wafa	1108
Fernández, Alicia	1417
Finazzi, Francesco	1471
Fontes, Xavier	1431
Frasser, Christiam F	883
Fuentes Castillo, Martha Elena	306
Furtado, Pedro	586
Gabdrakhmanova, Nailia	1367
Galan Prado, Fabio	883

García-García, José Carlos	373
García-Murillo, Daniel Guillermo	1281
García-Ródenas, Ricardo	373
García González, Gastón	1417
Gebert, Ole	625
Gerhart, Christoph	1483
German-Soto, Vicente	306
Gibberd, Alexander	1147
Gil-Swidarska, Agnieszka	1158
Giovannelli, Alessandro	1468
Giraldo, Eduardo	509, 519
Gloesekoetter, Peter	625
Gomes, Diogo	571
González Vasco, María Del Camino	677
Gorriz, Juan Manuel	331
Greiner, Martin	92
Guzman-Castillo, Maria	318
Gómez Sena, Gabriel	1417
Gómez Villegas, Miguel Ángel	1340
Górriz Sáez, Juan Manuel	1270
Gölles, Markus	170
Haber, Rana	408
Hahmann, Martin	1589
Halliday, Jamie	1519
Hamza, Asma	1108
Handayani, Susanti	444
Hanzlíková, Hana	1111
Hasan, Md. Irfanul	1680
Hasanov, Fakhri	795
Hassler, Uwe	1
Hazlett, Randy	1703
Heidari, Hadi	1300
Heming, Daniel	625
Heo, Jun-Haeng	1015
Herzallah, Randa	134
Hirata, Yoshito	550
Hobincu, Radu	1640
Hondula, David M.	1111
Hong, Taehoon	1720, 1724
Horsthemke, Ludwig	625
Huls, Albert Jan	158
Hungnes, Håvard	811
Hurtado, Juana Valeria	1728
Iqbal, Md. Shahid	1680
Ircha, Dobieslaw	1166, 1554
Irfan, Andri	444

Isern, Eugeni	883
Jakubík, Jozef	1265
Jankowski, Stanislaw	432
Javaherian, Mohsen	1010
Jensen, Kjeld	1469
Jerónimo, Carlos	876, 1451
Jothimani, Dhanya	1660
Joutz, Frederic	795
Jowaheer, Vandna	1135
Jung, Woojin	1724
Junuz, Emina	264
Kagraoka, Yusho	1507
Keshavarz Haddad, Gholamreza	1300
Khoroshiltseva, Marina	1468
Killian, Michaela	250
Killick, Rebecca	300
Kim, Hakpyeong	1724
Kim, Hyun-Jun	1125
Kim, Hyun-Soo	1125
Kim, Taereem	1015
Kivimäki, Mika	318
Kobialka, Hans-Ulrich	729
Koskinen, Hannu	907
Kowalczyk, Kamil	1063
Kozek, Martin	250
Kozia, Christina	134
Krakovská, Anna	1113
Kranz, Mathias	1612
Krstanovic, Sascha	895
Kuczynska-Siehien, Joanna	1063
Kunst, Robert	32
Kurisummoottil Thomas, Christo	1531
Kysely, Jan	1111
Lakkakula, Prithviraj	1576
Lasak, Katarzyna	753
Latruwe, Timo	302
Lazanyuk, Inna	921
Lee, Eun-Seong	1125
Lee, Minhyun	1720
Lee, Sang-Hyun	1125
Lefkir, Abd El Wahab	1091
Lehner, Wolfgang	1589
Lehnert, Thorsten	1009
Levy-Leduc, Celine	42
Lira-Loarca, Andrea	260, 598

