

Olga Valenzuela · Fernando Rojas ·
Luis Javier Herrera · Héctor Pomares ·
Ignacio Rojas *Editors*

Theory and Applications of Time Series Analysis and Forecasting

Selected Contributions from ITISE 2021

Contributions to Statistics

The series **Contributions to Statistics** contains publications in theoretical and applied statistics, including for example applications in medical statistics, biometrics, econometrics and computational statistics. These publications are primarily monographs and multiple author works containing new research results, but conference and congress reports are also considered.

Apart from the contribution to scientific progress presented, it is a notable characteristic of the series that publishing time is very short, permitting authors and editors to present their results without delay.

Olga Valenzuela • Fernando Rojas •
Luis Javier Herrera • Héctor Pomares •
Ignacio Rojas
Editors

Theory and Applications of Time Series Analysis and Forecasting

Selected Contributions from ITISE 2021



Springer

Editors

Olga Valenzuela 
Faculty of Sciences
University of Granada
Granada, Spain

Luis Javier Herrera 
ETSIIT, CITIC-UGR
University of Granada
Granada, Spain

Ignacio Rojas 
ETSIIT, CITIC-UGR
University of Granada
Granada, Spain

Fernando Rojas 
ETSIIT, CITIC-UGR
University of Granada
Granada, Spain

Héctor Pomares 
ETSIIT, CITIC-UGR
University of Granada
Granada, Spain

ISSN 1431-1968

Contributions to Statistics

ISBN 978-3-031-14196-6

ISBN 978-3-031-14197-3 (eBook)

<https://doi.org/10.1007/978-3-031-14197-3>

Mathematics Subject Classification: 62M10, 91B84, 68T09

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book gathers the extended versions of a selection of the best contributions presented in the seventh edition of the International Conference on Time Series and Forecasting, ITISE 2021, held in Gran Canaria (Spain) in July 2021. Unfortunately, the 2020 edition had to be canceled due to the recent (at the time) outbreak of the COVID-19 coronavirus. After the decision of this cancellation, we all took it for granted that the next edition, that of 2021, was going to be exempt from any problems derived from the pandemic and that it was going to be a congress that was going to be held in a totally normal way, with all face-to-face sessions. However, a year later, the situation was still quite serious, although the experience gained during the previous months on how to act in these very special situations allowed the congress to be held in a hybrid way, that is, with a mixture of online sessions and face-to-face sessions where both speakers and listeners, regardless of whether they were physically at the conference venue, could participate in reasonably good conditions. The organization of the congress was very complex, but it could be carried out successfully. This would not have been possible without the excellent predisposition of all the participants, and for this, we want to heartily thank them.

On the other hand, the pandemic itself turned out to be a very important source of information that is being exploited by many researchers to achieve the capability to analyze, for example, the way in which infections spread and how these pandemics influence the economy of each country or how the characteristics of a country influence the evolution of infections, and, from there, the ability to predict this evolution in the future. Although some contributions in this direction have already appeared (some even in this book), it is clear that the pandemic is not yet over and that there is still much work to be done. This congress, in some way, contributes its grain of sand to being able to answer the questions just raised above as the main objective of this conference is none other than to provide a friendly discussion forum for scientists, engineers, educators, and students to debate about the latest ideas and realizations in the foundations, theory, models, and applications in the field of time series analysis and forecasting. More specifically, the main topics in the last edition of the Conference were:

1. Time series analysis and forecasting

- Nonparametric and functional methods.
- Vector processes.
- Probabilistic approaches 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.
- Data preprocessing methods: data decomposition, seasonal adjustment, singular
- spectrum analysis, detrending methods, etc.

2. Econometrics and forecasting

- Econometric models
- Economic and econometric forecasting
- Real macroeconomic monitoring and forecasting
- Advanced econometric methods

3. Advanced methods and online learning in time series

- Adaptivity for stochastic models
- Online 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

4. High dimension and complex/big data

- Local vs global forecasts
- Dimension reduction techniques
- Multiscaling
- Forecasting Complex/Big data

5. Forecasting in real problems

- Health forecasting
- Atmospheric science forecasting
- Telecommunication forecasting
- Hydrological forecasting
- Traffic forecasting
- Tourism forecasting
- Marketing forecasting

- Modelling and forecasting in power markets
- Energy forecasting
- Climate forecasting
- Financial forecasting and risk analysis
- Forecasting electricity load and prices
- Forecasting and planning systems

High-quality candidate papers from the Conference ITISE 2021 (24 contributions) were invited to submit an extended version of their conference paper to be considered for this special publication in the Springer book series Contributions to Statistics. For the selection procedure, the information/evaluation of the chair of every session, in conjunction with the review comments and the summary of reviews, was taken into account.

So, now we are pleased to have reached the end of the whole process and present the readers with these final contributions that we hope will provide a clear overview of the thematic areas covered by the ITISE 2021 conference ranging from theoretical aspects to real-world applications.

For the sake of readability, the contributions presented in this book have been classified into different chapters according to their content. Some chapters of the book contain pure theoretical contributions. On the other hand, there are chapters with more practical contributions with the intention of providing the readers with a more real-world view of the field. As is common in these editions, a specific chapter of the book has been dedicated to econometrics, one of the most prominent applications of time series modelling & forecasting. In the following, we will make a short summary of what the reader may find in every chapter of the book:

- **Part 1. “Theoretical Aspects of Time Series”** Although in the field of time series it is difficult to separate the theoretical aspects from the practical ones, since the presentation of many of the theoretical developments usually ends with practical examples where these developments could be applied, the papers in this first chapter have been selected for being mainly theoretical. It is the largest chapter in this book, which is reasonable since theoretical contributions are the seed of numerous practical advances that can be derived from them. The chapter begins with a very original paper to forecast and detect structural breaks in time series using fuzzy natural logic. In the second contribution, the authors investigate approaches to automatically detect and replace anomalies in time series to enable accurate forecasts with special emphasis on energy consumption time series. In the third contribution the size and power of a large set of unit root tests on time series from the M4 competition data are evaluated and then a conditional random forest-based elimination algorithm is used to assess which tests should be combined to improve the performance of each individual test. The next contribution is dedicated to how to improve probabilistic forecasting accuracy in seasonal time series. To that end, the authors propose a framework in which a combination of several machine learning techniques is used to identify typical seasons and to forecast a probability distribution of the next season. Next, the fifth paper deals with a comparison between statistical and non-statistical

methods for the analysis, forecasting, and mining of time series. As the authors state, “our goal is not to beat statistical methods, but vice versa – to benefit from the synergy of both.” The last two contributions of this first chapter can also be of high interest for many readers: the first one deals with forecasting of time series whose samples are non-negative integer values (count processes), and the second one, and last contribution of this chapter, deals with discrete-time series observed at irregularly spaced times.

- **Part 2. “Econometric and Forecasting”.** This chapter aims at presenting some recent developments of time series research applied to forecasting methods in econometrics. Five contributions have been selected. The first one develops an economic policy uncertainty index for the USA and Canada using natural language processing methods which are capable of successfully capturing COVID-19-related uncertainty. In the second contribution, the authors propose two semi-nonparametric distributions to estimate the value-at-risk and the expected shortfall in four indices related to energy, metals, mining, and physical commodities. The third contribution deals with forecasting the long-term trend of housing prices in the Spanish cities with more than 25 thousand inhabitants, a total of 275 individual municipalities. According to the author, the results obtained give a comprehensible evolution of the long-term component of housing prices, and the model also provides a way to understand the main drivers of housing prices in each Spanish region. The next paper combines different models to obtain a unique prediction model for Bitcoin dollars time series. Finally, in the last contribution of this chapter, the authors estimate the historical cost of the Hungarian retail debt program, taking portfolio effects and risks into account so that they can afterwards simulate the future effects of retail debt based on security-level transaction data and a vector error correction macroeconomic model. Finally, the last contribution of this chapter makes a deep analysis of how important is to utilize information about exchange rate movements closer to the publication date to improve the prediction of the Exchange Rate Path.
- **Part 3. “Time Series Prediction Applications”.** The third chapter of the book is dedicated to real applications of time series forecasting different from the ones related to econometrics. The idea is to state explicitly that applications of time series prediction reach practically any scientific discipline imaginable. Four contributions were selected for this chapter. The first studies how to combine the forecasting capability of very different methods such as Neural Network Auto Regression, Box-Cox Transformation, ARMA residuals Trend and Seasonality, Trigonometric Box-Cox Transformation, Holt's Linear Trend, autoregressive integrated moving average, and cubic smoothing splines with the aim to improve the forecast of infection cases of COVID-19. The next two contributions deal with daily electricity demand. The first one, targeting Uruguay, makes use of Markov switching models, and the second one is based on calendar features and temporal convolutional networks for some regions in Germany. This chapter ends with a paper focused on network security situation awareness forecasting. The authors try first to estimate the influence of the loss function on network security situation awareness forecasting and then compare the performance of

both statistical and neural networks-based methods in network security situation awareness forecasting.

- **Part 4. “Advanced Applications in Time Series Analysis”.** The last chapter of the book is dedicated to specific applications of time series analysis. The first contribution of this chapter analyzes daily COVID-19 contagion time series of different countries using Markov switching models with ARMA structure. The second contribution investigates the temporal trends of the diffusion process of renewable energies, namely wind and solar, in leading countries for their consumption. The third contribution tries to find answers to the question of how to turn negative employment trends in the Croatian water transport system into positive ones. To that end, the authors make use of descriptive statistics and correlation and regression analysis to compare the state of employment and employment trends in the water transport system of the European Union and the Republic of Croatia. Finally, this chapter, and therefore this book, concludes with a contribution that analyzes the relationship between economic growth, demographic development, and CO₂ emissions for 30 industrialized countries from time-series data. According to the authors, the results confirm that GDP per capita growth rates of highly industrialized economies are significantly driven by the development of CO₂ emissions, population, and energy intensity.

Last but not least, we would like to point out that this edition of ITISE was organized by the University of Granada (UGR), Spain, together with the Spanish Chapter of the IEEE Computational Intelligence Society. The guest editors would also like to express their gratitude to all the people who supported them in the compilation of the book, and especially to the contributing authors for their submissions, the chairs of the different sessions, and to the anonymous reviewers for their comments and useful suggestions in order to improve the quality of the papers.

We wish to thank our main sponsors as well: the Department of Computer Architecture and Technology at the UGR, the Faculty of Science at the UGR, the Research Center for Information and Communications Technologies (CITIC-UGR), the Spanish Network on Time Series (RESeT), and the Ministry of Science and Innovation for their support and grants. Finally, we wish also to thank Prof. Alfred Hofmann, Vice President Publishing – Computer Science, Springer-Verlag, and Dr. Veronika Rosteck, Springer Editor, for their interest in editing a book series of Springer based on the best papers of ITISE 2021.

We hope the readers of this book find these contributions interesting and helpful.
Granada, Spain, January 2021

Granada, Spain

Olga Valenzuela
Héctor Pomares
Luis Javier Herrera
Fernando Rojas
Ignacio Rojas

Contents

Part I Theoretical Aspects of Time Series

An Improved Forecasting and Detection of Structural Breaks in Time Series Using Fuzzy Techniques	3
Thi Thanh Phuong Truong and Vilém Novák	
Anomaly Detection Algorithm Using a Hybrid Modelling Approach for Energy Consumption Time Series	19
Florian Rippstein, Steve Lenk, Andre Kummerow, Lucas Richter, Stefan Klaiber, and Peter Bretschneider	
Unit Root Test Combination via Random Forests	31
Luca Nocciola, Daniel Ollech, and Karsten Webel	
Probabilistic Forecasting of Seasonal Time Series	47
Colin Leverger, Thomas Guyet, Simon Malinowski, Vincent Lemaire, Alexis Bondu, Laurence Rozé, Alexandre Termier, and Régis Marguerie	
Nonstatistical Methods for Analysis, Forecasting, and Mining Time Series	65
Vilém Novák and Irina Perfilieva	
PMF Forecasting for Count Processes: A Comprehensive Performance Analysis	79
Annika Homburg, Christian H. Weiß, Layth C. Alwan, Gabriel Frahm, and Rainer Göb	
A Novel First-Order Autoregressive Moving Average Model to Analyze Discrete-Time Series Irregularly Observed	91
César Ojeda, Wilfredo Palma, Susana Eyheramendy, and Felipe Elorrieta	

Part II Econometric and Forecasting

Using Natural Language Processing to Measure COVID-19-Induced Economic Policy Uncertainty for Canada and the USA	107
Shafiullah Qureshi, Ba Chu, Fanny S. Demers, and Michel Demers	
Asymptotic Expansions for Market Risk Assessment: Evidence in Energy and Commodity Indices	123
Daniel Velásquez-Gaviria, Andrés Mora-Valencia, and Javier Perote	
Predicting Housing Prices for Spanish Regions	143
Paloma Taltavull de La Paz	
Optimal Combination Forecast for Bitcoin Dollars Time Series.....	161
Marwan Abdul Hameed Ashour and Iman A. H. Aldahhan	
The Impact of the Hungarian Retail Debt Program	175
Bianka Biró, Dávid Tran, András Stark, and András Bebes	
Predicting the Exchange Rate Path: The Importance of Using Up-to-Date Observations in the Forecasts	195
Håvard Hungnes	