Lisboa, Adriano	420, 602
Lita, Merry	444
Liu, Yuchen	119
Livina, Valerie	1089
Loechte, Andre	625
Lops, Ben	158
Losada, Miguel A.	598
Lototsky, Alexey	783, 1429
Lototsky, Vladimir	783, 1429
Lowe, David	134, 717
Lowther, Aaron	1469
Lu, Shuixiu	457
Luna Naranjo, David Felipe	499
Luo, Zongwei	457
Luzzu, Mauro	170
Lévy-Leduc, Céline	1011
López Montes, Juan Camilo	489
López-García, David	1270
Lütkebohmert, Eva	1483
Macedo, Pedro	1355
Machado, Sérgio	95
Madi Wamba, Gilles	982
Mahfoudi, Mabrouka	1108
Mahjoubi, Fatima Zahra	1574
Maldonado, Juan	1693
Malecka, Marta	99
Mamode Khan, Naushad	1135
Mandal, Anandadeep	146
Manzouri, Farrokh	327
Marino, Maria Adele	331
Martinez Vargas, Juan David	509
Martinez, Andrew	920
Martinez, Juan David	1728
Martins, Luis	1193
Martinez Tagliafico, Sergio	1417
Mattila, Petteri	76
Matyushok, Vladimir	921
Meisel, Christian	327
Mestre, Roman	274, 332
Mestre-Zhou, Yang	332
Meyer-Baese, Anke	331
Michalak, Pawel	1166, 1554
Mijatovic, Nenad	408
Mikayilov, Jeyhun	795
Mirarabi, Ali	1073
Miró Rodríguez, Conrado	1017, 1022
Miśkiewicz, Janusz	192

Moctezuma, Luis Alfredo	64
Molinas, Marta	64, 519
Mollah, Sabur	825
Monte-Moreno, Enrique	1543
Montillet, Jean-Philippe	654
Mora García, Antonio	1495
Morales, Diego	1693
Moreira Barreto de Oliveira, Abdinardo	146
Morán, Alejandro	883
Moussa, Zakaria	1507
Moyou, Mark	408
Muniaín, Peru	852
Muschick, Daniel	170
Muslun, Azmi	1135
Muñoz Gutiérrez, Pablo Andrés	509, 519
Méndez, Pablo	1693
Nakhaei, Mohammad	1073
Nassery, Hamid Reza	1073
Nethery, Rachel	469
Neupane, Bijay	1589
Niedzwiecki, Maciej	741
Nigitz, Thomas	170
Njah, Hasna	1108
Njemanze, Philip	1612
Noble, Jordan	1147
Nunes, Matt	1469
O'Flaherty, Martin	318
Oberst, Sebastian	457
Oh, Jeongyoon	1720, 1724
Ollech, Daniel	20
Olszewski, Wojciech	1166, 1554
Oussama, Abdel Khalek	1574
Paci, Lucia	1471
Padilla Buritica, Jorge Ivan	1308
Padilla, Jorge Iván	1728
Pajak, Katarzyna	1063
Paláncz, Béla	119
Pandey, Rajeev	1670
Paramaguru, Kanya	239
Pardo-Igúzquiza, Eulogio	645, 1406
Park, Hee-Jung	1125
Park, Se-Ik	1125
Paulheim, Heiko	895
Paurat, Daniel	729
Pedregal, Diego J.	994

Pereira, Isabel	997
Pereira, Leandro	876, 1451
Perkowski, Emilian	1166
Peter, Adrian M.	408
Pinho, André	586
Pinilla-Gil, Eduardo	1017, 1022
Pinker, Katja	331
Polukoshko, Svetlana	1096
Pulgarin-Giraldo, Juan Diego	1316
Pumo, Besnik	665
Pupavac, Drago	362
Pupavac, Justin	362
Pyatetsky, Valery	783
Pytlak, Radoslaw	864, 1166, 1554
Pytlak, Radosław	432
Pönkä, Harri	1051
Qarout, Yazan	717
Quoreshi, A.M.M. Shahiduzzaman	825
R. Silva, João Luis	420
Raiyn, Jamal	529
Ramirez, Javier	331
Ramírez Pérez de Inestrosa, Javier	1270
Rebai, Ahmed	1108
Reisen, Valdério	42, 1011
Restrepo, Francia	1728
Reyes Merlo, Miguel Ángel	1374
Ribeiro, Joana	571
Rissanen, Antti	76
Rizzo, Manuel	396
Roca, Miquel	883
Rodrigues, Paulo	1193
Rodríguez-Tovar, F. J.	645
Rodríguez-Tovar, Francisco Javier	1406
Romero, Francisco P.	373
Román-Román, Patricia	325
Rossello, Josep L	883
Ruz, María	1270
Rzepecka, Zofia	1601
Saastamoinen, Kalle	76
Sabbarese, Carlo	1038
Safari, Hossein	1010
Safi, Salah Al-Deen	1495
Saldanha, Rodney	420, 602
San Cristobal, Jose Ramon	111
Sanchez-Morales, J.	645

Santagostini, Pierre	665
Santamaría, Ignacio	1316
Santos, Alexandre	966
Santos, Cláudia	997
Santos, José	876, 1451
Sbihi, Boubker	1738
Schrader, Lisa	729
Schulze-Bonhage, Andreas	327
Scotto, Manuel	997
Selmani Bouayoune, Karima	1443
Serna, Andres C.	1036
Sheen, Jeffrey	197
Shin, Ju-Young	1015
Shiple, Martin	318
Siles Ajamil, Maria De Los Reyes	1374
Silva, Claudionor	95
Silva, Daniel	1431
Silva, Helena	586
Singh Patel, Shivshanker	1460
Skibinsky Gitlin, Erik Sebastian	883
Slock, Dirk	1531
Small, Michael	90
Smith, Anthony O.	408
Soeb, Md. Janibul Alam	1680
Spruit, Martijn A.	130
Sreekanth, K. J.	350
Stefanakos, Christos	1385
Sunecher, Yuvraj	1135
Suski, Damian	864
Szzechla, Eliza	432, 1166, 1554
Szymański, Zbigniew	432
Sánchez-Morales, José	1406
Tarnawski, Tomasz	864
Tawegoum, Rousseau	665
Terouzi, Wafa	1574
Terraza, Michel	274
Theunis, Jan	130
Thinová, Lenka	1038
Toledo, Marco	1693
Torra Porrás, Salvador	1543
Torres-Ruiz, Francisco	325
Tran, Dávid	798
Trapero, Juan R.	994
Truong, Young	469
Têtu, Amélie	1328
Tómasson, Helgi	54

Ulbricht, Robert	1589
Unterberger, Viktor	170
Urban, Aleš	1111
Ursu, Eugen	396
Van der Wee, Marlies	302
Van Vaerenbergh, Steven	1316
Vanleenhove, Pieter	302
Vansteenkiste, Henk	302
Veerasawmy, Isven	1135
Velasquez, Luisa	1344
Velasquez-Martinez, Luisa Fernanda	1291
Verbrugge, Sofie	302
Vermeersch, Sebastiaan	302
Victoria, Marta	92
Vieira, Douglas	420, 602
Villegas, Diego	994
Villegas, Marco A.	994
Vilmunen, Jouko	907
Walker, David	90
Wang, Ben	197
Wawrzyniak, Zbigniew	432, 864, 1166, 1554
Webel, Karsten	20
Weber, Marc	1483
Westerlund, Per	182
Whittaker, Hannah	318
Wouters, Emiel F.M.	130
Xia, Jianhong	119
Yadav, Surendra S.	1660
Yamauchi, Koichiro	1253
Yang, Tai-Ming	1026
Yu, Kegen	654
Zakharov, Eddy	1429
Zapata, Frank	1344
Zauner, Michael	250
Zawadzki, Tomasz	864
Zeiler, Wim	158
Zhang, Bao-Jun	1013, 1026
Zhang, Guoqiang	457
Zheng, Yi	1051
Ziel, Florian	852