Part III Time Series Prediction Applications

Development of Algorithm for Forecasting System Software	213
Mostafa Abotaleb and Tatiana Makarovskikh	
Forecasting High-Frequency Electricity Demand in Uruguay	227
Bibiana Lanzilotta and Silvia Rodríguez-Collazo	
Day-Ahead Electricity Load Prediction Based on Calendar Features and Temporal Convolutional Networks.....	243
Lucas Richter, Fabian Bauer, Stefan Klaiber, and Peter Bretschneider	
Network Security Situation Awareness Forecasting Based on Neural Networks	255
Richard Staňa, Patrik Pekarčík, Andrej Gajdoš, and Pavol Sokol	

Part IV Advanced Applications in Time Series Analysis

Modeling Covid-19 Contagion Dynamics: Time-Series Analysis Across Different Countries and Subperiods.....	273
Zorica Mladenović, Lenka Glavaš, and Pavle Mladenović	
Diffusion of Renewable Energy for Electricity: An Analysis for Leading Countries	291
Alessandro Bessi, Mariangela Guidolin, and Piero Manfredi	

The State and Perspectives of Employment in the Water Transport System of the Republic of Croatia	307
Drago Pupavac, Ljudevit Krpan, and Robert Maršanić	
Reversed STIRPAT Modeling: The Role of CO₂ Emissions, Population, and Technology for a Growing Affluence	321
Johannes Lohwasser, Axel Schaffer, and Tom Brökel	

Part I

Theoretical Aspects of Time Series

An Improved Forecasting and Detection of Structural Breaks in Time Series Using Fuzzy Techniques



Thi Thanh Phuong Truong and Vilém Novák

Abstract In this paper, we address nonstatistical methods for forecasting and detection of structural breaks in time series. Our methods are based on the application of the unique fuzzy modeling method called *fuzzy transform* (F-transform) and selected methods of *fuzzy natural logic* (FNL). The latter provides a formal model of the semantics of a part of natural language and methods for reasoning based on it. Using F-transform, we first estimate the trend-cycle. Then, using methods of FNL, we extract a sort of expert information that enables us to forecast the trend-cycle. Since F-transform also makes it possible to estimate the slope of time series over an imprecisely specified area (ignoring its volatility), we identify structural breaks through evaluation of changes in the slope by a suitable evaluative linguistic expression. We will demonstrate the effectiveness of our methods on several real time series and compare our results of forecasting with the classical ARIMA statistical method. Our methods are computationally very effective.

Keywords Time series · ARIMA · Fuzzy transform · Evaluative linguistic expressions · Fuzzy natural logic

The work was supported from ERDF/ESF by the project “Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region” No. CZ.02.1.01/0.0/0.0/17-049/0008414.

T. T. P. Truong (✉) · V. Novák

Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, Ostrava, Czech Republic

e-mail: phuong.truong@osu.cz; vilem.novak@osu.cz

1 Introduction

In this paper, we discuss a nonstatistical approach to forecasting time series and detection of structural breaks in them. We apply special techniques of fuzzy modeling, namely, the F-transform and methods based on the theory of fuzzy natural logic (FNL).

We stem from the assumption that we can decompose the time series additively into trend-cycle, seasonal component, and random noise. The fuzzy transform makes it possible to find the arbitrary shape of the trend-cycle and detect specific areas or intervals of monotonous behavior. We can also estimate the slope (average value of the first derivative) of the time series in a given though imprecisely delineated area and evaluate it using methods based on the FNL theory. A brief overview of our other methods is provided in the paper [5] (this volume). More details can be found in the citations therein and in the book [7].

The mentioned methods are applied to the data collected from the Internet using an experimental software LFL Forecaster.¹

The paper is structured as follows. In the next section, we introduce the decomposition of time series and provide a brief overview of the main concepts of the fuzzy transform and fuzzy natural logic. In Sect. 3, we explain our forecasting method. In Sect. 4, we describe our approach to location and identification of structural breaks. Section 5 is devoted to demonstrating the forecasting on real-time series and comparison with the ARIMA method. The second part of this section is devoted to demonstrating our method for detecting structural breaks.

2 Processing Time Series Using Fuzzy Modeling Methods

2.1 Time Series Decomposition

A time series X is a mapping

$$X : \mathbb{T} \times \Omega \rightarrow \mathbb{R},$$

where $\mathbb{T} = \{0, \dots, n\} \subset \mathbb{N}$ is a finite set of numbers interpreted as time moments and Ω is a nonempty set of elementary random events.

We assume that the time series is decomposed into four components:

$$X(\omega, t) = Tr(t) + C(t) + S(t) + R(\omega, t), \quad t \in \mathbb{T}, \quad (1)$$

¹ This is an experimental software developed in the Inst. for Research and Applications of Fuzzy Modeling of the University of Ostrava, Czech Republic, which implements the described method (see http://irafm.osu.cz/en/c110_lfl-forecaster/). Its author is Viktor Pavliska.

where $Tr(t)$ and $C(t)$ are trend and cyclic components of time series. These two components are usually joined into one component called *trend-cycle* $TC(t) = Tr(t) + C(t)$.

The $S(t)$ is the seasonal component and $R(\omega, t)$ is a random noise. The trend, cycle, and seasonal components are ordinary functions not having stochastic character. The noise $R(\omega, t)$ is assumed to be a sequence of independent random variables with the mean $\mu = 0$ and variance $\sigma^2 < +\infty$. Since in practice we have one realization for a fixed $\omega \in \Omega$ of the time series at disposal, we will omit ω from the arguments in (1).

2.2 Fuzzy Transform (F-Transform)

Recall that by a *fuzzy set* in the universe U , we understand a function $A : U \rightarrow [0, 1]$.² The set of all fuzzy sets on U is denoted by $\mathcal{F}(U) = \{A \mid A : U \rightarrow [0, 1]\}$.

The fuzzy transform is a technique for approximation of bounded continuous functions. In our case, it can be effectively applied to analysis and forecasting of time series. Let a bounded real continuous function $f : [a, b] \rightarrow [c, d]$ be given, where $a, b, c, d \in \mathbb{R}$. The fundamental concept is that of *fuzzy partition*.

Definition 1 Let $c_0 < \dots < c_n$ be fixed nodes in the interval $[a, b]$ where $c_0 = a$, $c_n = b$ with $n \geq 2$ and $a, b \in \mathbb{R}$. The set $\mathcal{A} = \{A_0, \dots, A_n\}$ of fuzzy sets is called a *fuzzy partition* of $[a, b]$ if the following conditions are fulfilled:

- $A_k : [a, b] \rightarrow [0, 1]$, $A_k(c_k) = 1$;
- $A_k(x) = 0$ if $x \notin (c_{k-1}, c_{k+1})$ (for $c_{-1} = a$ and $c_{n+1} = b$);
- A_k is continuous;
- A_k strictly increases on $[c_{k-1}, c_k]$ for $k = 1, \dots, n$ and A_k strictly decreases on $[c_k, c_{k+1}]$ for $k = 0, \dots, n-1$;
- $\sum_{k=0}^n A_k(x) = 1$ for all $x \in [a, b]$;
- Let $c_k = a + hk$, where $h = (b - a)/n$, $A_k(c_k - x) = A_k(c_k + x)$, for all $x \in [0, h]$ and $k = 1, \dots, n-1$;
- $A_k(x) = A_{k-1}(x - h)$, $A_{k+1}(x) = A_k(x - h)$ for $k = 1, \dots, n-1$ and $x \in [c_k, c_{k+1}]$.

The fuzzy sets A_k are also called *basic functions*. Note that Definition 1 specifies their properties but not their shape. Most usual are triangular A_k , but any shape fulfilling Definition 1 can be considered. Note that the width of basic functions is equal to $2h$.

The F-transform has two phases: direct and inverse.

² The interval $[0, 1]$ can be replaced by a proper bounded lattice.

Definition 2 Given a fuzzy partition by Definition 1 and $f : [a, b] \rightarrow [c, d]$ be a continuous function on $[a, b]$. The $(n + 1)$ -tuple $\mathbf{F}^m[f] = (F_0^m[f], \dots, F_n^m[f])$ is called m -th degree *direct fuzzy transform* of f if

$$F_k^m[f](x) = \beta_k^0[f] + \beta_k^1[f](x - c_k) + \dots + \beta_k^m[f](x - c_k)^2, \quad (2)$$

for all $k = 0, \dots, n$. We call $F_k^m[f]$ in (2) *components* of the fuzzy transform. Precise computation of the components (2) is in detail described in [8, 9] and elsewhere.

Definition 3 Given a fuzzy partition due to Definition 1 and $\mathbf{F}^m[f]$ be the direct F-transform of f due to Definition 2. Then the function $\hat{f} : [a, b] \rightarrow \mathbb{R}$ denoted by $\hat{f}(x) = \sum_{k=0}^n F_k^m[f]A_k(x)$ is called the *inverse fuzzy transform* of f .

The fuzzy transform is linear, has a universal approximation property, and has a linear computational complexity (cf. [2, 8, 9]).

2.3 Fuzzy Natural Logic

In our applications to time series, we will also use some methods of *fuzzy natural logic* (FNL). The latter is a set of special theories of mathematical fuzzy logic whose aim is to provide a mathematical model of common-sense human reasoning that is based on the use of natural language. Our methods for time series analysis apply two theories of FNL: the theory of evaluative linguistic expressions and fuzzy/linguistic IF-THEN rules.

Evaluative linguistic expressions are special expressions of natural language in the form

$$\langle \text{linguistic hedge} \rangle \langle \text{TE-adjective} \rangle$$

where $\langle \text{linguistic hedge} \rangle$ is a special adverb standing before $\langle \text{TE-adjective} \rangle$ that makes the adjective more or less specific. Examples of the former are “roughly, very, quite, significantly,” etc., and examples of the latter are canonical adjectives “small, medium big,” but also “shallow, medium deep, deep,” and many other ones.

To determine the semantics of evaluative expressions, we need the concept of *context* that, in our case, is the interval $w = [v_L, v_S] \cup [v_S, v_R]$ where $v_L, v_S, v_R \in \mathbb{R}$. The numbers have the following meaning: v_L is the left bound, v_S is a typical middle value, and v_R is the right bound.

The meaning of an evaluative expression \mathcal{A} is modeled by a function $W \rightarrow \mathcal{F}(\mathbb{R})$ where W is a set of all contexts. Such a function is called *intension* of \mathcal{A} .

If the context $w \in W$ and a value $x \in \mathbb{R}$ are given, then we can generate an evaluative expression Ev characterizing linguistically x w.r.t. w using a special function of *local perception*:

$$Ev = \text{LPerc}(x, w). \quad (3)$$

A special class of evaluative expressions are those characterizing trend:

$$\text{Trend is } \langle \text{direction} \rangle \quad (4)$$

where

- $\langle \text{direction} \rangle := \text{stagnating} \mid \langle \text{special hedge} \rangle \langle \text{sign} \rangle$
- $\langle \text{sign} \rangle := \text{increasing} \mid \text{decreasing}$
- $\langle \text{special hedge} \rangle := \emptyset \mid \text{negligibly} \mid \text{slightly} \mid \text{somewhat} \mid \text{clearly} \mid \text{roughly} \mid \text{sharply}$
 $\mid \text{quite largely} \mid \text{fairly large} \mid \text{hugely} \mid \text{significantly}.$

We must also consider the context w_{tg} for tangent (cf. Subsection 2.3) that is here extended to have two parts: positive w_{tg}^+ for increase of time series and negative w_{tg}^- for its decrease.

Fuzzy/linguistic IF-THEN rules are the main tool used in the forecasting. These rules are taken as conditional sentences of natural language having the form “IF X is \mathcal{A} THEN Y ” is \mathcal{B} ” where \mathcal{A} and \mathcal{B} are evaluative linguistic expressions. The relation between the antecedent “ X is A ” and the consequent “ Y is B ” is modeled using a fuzzy implication. A set of fuzzy/linguistic IF-THEN rules is called *linguistic description* since it describes in natural language the way how forecast was obtained. We have developed also a procedure for learning linguistic description from data (i.e., from the given time series).

A special reasoning method based on linguistic description is called *perception-based logical deduction* (PbLD) and is used in forecasting. For the details about the theory of evaluative linguistic expressions, fuzzy/linguistic IF-THEN rules, and PbLD, see [7].

3 Forecasting Time Series

By forecasting, we understand determining future values of the time series on the basis of the previous available values. To do it, we divide the time series into two subsets: *learning set* and *validation set*. The learning set is used for searching the best model of time series. The quality of the model is evaluated using special indexes computed on the basis of the validation set. The learning and validation sets form in-samples.

To test whether our forecast really works, we can cut off the last part of the time series and form a *testing set* (out-samples) that is not used in the computations but only to test the quality of our forecast. We use the standard quality indexes: root mean square error (RMSE) or symmetric mean absolute percentage error (SMAPE).

Forecasting Procedure

- From a time series X in (1), compute the F-transform components $F_1[f|A_h], \dots, F_n[f|A_h]$. With respect to the model (1), we can forecast the trend-cycle TC or trend T and the seasonal component (S). This depends on the choice of the distance h between nodes. As has been proved in [6], to remove all frequencies higher than a given one, we have to set $h = dT$ where T is the corresponding periodicity (found using periodogram) and d is a number that we usually set as $d \in \{1, 2\}$.
- The forecast is obtained by combination of the theory of F-transform, learning a linguistic description and PbLD inference. The variables in the latter are the F-transform components above, and their first and second differences:

$$\Delta F[f|A_h]_i = F[f|A_h]_i - F[f|A_h]_{i-1}, i = 2, \dots, n-1. \quad (5)$$

$$\Delta^2 F[f|A_h]_i = \Delta F[f|A_h]_i - \Delta F[f|A_h]_{i-1}, i = 3, \dots, n-1. \quad (6)$$

The learned linguistic description consists of the rules having the form

$$\text{IF } X_{i-1} \text{ is } \mathcal{A}_{i-1} \text{ AND } X_i \text{ is } \mathcal{A}_i \text{ THEN } X_{i+1} \text{ is } \mathcal{B}_{i+1} \quad (7)$$

where X_i stands either for the components $F[f|A_h]_i$, their first (5), or second differences (6).

The linguistic description gives us information both about dynamic behavior of the time series as well as logical dependencies inside the trend-cycle (or trend).

- The seasonal component is assumed to be stationary. The future vector S_{p+1} is computed as a linear combination of p previous vectors. We can also use other methods for its forecast, i.e. ARIMA or neural nets.

4 Detection of Structural Breaks in Time Series

Detection of structural breaks in time series means determining time moments when the course of the time series is abnormally changed. Our detection method is based on finding short intervals with a steep slope of trend (big tangent) and on the characterization of their steepness. The intervals are characterized by fuzzy sets, i.e., imprecisely. They are constructed from the last time moment of the time series to the first one. Each found interval is characterized by a specific evaluative expression (4). If the expression is of the kind “fairly large (hugely) decreasing/increasing,” this is the candidate for a structural break.

Let X be a time series and $\bar{T} \subseteq \mathbb{T}$ be a time interval, $\beta^1[X|\bar{T}]$ be the slope of trend of X over the period \bar{T} , and w_{tg}^-, w_{tg}^+ be the corresponding negative and positive parts of the context, respectively. Then, the evaluative expression $\pm Ev[X|\bar{T}]$ is obtained using the function of *local perception*: $\pm Ev[X|\bar{T}] :=$

$L\text{Perc}(\pm\beta^1[X|\bar{T}], w_{tg}^\pm)$. We thus decompose the time domain \mathbb{T} into a set of intervals

$$\mathcal{T} = \{\bar{T}_i \mid i = 1, \dots, s\}, \quad \bigcup \mathcal{T} = \mathbb{T}$$

Let $\bar{T}_i = \{t_{i_1}, \dots, t_{i_m}\}$ be processed being evaluated by the expression $\pm Ev[X|\bar{T}_i]$. The interval $\bar{T}_i \in \mathcal{T}$ is an area of a structural break in the course of X if the trend of X in the interval \bar{T}_i is hugely increasing (decreasing). The algorithm for finding structural breaks is in detail described in [3, 4].

5 Demonstration of Nonstatistical Forecast and Detection of Structural Breaks on Real Data

In this section, we will work with two datasets. The first one contains four monthly time series that are used in our experiments. The dataset is picked up from Time Series Data Library (TSDL) on the Internet.³ It will be used for forecasting time series by using both methods: ARIMA and LFL Forecaster tool.

The second dataset also contains four time series that will be used for detection of the structural breaks' task.

5.1 ARIMA Model

The ARIMA(p, d, q) (autoregressive integrated moving average)⁴ (see [1]) is one of the most successful forecasting models. The p is the order of AR (autoregressive) term, q is the order of MA (moving average) term, and d is the number of differencing needed for the time series to be stationary. To fit the ARIMA model in this paper, we applied the function “auto.arima()” in R program. By combination of order parameters, “auto.arima()” can choose the triple (p, d, q) that optimizes and fits the model.

Figure 2 shows results of forecasting horizons that are represented by testing data (out-sample date) of the four time series depicted in Fig. 1. The parameters of the ARIMA model were as follows: (a) ARIMA(4,1,4), (b) ARIMA(0,1,2), (c) ARIMA(5,1,1), (d) ARIMA(0,1,2). In Table 1 are the corresponding RMSE errors.⁵

³ <https://robjhyndman.com/tsdl/>.

⁴ Also Box-Jenkins model.

⁵ $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}$, where y_i are the predicted values, x_i are the actual values, and n is the number of observations.

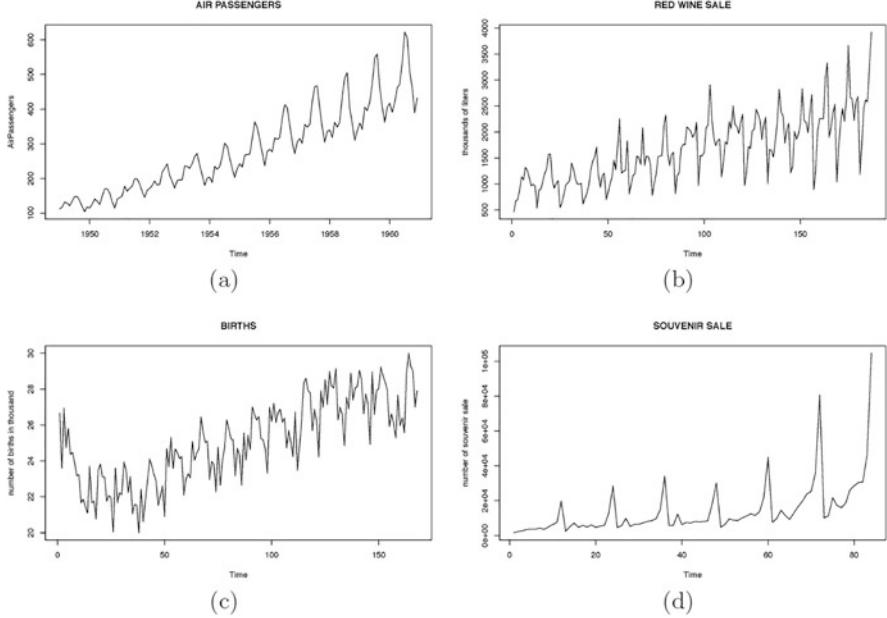


Fig. 1 Panel (a) describes the data that contain 144 observations, which is the number of monthly international airline passengers (in thousands) from January 1949 to December 1960. Panel (b) describes data of red wine that contains the monthly sale of red wine (in thousands of liters) in Australia from January 1980 to December 1995. Panel (c) describes data of the number of births per month in New York City, from January 1946 to December 1959 (originally collected by Newton). Panel (d) describes data that contain monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, for January 1987–December 1993 (original data from Wheelwright and Hyndman (1998))

5.2 Forecasting Using LFL Forecaster

In this section, we present the results obtained using the experimental software LFL Forecaster developed in the IRAFM of the University of Ostrava. Using it, we divide the time series into in-samples and out-samples (testing data). The in-samples are then divided into the learning and testing part, where the latter is used for testing the best model.

In our case, the learning set consists of 120 observations, 12 validation sets, and 12 testing sets. The quality of forecasting process is measured both by RMSE and by SMAPE errors.

Unlike ARIMA, LFL Forecaster provides, besides the forecast, also linguistic description of how the forecast was obtained. The description consists of fuzzy/linguistic IF-THEN rules of the form (7). The obtained description of the forecasts of the mentioned four time series are in Figs. 3, 4, 5, and 6.

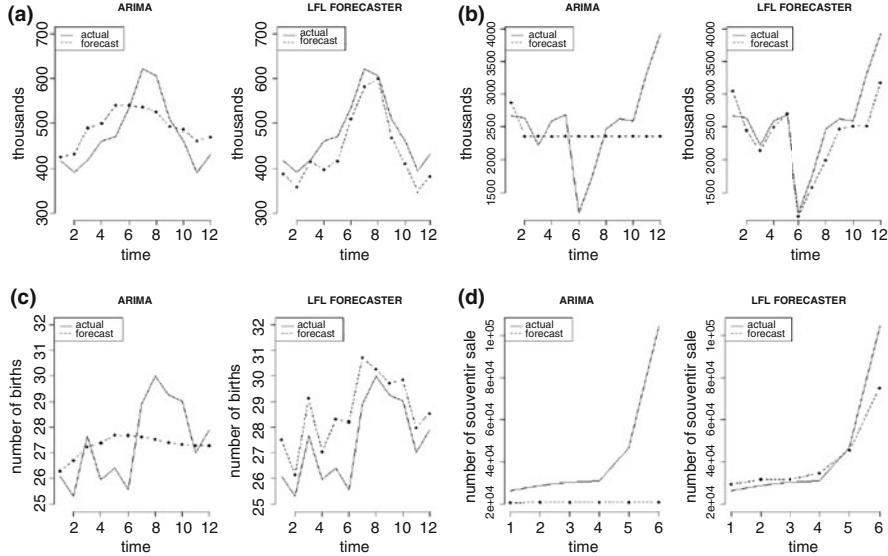


Fig. 2 Forecasts (dotted line) of four selected time series—comparison within the testing (out-sample) data (black line): (a) Air passengers, (b) red wine, (c) births, (d) souvenir sale. The forecast of (a)–(c) is 12 months, (d) is 6 months (cf. Fig. 1)

Table 1 Comparison of ARIMA and LFL Forecaster by RMSE and SMAPE errors

Data	RMSE		SMAPE	
	ARIMA	LFL	ARIMA	LFL
Air passengers	53.36	40.80	0.0938	0.0840
Red wine	677.42	373.99	0.2071	0.1022
Births	1.44	1.36	0.0455	0.0433
Souvenir	36,541	12,409	0.5746	0.1207

5.3 Demonstration of Found Structural Breaks

Datasets In this section, we work with four time series (see Figure 7). Two time series in panel (a) and (b) are the real data and taken from micro subset of time series from the M4-Competition published on the Internet.⁶ The last two time series in panel (c) and (d) are real data on ECG heartbeat taken from the MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database published on the Internet.⁷

⁶ <https://forecasters.org/blog/2018/01/19/m4-competition/>.

⁷ <https://www.kaggle.com/shayanfazeli/heartbeat>.

Signature: $S(t) \& dS(t) \rightarrow dS(t+1)$ {								
1	&	ze	&	ze	&	\rightarrow	&	vr sm
2	&	ml sm	&	ra me	&	\rightarrow	&	ra me
3	&	vr sm	&	ml me	&	\rightarrow	&	ra me
4	&	ra me	&	ml me	&	\rightarrow	&	si sm
5	&	ra me	&	ra me	&	\rightarrow	&	ve bi
6	&	vr bi	&	ex bi	&	\rightarrow	&	ra bi
7	&	ra bi	&	si bi	&	\rightarrow	&	ra me
8	&	ex bi	&	vr bi	&	\rightarrow	&	ml me

Fig. 3 Linguistic description of the air passengers time series forecast. The width of basic functions is $2h = 27$, i.e., $h = 13.5$. For example, from Rule 4, we learn that, in the current 2 years, rather medium number of passengers and its more or less increase will lead to significantly small increase in the following 2 years

Signature: $\bar{S}(t) \& dS(t) \& dS(t-1) \rightarrow \bar{dS}(t+1)$ {								
1	&	ze	&	ra me	&	ex bi	&	\rightarrow
2	&	ze	&	ze	&	ml me	&	\rightarrow
3	&	ml me	&	ra bi	&	ze	&	\rightarrow
4	&	ty me	&	ra me	&	ml me	&	\rightarrow
5	&	qr bi	&	ex bi	&	vr sm	&	\rightarrow
6	&	ra bi	&	ra me	&	vr bi	&	\rightarrow
7	&	si bi	&	vr sm	&	ra me	&	\rightarrow
8	&	ex bi	&	ml sm	&	ro sm	&	\rightarrow

Fig. 4 Linguistic description of the red wine sale time series forecast. The width of basic functions: $2h = 36$, i.e., $h = 18$. For example, from Rule 4 we learn that if, in the current 3 years, typically medium amount of red wine is sold and the medium increase is encountered for the past 6 years, then we may expect significantly big increase in the following 3 years

Signature: $S(t) \& dS(t) \& dS(t-1) \rightarrow dS(t+1)$ {								
1	&	ra sm	&	-ex bi	&	-si bi	&	\rightarrow
2	&	ro ze	&	-ro sm	&	-ex bi	&	\rightarrow
3	&	ze	&	-ex sm	&	-ro sm	&	\rightarrow
4	&	vr sm	&	vr bi	&	-ex sm	&	\rightarrow
5	&	ty me	&	ra me	&	vr bi	&	\rightarrow
6	&	ra me	&	ml sm	&	ra me	&	\rightarrow
7	&	vr bi	&	ra me	&	ml sm	&	\rightarrow
8	&	qr bi	&	qr sm	&	ra me	&	\rightarrow
9	&	ra bi	&	ra me	&	qr sm	&	\rightarrow
10	&	ex bi	&	ra me	&	ra me	&	\rightarrow

}]Linguistic[Rules count: 9,

Fig. 5 Linguistic description of the births time series forecast. The width of basic functions: $2h = 26$. For example, from Rule 7, we learn that if, in the current 2 years, very roughly big number of births happens and rather medium increase and more or less small decrease in the previous 2 years are encountered, then we may expect rather medium increase in the following 3 years

The following periodicities were detected using periodogram:

- Time series (a): 9.2, 11.8, 15.3, 21.8, 28, 32.5, 38.1, 45.8, 57.2, 86.6, 150;
- Time series (b): 7.9, 10.5, 14.4, 18.9, 27.2, 47.9;
- Time series (c): 7.5, 11.2, 18, 22.4, 29.7, 38.4, 45.9, 51.8, 57.5, 61, 67.3, 86.5, 91.6, 102.6;
- Time series (d): 5, 8.8, 13.1, 17.7, 25.7, 33.1, 39.2, 53, 57.6, 69.3, 87.5, 115;

Signature: $S(t) \& dS(t) \rightarrow dS(t+1)$ {							
1	&	ze	&	ml me	&	\rightarrow	&
2	&	vr sm	&	ra me	&	\rightarrow	&
3	&	ro sm	&	-vr sm	&	\rightarrow	&
4	&	ra me	&	vr bi	&	\rightarrow	&
5	&	ml me	&	-ra me	&	\rightarrow	&
6	&	vr bi	&	ro bi	&	\rightarrow	&
7	&	ra me	&	-vr bi	&	\rightarrow	&
8	&	ml me	&	ml me	&	\rightarrow	&
9	&	ra me	&	-ro sm	&	\rightarrow	&
10	&	ex bi	&	ex bi	&	\rightarrow	&
11	&	ex bi	&	ro ze	&	\rightarrow	&

\LinguisticRules count: 9.

Fig. 6 Linguistic description of the souvenir sales time series forecast. The width of basic functions: $2h = 12$. For example, from Rule 10, we learn that if in this year the number of souvenir sales is extremely big and the increase is also extremely big, then we may expect that next year the increase will be zero

On the basis of the detected periodicities (cf. [6]), we put $h_{(a)} = 19$, $h_{(b)} = 9$, $h_{(c)} = 22$, and $h_{(d)} = 20$.

To find the structural breaks, we need to specify the width of the basic functions ($2h$) and the increment interval p , i.e., a shift of the basic function along the time axis. We used the following parameters: (a) $2h = 5$ and $p = 2$; (b) $2h = 5$ and $p = 3$; (c) $2h = 4$ and $p = 2$; and (d) $2h = 5$ and $p = 1$.

Tables 2 and 3 show the results for detection of structural breaks in four time series using the LFL Forecaster software. The intervals typeset in bold font and characterization of their trend point out the sudden change (hugely increasing and hugely decreasing) in the course of the time series. The results above are applied in Fig. 8 to show the structural breaks. Intervals with the increasing trend are depicted in a continuous line, and the decreasing ones are in a dashed line.

6 Conclusion

In this paper, we gave an overview of fuzzy modeling methods for analysis, forecasting, and detecting structural breaks in time series. The methods are based on the theory of fuzzy transform and selected methods of fuzzy natural logic. We provided a demonstration of the forecast and detection of structural breaks on real data taken from the Internet. We compared the forecasts using our methods with those of the ARIMA model. The results demonstrate that our forecasts are well comparable with those obtained using the classical statistical methods. An additional asset of our methods is in the ability to provide information about time series in natural language.

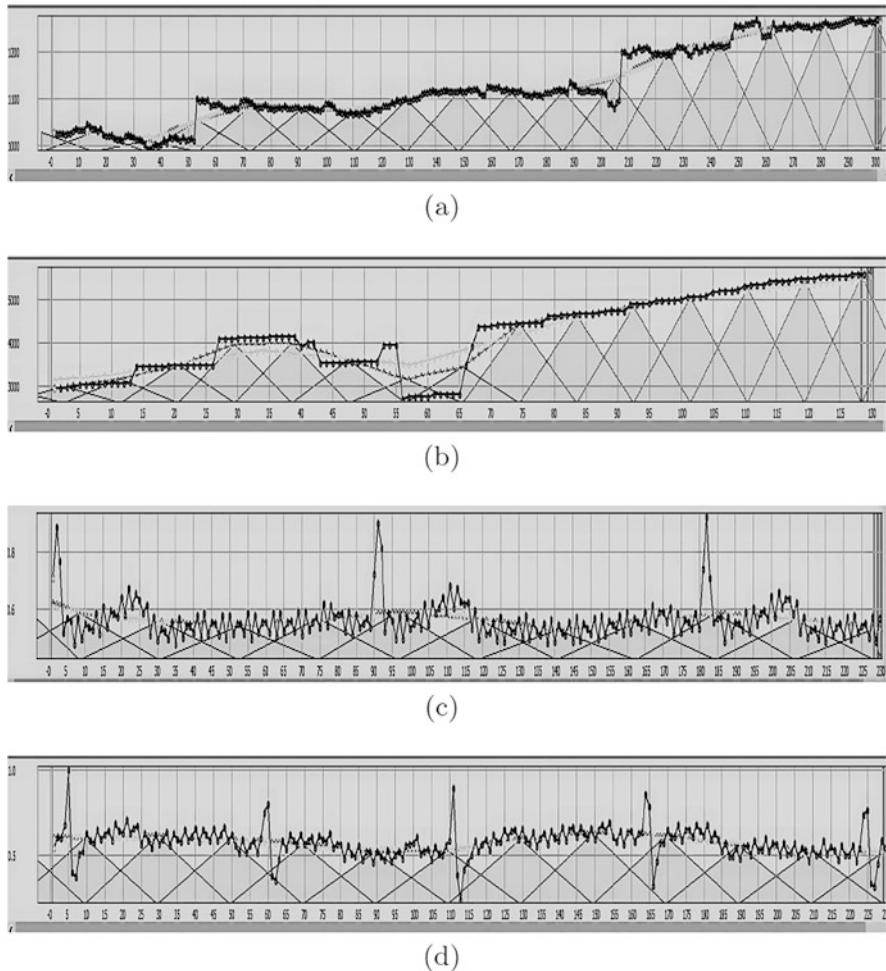


Fig. 7 Four time series with various kinds of structural breaks. Time series (a) and (b) appear breaks increase and decrease suddenly. Time series (c) and (d) breaks occur in roughly equal time periods, and breaks are symmetric and have almost similar shapes

Table 2 Intervals with monotonous trend detected in time series (a) and (b). The evaluative expressions characterizing trend are derived using the function of local perception (3) w.r.t. the context w_{tg} determined by the minimal and maximal values of the time series

Interval	Trend characterization (a)	Interval	Trend characterization (b)
(46,50)	Very little decreasing	(25,29)	Hugely increasing
(50,54)	Hugely increasing	(29,33)	Negligibly increasing
(54,58)	A little decreasing	(41,45)	Hugely decreasing
(204,210)	Hugely increasing	(45,49)	Somewhat increasing
(210,114)	Clearly increasing	(53,57)	Hugely decreasing
(246,250)	Hugely increasing	(57,61)	A little increasing
(250,254)	A little increasing	(65,69)	Hugely increasing

Table 3 Intervals with monotonous trend detected in time series (c) and (d)

Interval	Trend characterization (c)	Interval	Trend characterization (d)
(1,5)	Hugely decreasing	(1,5)	Hugely increasing
(5,8)	Clearly decreasing	(58,63)	Hugely decreasing
(88,91)	Hugely increasing	(107,112)	Hugely increasing
(91,94)	Hugely decreasing	(112,117)	Hugely increasing
(94,97)	Clearly decreasing	(163,168)	Hugely decreasing
(178,181)	Hugely increasing	(217,221)	Stagnating
(181,184)	Hugely decreasing	(221,225)	Hugely increasing

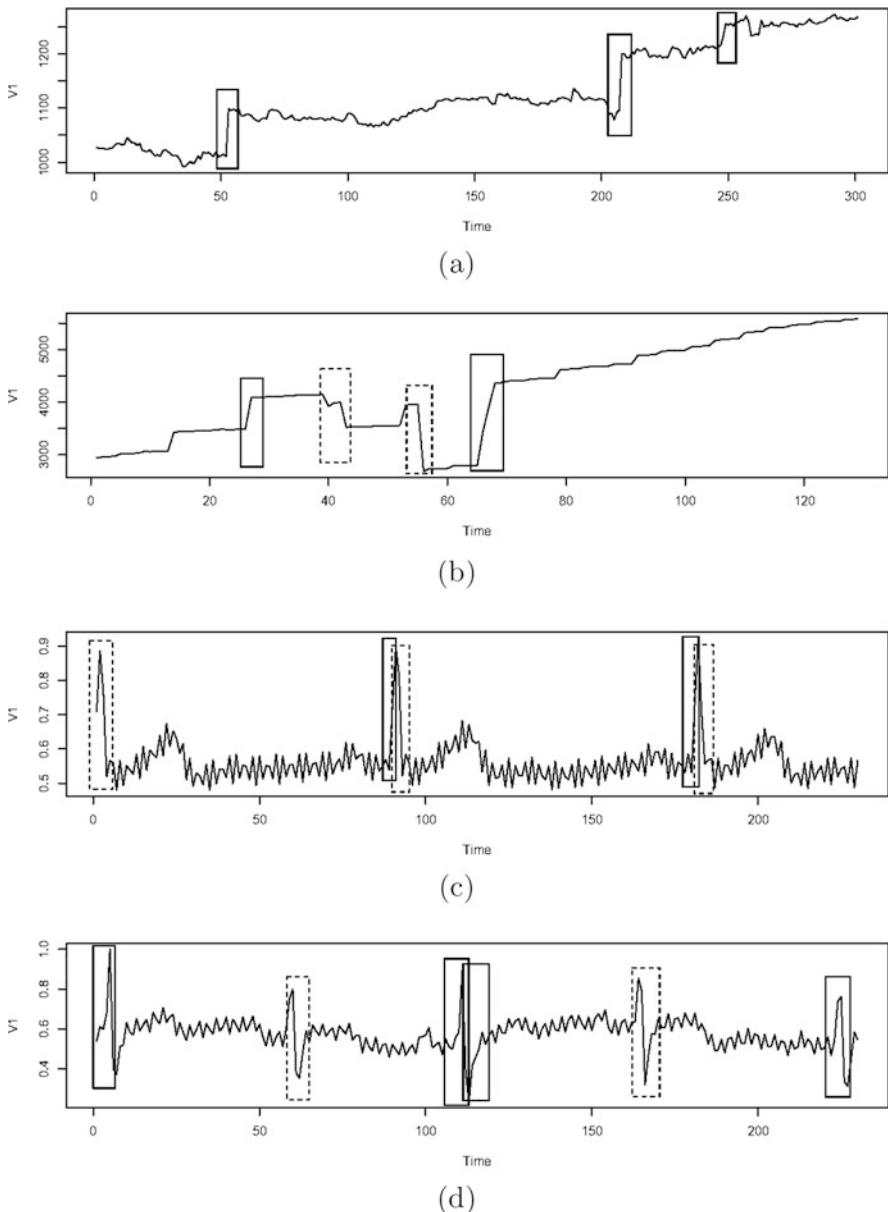


Fig. 8 Structural breaks detected in the four selected time series

References

1. Chatfield, C.: Time Series Forecasting. Chapman& Hall/CRC, Boca-Raton (2000)
2. Kreinovich, V., Perfilieva, I.: Fuzzy transforms of higher order approximate derivatives: A theorem. *Fuzzy Sets Syst.* **180**, 55–68 (2011)
3. Novák, V.: Detection of structural breaks in time series using fuzzy techniques. *Int. J. Fuzzy Logic Intell. Syst.* **18**(1), 1–12 (2018)
4. Novák, V., Pavliska, V.: Time series: how unusual local behavior can be recognized using fuzzy modeling methods. In: Kreinovich, V. (ed.) Statistical and Fuzzy Approaches to Data Processing, with Applications to Econometrics and Other Areas, pp. 157–177. Springer, Berlin (2021)
5. Novák, V., Perfilieva, I.: Non-statistical methods for analysis, forecasting and mining time series. In: Proc. 7th Int. Conference on Time Series and Forecasting. Springer (2021)
6. Novák, V., Perfilieva, I., Holčapek, M., Kreinovich, V.: Filtering out high frequencies in time series using F-transform. *Information Sciences* **274**, 192–209 (2014)
7. Novák, V., Perfilieva, I., Dvořák, A.: Insight into Fuzzy Modeling. Wiley & Sons, Hoboken, New Jersey (2016)
8. Perfilieva, I.: Fuzzy transforms: theory and applications. *Fuzzy Sets Syst.* **157**, 993–1023 (2006)
9. Perfilieva, I., Daňková, M., Bede, B.: Towards a higher degree F-transform. *Fuzzy Sets Syst.* **180**, 3–19 (2011)

Anomaly Detection Algorithm Using a Hybrid Modelling Approach for Energy Consumption Time Series



Florian Rippstein, Steve Lenk, Andre Kummerow, Lucas Richter,
Stefan Klaiber, and Peter Bretschneider

Abstract Many energy time series captured by real-time systems contain errors or anomalies that prevent accurate forecasts of time series evolution. However, accurate forecasting of load time series and fluctuating renewable energy feed-in as well as subsequent optimisation of the dispatch of controllable generators, storage and loads is crucial to ensure a cost-effective, sustainable and reliable energy supply. Therefore, we investigate methods and approaches for a system solution that automatically detect and replace anomalies in time series to enable accurate forecasts. Here, we introduce a hybrid anomaly detection system for energy consumption time series, which consists of two different neural networks (Seq2Seq and autoencoder) and two more classical approaches (entropy, SVM classification). This network is able to detect different types of anomalies, namely, outliers, zero points, incomplete data, change points and anomalous (parts of) time series. These types are defined for the first time mathematically. Our results show a clear advantage of the hybrid modelling approach for detecting anomalies in previously unknown energy time series compared to the single approaches. In addition, due to the generalisation capability of the hybrid model, our approach allows very good estimation of energy values without requiring a large amount of historical data to train the model.

Keywords Anomaly detection · Energy consumption · Time series processing · Seq2Seq · Autoencoder hybrid neural network

F. Rippstein (✉) · S. Lenk · A. Kummerow · L. Richter · S. Klaiber · P. Bretschneider
Fraunhofer-Institut Optronik, Systemtechnik und Bildauswertung (IOSB), Ilmenau, Germany
e-mail: Florian.Rippstein@iosb-ast.fraunhofer.de; Steve.Lenk@iosb-ast.fraunhofer.de;
Andre.Kummerow@iosb-ast.fraunhofer.de; Lucas.Richter@iosb-ast.fraunhofer.de;
Stefan.Klaiber@iosb-ast.fraunhofer.de; Peter.Bretschneider@iosb-ast.fraunhofer.de

1 Introduction

Many energy data sets of real-time systems include errors or anomalies, which hinder an appropriate prediction. However, the prediction and the following optimisation of energy load, generation and storage are crucial to prevent blackouts or brownouts due to unbalanced fluctuations in the energy grid [9]. For critical infrastructures, e.g. the energy sector, new challenges arise due to the increasing amount of data to handle, the increasing automation level and possible threats by cyberattacks. Thus, resilience, i.e. to be prepared for and to prevent threats, to protect systems against them, to respond to threats and to recover from them, became more and more important.

Therefore, we study a system which automatically detects and replaces anomalies in time series to enable accurate predictions.

Thereby, we define anomalies as data, which do not belong to the normal characteristics of time series, whereas errors are normal or anomalous parts of time series, which are known to be erroneous due to external information, e.g. information of fallen power pole.

To classify anomalies, we distinguish outliers, zero points, incomplete data, change points and anomalous (part of) time series similarly to [3, 10], but we formalised their definitions mathematically (see Sect. 2). To study our detection methods, we manipulated real, highly accumulated energy consumption time series, which were manually verified and corrected [1].

An example is shown in Fig. 1 in which a part of such an accumulated energy consumption time series [1] (green) is shown. A classical approach to detect anomalies is to calculate the difference between a prediction and an observation [15]. This difference is called “surprise” by Goldberg et al. [4] and is calculated as the difference between the true and the observed values. Unfortunately, this approach is only applicable if a precise prediction can be calculated, which in case of a regression needs sufficient amount of data. Alternatively, neural networks show good results using unknown data, either by default or by techniques such as domain adaption [16].

Three approaches to detect anomalies in energy data sets were suggested by Zhang et al. [19], namely, using Shannon entropy, classification or a regression approach. For unknown data sets, the regression approach is obviously inadequate since the amount of training data is too small. However, using the well-known Shannon entropy from information theory [12] to measure the surprise or uncertainty of data points in a time series, it is possible to detect anomalous data points in previously unknown time series to a limited amount. The entropy H is calculated as:

$$H(x) = \sum_{i=1}^n p(x_i) \log_b p(x_i),$$

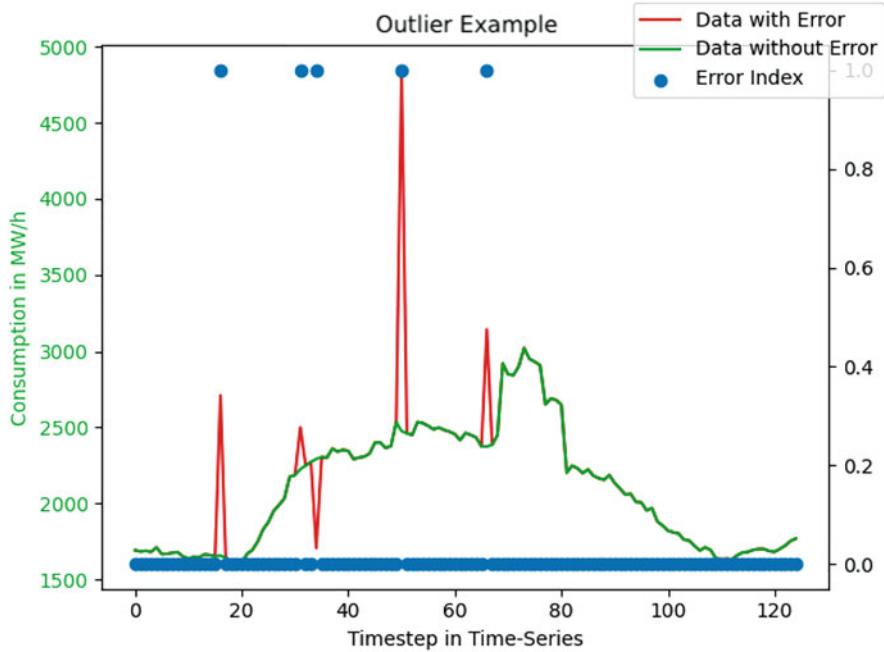


Fig. 1 Example of an anomalous time series including outliers with different anomaly delta

where p is the probability of the energy consumption x . We also have b as the base of the logarithm. The two common used bases are 10 or 2 [12]. However, this measured accuracy and precision is not as high as a regression approach.

A neural network approach can be created by using Seq2Seq networks, which are able to predict values of time series [5, 6]. Thus, we can classify by using the surprise.

Autoencoders otherwise show strong in the reconstruction of data in general [14] and also in time series [11]. Hence, it can also be used to evaluate a time series by calculating a surprise based on the reconstruction error. Furthermore, support vector machines (SVM) have a strong theoretical foundation and are fast implementable to classify data. Yet, SVM have some disadvantages, like overfitting and the need for labelled data, which are the common weaknesses of supervised learning. Additionally, SVM needs good kernel (function) to separate between classes [17], i.e. normal data and anomalies.

To overcome the limitations and drawbacks of these approaches, a hybrid model was developed for all defined anomalies.

2 Our Definitions

In general, we consider a time series X as a sequence of n -tuples:

$$((c_1, t_1), \dots, (c_n, t_n)).$$

The discussed anomalies are defined in the following:

Definition 1 (Noise Data) Noise data is incomprehensible for either computers or unstructured data. These can be logical errors or inconsistent data [3], e.g. string in databases, not detected bit flips.

Definition 2 (Outlier) A time series X^* with outlier can be created by modifying tuples of X by multiplying c_i with factor $o_i \in \mathbb{R}_0^+ \setminus [0.9, \dots, 1.1]$ to the left elements of the chosen tuples were the predecessor and successor of the single tuples are not modified, i.e.

$$(o_i * c_i, t_i), \text{ where } i \in \{2, \dots, n-1\}$$

Then the modified tuple is an outlier.

Definition 3 (Zero Point) Based on Definition 2, an outlier is called zero point if the modifying factor o_i is 0 instead.

Definition 4 (Change Point) For given time series X is $2 \leq m \leq n-2$. Then a time series X^* with change points can be created by replacing a consecutive m -sub-sequence of X by $o_i \in \mathbb{R}_0^+$. Additionally, the first modifier o_j of the sub-sequence has to satisfy $o_j \notin [0.9, \dots, 1.1]$, to the left elements of the chosen tuples were the predecessor and successor of this m -sub-sequence are not modified, i.e.

$$\begin{aligned} & (o_i * c_i, t_i), \text{ where } i \in \{j, \dots, j+m-1\}, |o_i - 1| > |o_{i+1} - 1| \text{ and:} \\ & \quad o_i > 1 \text{ and } o_{i+1} > 1 \text{ or} \\ & \quad o_i < 1 \text{ and } o_{i+1} < 1, \quad \forall i \in \{j, \dots, j+m-1\}. \end{aligned}$$

The points of this consecutive m -sub-sequence are called change points.

Definition 5 (Incomplete Data) For given time series X is $2 \leq m \leq n-2$. A time series X^* with incomplete data can be created by replacing a consecutive m -sub-sequence of X by using factors $o_i \in \mathbb{R}_0^+ \setminus [0.9, \dots, 1.1]$, with o_j being the first modifier of the m -sub-sequence and $o_j = o_i$, where $i \in \{j, \dots, j+m-1\}$,

to the left elements of the chosen tuples were the predecessor and successor of this m -sub-sequence are not modified, i.e.

$$(o_i * c_i, t_i), \text{ where as } i \in \{j, \dots, j + m - 1\}$$

The points of this consecutive m -sub-sequence are called incomplete data.

Definition 6 (Anomalous Time Series/Outlier Type B) For given time series X is $2 \leq m \leq n - 2$. An anomalous time series X^* can be created by replacing a consecutive m -sub-sequence of the n -sequence X by multiplying factors $o_i \in \mathbb{R}$, with o_j being the first modifier of the m -sub-sequence and $o_i \neq 1$, where $i \in \{j, \dots, j + m - 1\}$, to the left elements of the chosen tuples were the predecessor and successor of this m -sub-sequence are not modified, and where the sub-sequence is either incomplete data or change point, i.e.

$$(o_i * c_i, t_i), \text{ where as } i \in \{j, \dots, j - m - 1\}$$

The points of this consecutive m -sub-sequence are called incomplete data.

Information: Anomalous time series are similar to a set of outliers; therefore, we decided to use the name outlier type B.

3 Our Hybrid Model

Our developed architecture is shown in Fig. 2.

It contains of the two previously mentioned neural networks, an autoencoder and a Seq2Seq networks, and the Shannon entropy and SVM as more classical approaches.

Autoencoder is able to reconstruct time series to find anomalous data points [2]. Thus, autoencoder can be trained to reconstruct a time series, and such a reconstructed time series can be compared with the original time series using the mean squared error (MSE) or alternatives like RMSE to classify them.

We improved this approach by calculating the (squared) difference of every single data point and using this as input for a convolutional neural network (CNN), which is trained together with the autoencoder. The training process utilises loss weight to comply with the fact that a good classification is more important than a good reconstruction. To evaluate a whole time series, we used a rolling window (standard size 24 time stamps) to evaluate each single data point with the single autoencoder.

Additionally, we created a Seq2Seq prediction network similar to the network by Hwang et al. [6]. Seq2Seq networks are well known for their strong capabilities in the field of natural language processing [8].

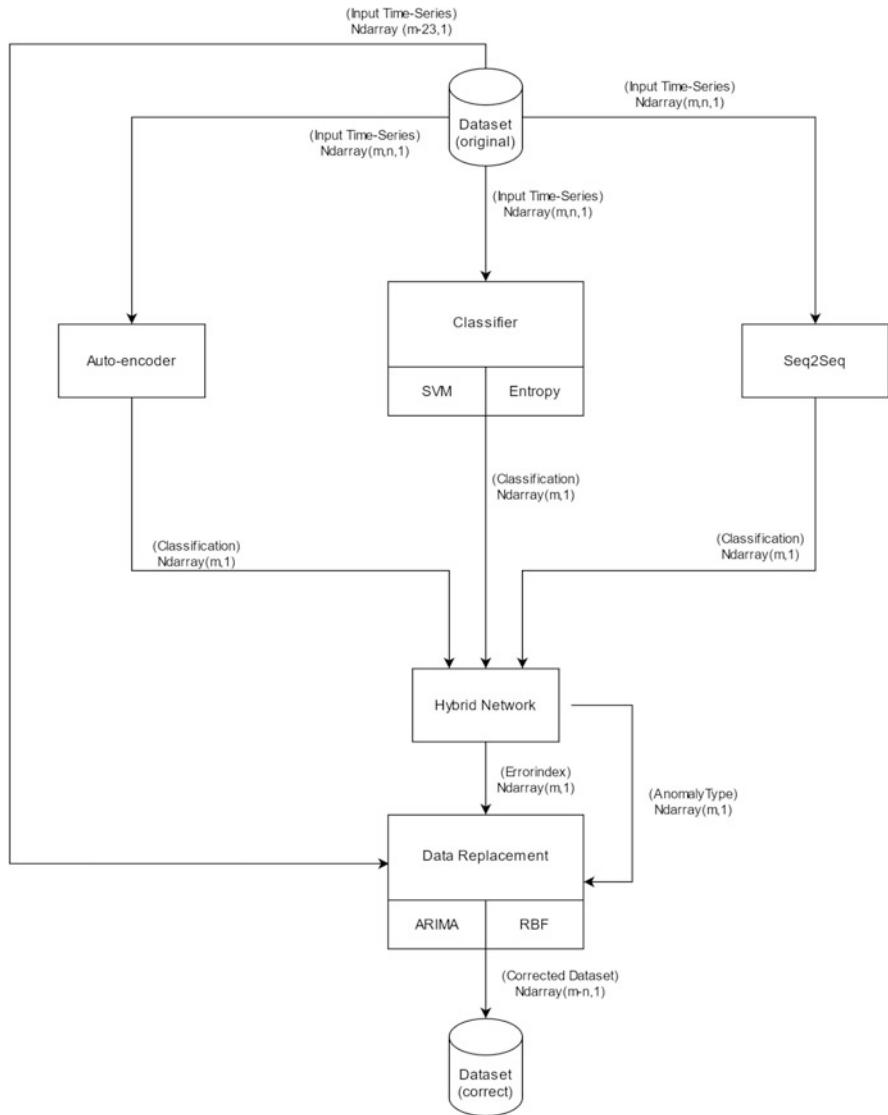


Fig. 2 Our solution

The Seq2Seq networks use the unrolling properties of RNN [13] to evaluate an input. Again, a full time-set evaluation was be done using a rolling window.

By combining the two classical approaches (entropy and SVM) and the two neural networks (autoencoder and Seq2Seq), a hybrid model was built (as shown in Fig. 2), which takes advantage of each of the single approaches. The hybrid network in Fig. 2 itself is a SVM, which evaluates the different results and computes a more

precise final decision. Decision trees or a neural network could be used as well. These approaches have shown similar or even better scores in other tasks [7]. The next step was used to substitute all detected anomalies by using either interpolation, extrapolation or an autoencoder, depending which of those replacement algorithms is suited best for a given time series.

4 Results

Before we show the hybrid results, we explain some benefits of our hybrid solution.

In Fig. 3, we plotted the MSE of anomalies and of normal data after reconstruction by the autoencoder as orange and blue lines, respectively. Here, anomalies have a MSE of approx. 1.0, whereas for normal data, it fluctuates around 0.1. A classification based on the plotted MSE was done by using, e.g 0.4 as the limit for normal data. This approach yields F1-scores around 0.8, but some data points are wrongly classified.

Here, we developed a different approach based on CNN as described in Sect. 3. Instead of using the MSE, we used the squared error in a CNN for each single data point which improved the F1-score. However, the reconstruction result of

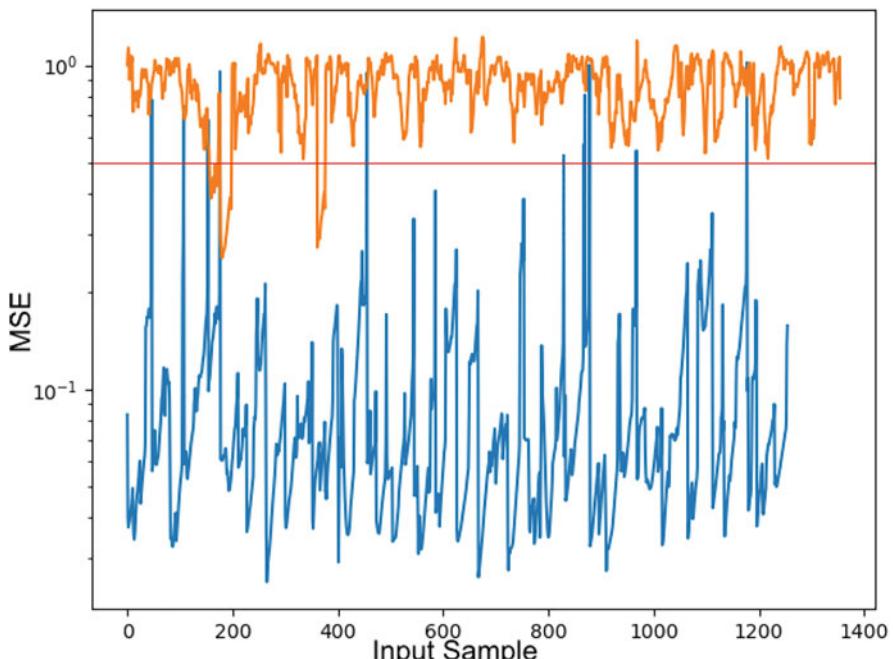


Fig. 3 MSE output of the autoencoder

the autoencoder is no longer usable for replacing the abnormal data, since both networks, autoencoder and CNN, are trained together focusing on MSE for classification. Thus, it will yield a large difference between MSE of normal and abnormal data points but not necessarily anomalies will have a larger MSE.

The Seq2Seq network used the introduced surprise calculating approach. Therefore, the network classifies data by building an internal confidence window [18]. Additionally, we used a similar CNN-based approach as for the autoencoder. This approach showed that the prediction accuracy of a Seq2Seq network depends on the placement of the data point within the sample window, i.e. the closer to the window borders, the worse the prediction accuracy. For better classification results, we combine the different anomaly detection results for a single data point, i.e. 24 decisions for each data point due to a standard rolling window size of 24. The result of an (part of an) energy consumption time series is shown in Fig. 4 as green line. In this figure, the time series is shown as red line and the time stamp of generated anomalies (as a Boolean index in the (not-shown) range between 0 and 1). It is observable that abrupt changes in the time series result in an increased detection rate by the Seq2Seq network as desired. Thus, points with a higher surprise are detected more often than normal data. This Seq2Seq worked to a certain degree as seen in Fig. 4. Here, the network detected a normal spike on data point 30 as outlier, but detected the real outlier only six times. This behaviour is explainable because

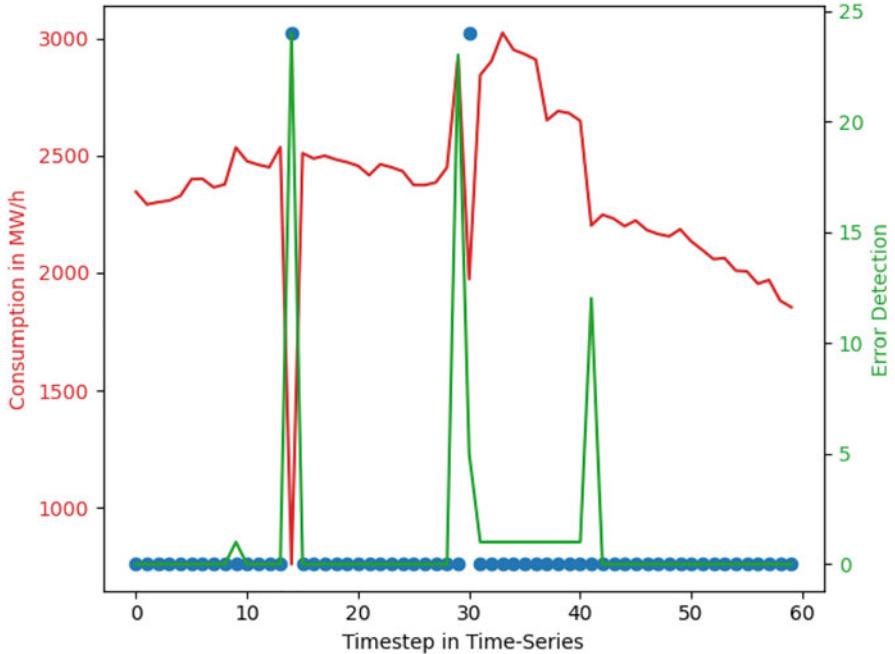


Fig. 4 Seq2Seq output

the network learned that outliers are always single points, and, thus, it is not capable to distinguish correctly between the two data points with high surprise. After adding change points or incomplete data to our train set, this behaviour was not observed anymore. Unfortunately, Seq2Seq networks, trained only with long anomalies, are always detected at least 3 points as anomalies in test with single-point outliers. Our approach of deciding upon majority votes can be used to decrease the amount of false positive or false negative. The hybrid solution is trained on using a higher or lower limit depending on the Seq2Seq networks.

It is notable that the capability of the Seq2Seq network to generalise is not as high as in case of the autoencoder. Therefore, only inter-domain tests can be well detected by the Seq2Seq network. A domain transfer approach is highly recommended to get Seq2Seq networks, which can be usable for a larger variety of data.

So far, we have shown two approaches for detecting anomalies separately, yielding reasonable results, but still improvable ones.

In consequence, this leads to our hybrid network, which combines both approaches. Before presenting the results, we want to emphasise that for the achieved results, our hybrid model was trained with manipulated energy consumption data from Germany and tested it with manipulated consumption data from Austria. So, the evaluation was done with unknown data. The results for the Germany consumption test set showed slightly better results. An example of the F1-score for our networks and the hybrid network can be found in Table 1 and in Fig. 5. Here, we were able to reach F1-scores for outliers above 0.99. Additionally, we studied the influence of the ratio between normal and abnormal values, here called anomaly delta. As shown in Table 1, even anomalies with a deviation of only 5% are detectable by the presented hybrid model. The accuracy for the substitution of outliers is already satisfying as seen in Fig. 5 by comparing the real (broken yellow line) and corrected data (black solid line). The substitution was done with an RBF Interpolation.

If domain adaption techniques were used, the F1-Score of the hybrid solution was decreased by 0.01.

Also the results of the other anomaly types are shown in Table 2.

Table 1 F1-scores for different anomaly deltas

Anomaly delta	Hybrid result	F1-score	Type
10%	0.9976	0.848	Autoencoder
		0.899	Seq2Seq
7.5%	0.9948	0.748	Autoencoder
		0.812	Seq2Seq
6%	0.9917	0.564	Autoencoder
		0.845	Seq2Seq
5%	0.9908	0.567	Autoencoder
		0.904	Seq2Seq

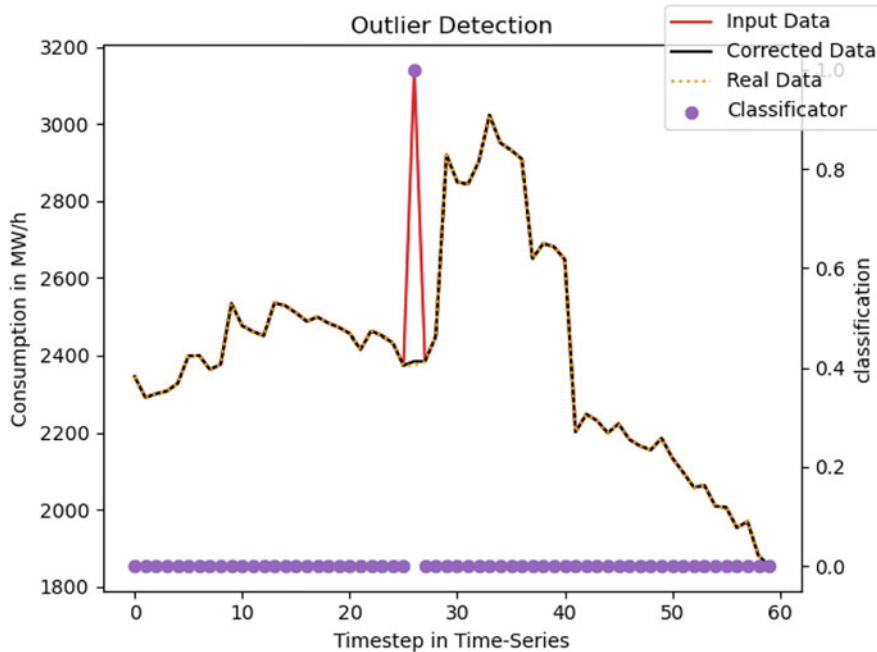


Fig. 5 Example of the hybrid solution with anomaly delta of 10%

Table 2 Comparison of different anomaly types for a delta of 10%

Anomaly type	Autoencoder	Seq2Seq	Hybrid network
Outlier	0.5715	0.8403	0.9941
Incomplete data	0.7970	0.6209	0.8805
Change points	0.8170	0.7623	0.9622

5 Summary

We presented a hybrid model approach that uses two classical mathematical approaches and neural networks to detect anomalies and substitute them with an appropriate algorithm. The results showed clear advantages of the hybrid model for detecting anomalies in previously unknown energy time series compared to the single approaches for outliers, but also for other types of anomalies. In addition, due to the generalisation capability of the hybrid model, this approach allows very good estimation of energy values without requiring a large amount of historical data to train the model.

Our anomaly definitions were defined mathematically based on examples of anomalies and will be adapted to better reflect statistical properties of time series and their anomalies in future studies.

Acknowledgments The work was financially supported by BMBF (Bundesministerium für Bildung und Forschung) under the project “reDesigN - Resilience By Design for IoT Platforms in Distributed Energy Management” [1] (support code 01IS18074D) and Fraunhofer Cluster of Excellence Integrated Energy Systems (CINES). The authors want to acknowledge Prof Mäder and M. Sc. Martin Rabe (TU Ilmenau) for their supervision of the master thesis “Automatic energy data processing based on machine learning algorithms” of one of us (F.R.). Additionally, we thank B. Sc. Jonathan Schäfer (FSU Jena) for the fruitful discussions about the mathematical definition of the anomaly types.

References

1. Bundesnetzagentur.: SMARD | SMARD - Strommarktdaten, Stromhandel und Stromerzeugung in Deutschland, Mar 2021. [Online; accessed 12. Mar. 2021]
2. Chaitanya, C.R.A., Kaplanyan, A.S., Schied, C., Salvi, M., Lefohn, A., Nowrouzezahrai, D., Aila, T.: Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Trans. Graph. (TOG)* **36**(4), 1–12 (2017)
3. Chen, W., Zhou, K., Yang, S., Wu, C.: Data quality of electricity consumption data in a smart grid environment. *Renewable and Sustainable Energy Reviews* **75**, 98–105 (2017)
4. Goldberg, D., Shan, Y.: The importance of features for statistical anomaly detection. In: 7th {USENIX} Workshop on Hot Topics in Cloud Computing (HotCloud 15) (2015)
5. Gong, G., An, X., Mahato, N. K., Sun, S., Chen, S., Wen, Y.: Research on short-term load prediction based on seq2seq model. *Energies* **12**(16), 3199 (2019)
6. Hwang, S., Jeon, G., Jeong, J., Lee, J.: A novel time series based seq2seq model for temperature prediction in firing furnace process. *Procedia Comput. Sci.* **155**, 19–26 (2019)
7. Kirkos, E., Spathis, C., Manolopoulos, Y.: Support vector machines, decision trees and neural networks for auditor selection. *J. Comput. Methods Sci. Eng.* **8**(3), 213–224 (2008)
8. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: Open-source toolkit for neural machine translation. In: Proceedings of ACL 2017, System Demonstrations (July 2017), pp. 67–72
9. Kummerow, A., Klaiber, S., Nicolai, S., Bretschneider, P., System, A.: Recursive analysis and forecast of superimposed generation and load time series. In: International ETG Congress 2015; Die Energiewende - Blueprints for the New Energy Age, pp. 1–6 (2015)
10. Laptev, N., Amizadeh, S., Flint, I.: Generic and scalable framework for automated time-series anomaly detection. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1939–1947 (2015)
11. Liguori, A., Markovic, R., Dam, T. T. H., Frisch, J., van Treeck, C., Causone, F.: Indoor environment data time-series reconstruction using autoencoder neural networks. Preprint (2020). arXiv:2009.08155
12. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
13. Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Phys. D Nonlinear Phenomena* **404**, 132306 (2020)
14. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1274–1283 (2017)
15. von Werra, L., Tunstall, L., Hofer, S.: Unsupervised anomaly detection for seasonal time series. In: 2019 6th Swiss Conference on Data Science (SDS), pp. 136–137 (2019)
16. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153 (2018)

17. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2**(2), 165–193 (2015)
18. Yu, Y., Zhu, Y., Li, S., Wan, D.: Time series outlier detection based on sliding window prediction. *Math. Prob. Eng.* **2014**, Article ID 879736 (2014). <https://doi.org/10.1155/2014/879736>
19. Zhang, Y., Chen, W., Black, J.: Anomaly detection in premise energy consumption data. In: 2011 IEEE Power and Energy Society General Meeting, pp. 1–8 (07 2011)

Unit Root Test Combination via Random Forests



Luca Nocciola, Daniel Ollech, and Karsten Webel

Abstract There is a wide variety of non-seasonal and seasonal unit root tests. However, it is not always obvious which tests can be relied upon due to uncertainties in identifying the data generating process, often with respect to the presence of deterministic terms and the initial conditions. We evaluate the size and power of a large set of unit root tests on time series that are simulated to be representative of economic time series in the M4 competition data. Furthermore, using a conditional random forest-based elimination algorithm, we assess which tests should be combined to improve the performance of each individual test.

Keywords ARIMA time series · Conditional inference trees · Monte Carlo simulation · Sequential testing · Supervised machine learning

1 Introduction

The question of whether a given macroeconomic time series contains a unit root is important since the presence (absence) of such a root implies that exogenous shocks have a persistent (transient) effect on the data. It is also well-known that unit roots

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the European Central Bank. The authors thank Uwe Hassler, Mehdi Hosseinkouchack and participants of the 2020 Conference on Machine Learning of Dynamic Processes and Time Series Analysis and of the 2021 International Conference on Time Series and Forecasting for valuable comments.

L. Nocciola (✉)
European Central Bank, Frankfurt, Germany
e-mail: luca.nocciola@ecb.europa.eu

D. Ollech and K. Webel
Deutsche Bundesbank, Frankfurt, Germany
e-mail: daniel.ollech@bundesbank.de; karsten.webel@bundesbank.de

cause problems in inference and forecasting. First, we know that, for independent random walks, we cannot rely on sample correlations and regression coefficients as consistent measures of the relationships within the population, since they can spuriously take on any value within the interval $[-1, 1]$ and \mathbb{R} , respectively, even asymptotically. Second, the rejection rate of standard t -tests for the null hypothesis of a unit root increases with sample size. Accordingly, these tests are not helpful due to the fact that t -statistics do not converge to any asymptotic distribution unless properly normalised at a customised rate [16, 20, 23, 36, 49]. Third, forecast accuracy under unit roots in the autoregressive polynomial steadily deteriorates as the horizon increases, since the forecast error variance grows linearly with the forecast horizon and tends towards infinity in the limit [6], contrary to the stationary case.

The large number of competing non-seasonal and seasonal unit root tests can complicate the researcher's decision regarding which test to apply in a given situation, even if the seasonal status of the data is known, or suggested by a seasonality test. Size distortions and power issues have been reported for various scenarios of practical importance [48], including finite samples [7], near-unit root behaviour and particularly large negative roots in the moving average polynomial [30, 40]. Contradictory conclusions have also been repeatedly reported when several tests are applied to the same data, especially for seasonal unit root tests as a result of the different model specifications that they impose [40]. Uncertainty surrounding the deterministic mean function and the initial conditions is also known to affect the tests' performance [14, 29, 47] and, moreover, make the assumptions of standard testing procedures unlikely to be met. Various strategies for dealing with these types of uncertainty have been suggested for non-seasonal unit root tests, including sequential pre-testing for trend specifications and unit roots, data-dependent weighted averaging of unit root tests and running union-of-rejections decision rules [1, 22, 35]. Although these strategies may involve multiple tests, they usually work with only a small preselection of tests, often from the same family. We elaborate on this approach by reinterpreting the unit root hypothesis as a classification problem with two classes. Our sequential approach utilises the conditional random forest classifier to identify, rank and combine the most informative tests. It is thus capable of incorporating a larger set of unit root tests and naturally extends to seasonal unit root tests. Our approach can help practitioners to select unit root tests with lower misclassification rates.

The remainder of this paper is organised as follows: Sect. 2 reviews widely applied tests for non-seasonal and seasonal unit roots, Sect. 3 provides basic information about random forests, Sect. 4 explains our proposed testing strategy, Sect. 5 reports our results, and, finally, Sect. 6 summarises.

2 Unit Root Tests

Let $\{y_t\}$ be a discrete time series with τ observations per year and assume that the series can be adequately described by the model:

$$\phi(B)(y_t - \mu_t) = \theta(B)\varepsilon_t , \quad (1)$$

where B is the backshift operator, $B^k x_t = x_{t-k}$, $\phi(B)$ is an autoregressive (AR) polynomial that has roots on or outside the unit circle, $\theta(B)$ is a moving average (MA) polynomial that has roots outside the unit circle, $\{\mu_t\}$ is a deterministic function of time and $\{\varepsilon_t\}$ is a zero-mean white noise process with finite variance.

The series $\{y_t\}$ is said to be integrated of order $d \in \mathbb{N}_0$, denoted by $\{y_t\} \sim I(d)$, if $\phi(B)$ contains the factor $(1 - B)^d$. Similarly, the series is said to be seasonally integrated of order $D \in \mathbb{N}_0$, denoted by $\{y_t\} \sim SI(D)$, if $\phi(B)$ contains the factor $(1 - B^\tau)^D$. Thus, $\{y_t\}$ is stationary around $\{\mu_t\}$ if $(d, D) = (0, 0)$, non-stationary if $(d, D) \neq (0, 0)$ and invertible in either case.

2.1 Non-seasonal Unit Roots

The problem of testing for a non-seasonal unit root can be stated as

$$\mathcal{H}_0 : \{y_t\} \sim I(1) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim I(0) \quad (2)$$

and several tests have been suggested under different assumptions for the AR and MA polynomials in (1), with the common assumption being that there is no seasonality, or at least no seasonal unit root ($D = 0$), in the data. The function $\{\mu_t\}$ is usually represented by

$$\mu_t = \mu + \beta t , \quad (3)$$

covering absence of deterministic terms ($\mu = \beta = 0$) as well as deviations of $\{y_t\}$ from a constant mean ($\beta = 0$), from a linear trend with a zero mean ($\mu = 0$) and from a linear trend with a non-zero mean (unrestricted model).

The Dickey-Fuller (DF) test [10, 11] considers the pure first-order AR case, i.e. $\phi(B) = 1 - \rho B$ and $\theta(B) = 1$, so that (2) is reduced to testing $\rho = 1$ against $\rho < 1$. The DF test is based on the time series regression:

$$y_t = \mu_t + \rho y_{t-1} + \eta_t , \quad (4)$$

where $\{\eta_t\}$ coincides with $\{\varepsilon_t\}$ of (1) under Gaussianity and the initial condition is $y_0 = 0$. The two proposed one-sided tests are the conventional t -statistic $DF_\tau = (\hat{\rho} - 1) \times \hat{\sigma}_{\hat{\rho}}^{-1}$ and $DF_\rho = T(\hat{\rho} - 1)$, where T is the sample size and estimates are

obtained by ordinary least squares (OLS). However, the null distribution of either statistic depends on the form of $\{\mu_t\}$ and is non-standard in each case, so that revised sets of critical values apply [11, 18].

Higher-order AR and ARMA models can be dealt with by the augmented Dickey-Fuller (ADF) test [41], even for unknown model orders. Letting $\Delta = 1 - B$, this test is usually carried out by testing $\rho^* = 0$ against $\rho^* < 0$ in the augmented regression:

$$\Delta y_t = \mu_t + \rho^* y_{t-1} + \sum_{j=1}^{k-1} \phi_j^* \Delta y_{t-j} + \varepsilon_t , \quad (5)$$

using the OLS t -statistic $ADF_\tau = \hat{\rho}^* \times \hat{\sigma}_{\hat{\rho}^*}^{-1}$. If the lag length satisfies $k = p$ for pure AR(p) models and $k \rightarrow \infty$ at a controlled rate as $T \rightarrow \infty$ for ARMA models, then the DF critical values apply. Counterparts to critical F -values for testing against specific trend alternatives have also been tabulated [11].

Allowing for correlated and possibly heterogeneously distributed innovations, the Phillips-Perron (PP) statistics PP_τ and PP_ρ [37, 38] rely on non-parametric transformations rather than augmentation to correct the DF_τ and DF_ρ statistics for the effects of nuisance parameters associated with the distribution of $\{\eta_t\}$. The PP statistics also follow the DF limiting null distributions and thus the DF critical values apply for each form of $\{\mu_t\}$.

Several extensions of the (A)DF and PP tests have been suggested, especially with respect to the proper treatment of the uncertainty surrounding $\{\mu_t\}$. For example, the Zivot-Andrews (ZA) test [50] allows for the estimation of breakpoints in $\{\mu_t\}$ under the alternative hypothesis, and the Elliott-Rothenberg-Stock (ERS) tests [15] are essentially feasible asymptotically point-optimal DF_ρ tests of $\rho = \bar{\rho}$ with $\bar{\rho} < 1$ being fixed against local-to-unity alternatives, allowing for polynomial trends. The test statistic is $ERS_\rho = [S(\bar{\rho}) - \bar{\rho} S(1)] \times \hat{\omega}^{-2}$, where $S(\alpha)$ is the sum of the squared quasi- α -detrended series $\{y_t - \hat{\mu}_t^{(\alpha)}\}$ and $\hat{\omega}^2$ is a consistent estimator of the sum of the covariances of $\{\eta_t\}$. A special member of the ERS family is a modified ADF $_\tau$ test that results from a specific choice of $\bar{\rho}^*$ given T and nominal size. This test, denoted by ERS_τ , is essentially the ADF $_\tau$ test applied to the quasi- $\bar{\rho}^*$ -detrended series in (5) without $\{\mu_t\}$.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [26] considers the problem of testing for (trend) stationarity against non-stationarity, i.e.

$$\mathcal{H}_0 : \{y_t\} \sim I(0) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim I(1) ,$$

assuming that $\{y_t\}$ is decomposable into a deterministic trend, random walk and stationary error. The test regression is essentially model (4) with $\rho = 0$ and $\{\mu_t\}$ being replaced with a random walk $\{\tilde{\mu}_t\}$:

$$y_t = \tilde{\mu}_t + \beta t + \eta_t \text{ with } \tilde{\mu}_t = \tilde{\mu}_{t-1} + \kappa_t , \quad (6)$$

where $\{\kappa_t\}$ is white noise with a zero mean and finite variance σ_κ^2 . The KPSS test is the one-sided upper-tail Lagrange multiplier (and also the locally best invariant) test of $\sigma_\kappa^2 = 0$ against $\sigma_\kappa^2 > 0$, i.e. $\text{KPSS} = \hat{\sigma}_\eta^{-2} \sum_{t=1}^T S_t^2$, where $\{S_t\}$ is the partial-sum process of the residuals estimated from (6) under the null hypothesis and $\hat{\sigma}_\eta^2$ is the sum of the squared residuals divided by T . The asymptotic distribution of the KPSS statistic had been derived initially under the assumption that $\{\eta_t\}$ is Gaussian white noise with finite variance but has then been shown to also hold under weaker (strong mixing) regularity conditions of the PP tests [38].¹

As an alternative to classical unit root tests, the Gómez-Maravall (GM) algorithm [19] determines the appropriate non-seasonal differencing order for ARIMA models by iteratively analysing the roots in the AR and MA polynomials of different predefined seasonal ARMA models. Any root of the characteristic polynomial is defined to indicate a unit root if its modulus is larger than a predefined threshold that, by default, depends on the ARMA model under consideration.

2.2 Seasonal Unit Roots

In analogy to (2), several tests for the general problem

$$\mathcal{H}_0 : \{y_t\} \sim SI(1) \text{ versus } \mathcal{H}_1 : \{y_t\} \sim SI(0) \quad (7)$$

exist, where the deterministic component $\{\mu_t\}$ may now also contain seasonal dummies in (1), so that (3) is extended to

$$\mu_t = \mu + \beta t + \boldsymbol{\gamma}^\top \mathbf{d}_t , \quad (8)$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_\tau)^\top$ is the vector of the τ fixed seasonal effects and $\mathbf{d}_t = (d_{1t}, \dots, d_{\tau t})^\top$ with $d_{it} = 1$ if t falls within the i -th month or quarter and $d_{it} = 0$ otherwise.

The Dickey-Hasza-Fuller (DHF) test [12] is essentially a seasonal variant of the DF test as it considers the pure first-order seasonal AR case, i.e. $\phi(B) = 1 - \rho_\tau B^\tau$ and $\theta(B) = 1$, so that (7) is reduced to testing $\rho_\tau = 1$ against $\rho_\tau < 1$ in the time series regression:

$$y_t = \mu_t + \rho_\tau y_{t-\tau} + \eta_t ,$$

where $\{\mu_t\}$ is now given by (8). The two proposed one-sided tests are based on the conventional t -statistic $DHF_\tau = (\hat{\rho}_\tau - 1) \times \hat{\sigma}_{\hat{\rho}_\tau}^{-1}$ and on $DHF_\rho = T(\hat{\rho}_\tau - 1)$. OLS and symmetric least squares variants (including revised sets of critical values) are provided for the zero-mean ($\mu = \beta = 0$ and $\boldsymbol{\gamma} = \mathbf{0}$), constant-mean ($\mu \neq 0, \beta = 0$

¹ To accommodate weaker assumptions about the errors, the KPSS test requires a consistent long-variance estimator rather than a variance estimator.

and $\boldsymbol{\gamma} = \mathbf{0}$) and seasonal-means cases ($\mu = \beta = 0$ and $\boldsymbol{\gamma} \neq \mathbf{0}$). An ADF-like extension, which is referred to as the augmented DHF (ADHF) test, can be obtained via augmentation with lags of $\{\Delta_\tau y_t\}$ as in (5), where $\Delta_\tau = 1 - B^\tau$.

The Osborn-Chui-Smith-Birchenhall (OCSB) test [32, 40] considers $\phi(B) = (1 - \rho B)(1 - \rho_\tau B^\tau)$ and tests $(\rho, \rho_\tau) = (1, 1)$ against $(\rho, \rho_\tau) \neq (1, 1)$ in the (re-parameterised) regression:

$$\Delta \Delta_\tau y_t = \mu_t + \beta_1 \Delta_\tau y_{t-1} + \beta_2 \Delta y_{t-\tau} + \eta_t \quad (9)$$

with the F -test for $(\beta_1, \beta_2) = (0, 0)$. However, model (9) also allows for sequential t -tests against one-sided (stationary) alternatives. If $\beta_2 = 0$, then the t -test for $\beta_1 = 0$ is the (A)DF test for the need of Δ in addition to Δ_τ . Similarly, if $\beta_1 = 0$, then the t -test for $\beta_2 = 0$ is the DHF test for the need of Δ_τ in addition to Δ . Revised critical values apply in either case. Moreover, assuming the validity of Δ , (9) can be rewritten as

$$\Delta_\tau z_t = \mu_t + \beta_1 S(B) z_{t-1} + \beta_2 z_{t-\tau} + \eta_t ,$$

where $\{z_t\} = \{\Delta y_t\}$ and $S(B) = 1 + B + \dots + B^{\tau-1}$ is the annual aggregation operator, allowing for separate treatments of the roots in $\Delta_\tau = \Delta S(B)$.

The Hylleberg-Engle-Granger-Yoo (HEGY) test [2, 17, 25] extends this factorisation principle even further as it expands $\phi(B)$ about all roots of Δ_τ , additionally allowing for individual assessments (and different moduli) of the $\tau - 1$ unit roots of $S(B)$. The test regression reads

$$\Delta_\tau y_t = \mu_t + \sum_{i=1}^{\tau} \pi_i x_{i,t-1} + \eta_t ,$$

where the processes $\{x_{i,t-1}\}$ are non-singular linear transformations of lagged versions of $\{y_t\}$. The DHF null hypothesis thus implies that $\pi_i = 0$ for all $i \in \{1, \dots, \tau\}$, but specific sub-hypotheses for single (real-valued or pairs of complex-valued) unit roots in Δ_τ can also be tested with separate t -tests and F -tests, respectively. Depending on the specified alternative and the form of $\{\mu_t\}$, the DF and DHF critical values apply except for $\boldsymbol{\gamma} \neq \mathbf{0}$, where critical values are found through simulations by [25].

The GM algorithm discussed above can also be used to determine the seasonal differencing order.

3 Random Forests

A random forest (RF) [4] is a supervised machine learning algorithm. It is a collection of classification trees constructed on bootstrap samples of the original training data. Additionally, at each split, the trees are restricted in that only a

subsample of the available predictors is taken into consideration to determine the optimal split. A conditional RF [24] enhances the classical RF in the presence of correlated predictors.

3.1 Classical Random Forests

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_P)$ be a set of predictors with $\mathbf{x}_j = (x_{1j}, \dots, x_{Nj})^\top$ for all $j \in \{1, \dots, P\}$ and $\mathbf{y} = (y_1, \dots, y_N)^\top$ be a vector of N responses with $y_i \in \{0, \dots, k\}$ for all $i \in \{1, \dots, N\}$, constituting our training set \mathcal{L} .² Let \mathcal{L}_b be a bootstrap sample of size N drawn from \mathcal{L} , from which we grow an unpruned classification tree \mathcal{T}_b with M terminal nodes corresponding to M classification regions. To create a binary split of any terminal node m , we draw a random sample $\tilde{\mathbf{X}}$ of size $\tilde{P} < P$ without replacement from \mathbf{X} . For each sampled predictor $\tilde{\mathbf{x}}_j$, we determine the best split of node m amongst all possible splits of m , and the predictor that generates it (say $\tilde{\mathbf{x}}_{j*}$) is chosen as the splitting predictor for m . The optimal split is identified by, for example, the Gini index. Let $\hat{q}_{mk} = N_m^{-1} \sum_{\mathbf{x}_i \in R_m} \mathbb{I}\{y_i = k\}$ be the share of training data in node m from class k , where $N_m = \sum_i \mathbb{I}\{\mathbf{x}_i \in R_m\}$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{iP})$ and R_m denote the i -th observation of the P predictors and the classification region corresponding to m , respectively. The Gini index is given by

$$Q_m(\mathcal{T}_b) = \sum_k \hat{q}_{mk}(1 - \hat{q}_{mk}) .$$

We stop the trees from growing whenever a pre-specified minimum number of observations in the terminal nodes is reached (say N_{\min}) or node impurity cannot be decreased further. Classification in the RF is obtained by an unweighted mode of the tree classifications. Moreover, we can determine predictor importance in an RF, for example, via the mean decrease in prediction accuracy after randomly permuting the values of the predictor \mathbf{x}_j in the out-of-bag samples (i.e. the training data not bootstrapped in \mathcal{L}_b), denoted by \mathcal{O}_b , to mimic the absence of \mathbf{x}_j . Let $\hat{y}_i(\mathcal{T}_b, \mathbf{x}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{x}_{\pi(j)})$ be the predicted classes of y_i obtained from \mathcal{T}_b before and after random permutation of the values of \mathbf{x}_j in \mathcal{O}_b , where $\pi(\cdot)$ is the permutation scheme. The predictor importance of \mathbf{x}_j is then given by

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

where B is the number of classification trees in the forest.

² In our case, y_i is binary ($k = 1$). However, $k = 3$ could also be used if we considered the combination of possibilities of having seasonal and non-seasonal unit roots.

3.2 Conditional Random Forests

The conditional RF introduces two adaptations with respect to predictor selection and the predictor importance measure. Predictors that are not related to \mathbf{y} (identified via a test of independence between \mathbf{y} and $\tilde{\mathbf{x}}_j$) are excluded in advance, whilst, amongst the predictors that exhibit association with \mathbf{y} (via the same test), the predictor with the strongest association with \mathbf{y} , say $\tilde{\mathbf{x}}_{j*}$, is chosen as a splitting predictor used to determine the optimal split of node m . Let each node m be represented by a vector of integer case weights $\mathbf{w}_m = (w_{m,1}, \dots, w_{m,N})^\top$, where $w_{m,i} > 0$ if (\mathbf{x}_i, y_i) belongs to m and equals zero otherwise. Formally, the global null hypothesis can be formulated as

$$\mathcal{H}_0^{(m)} : \bigcap_{j=1}^{\tilde{P}} \mathcal{H}_0^{(m,j)} \text{ with } \mathcal{H}_0^{(m,j)} : \mathcal{D}(\mathbf{y}|\tilde{\mathbf{x}}_j, \mathbf{w}_m) = \mathcal{D}(\mathbf{y}|\mathbf{w}_m) ,$$

where $\mathcal{D}(\cdot)$ denotes an arbitrary distribution. The rejection rule of this hypothesis is based on the minimal (adjusted) p -value for rejecting the local hypothesis $\mathcal{H}_0^{(m,j)}$. The predictor $\tilde{\mathbf{x}}_{j*}$ can be selected from the local null hypothesis $\mathcal{H}_0^{(m,j*)}$ rejected at the smallest (adjusted) p -value. Finally, $\tilde{\mathbf{x}}_{j*}$ is used as a splitting predictor to find the best binary split of m according to a pre-specified splitting criterion, after which the case weights \mathbf{w}_{m_L} and \mathbf{w}_{m_R} of the left and right descendants of m are computed.

To determine predictor importance, the conditional RF uses a conditional permutation taking into account correlations amongst predictors, thus preventing seemingly influential predictors (due to their correlation with the truly influential predictors) from being attached high importance. The random permutation $\pi(\cdot)$ is now applied to the values of \mathbf{x}_j only within subgroups of observations, say \mathbf{x}_j^C ³. The conditional permutation-based predictor importance is given by

$$PI^C(\mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^B \sum_{i \in \mathcal{O}_b} \left[\frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_{\pi(j)}|\mathbf{x}_j^C)\}}{|\mathcal{O}_b|} - \frac{\mathbb{I}\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{x}_j)\}}{|\mathcal{O}_b|} \right] , \quad (10)$$

where $\pi(\cdot)|\mathbf{x}_j^C$ is the conditional permutation scheme and, for each tree \mathcal{T}_b , the permutation grid for \mathbf{x}_j is given by the cut-points of \mathbf{x}_j^C in \mathcal{T}_b .

³ Alternatively, though less conservative, we can condition only on those predictors \mathbf{x}_j^C whose correlation with \mathbf{x}_j exceeds a certain threshold.

4 Test Evaluation

We aim to evaluate the unit root tests on ARIMA time series that are representative of the data used in the M4 competition [27, 28]. To this end, we simulate correlated Gaussian ARMA parameters and transform them so that their distribution is similar to the empirical distribution estimated from the M4 competition data [31]. This is achieved by combining the “NORmal-To-Anything” (NORTA) algorithm [5] with logspline density estimation [44]. We simulate ARMA parameters and use these to simulate 120,015 time series. The set consists of 40,005 time series with 5, 10 and 20 years of monthly data, respectively. The set of time series is balanced in the sense that half of them contain a (seasonal) unit root. We then run the unit root tests on the raw and first-differenced simulated data and use their p -values as predictors in the conditional RF in order to classify the simulated data into time series with and without unit roots.⁴ Finally, we use (10) to quantify the importance of each individual test.

Based on the test results, we run a conditional RF-based recursive feature elimination (CRFE) algorithm [46]. The key idea is to start with the full set of candidate tests and to grow a sequence of conditional RF by (1) eliminating the least important test in each round and (2) re-estimating the conditional RF with the reduced set of tests in the next round. This scheme is repeated until an “optimal” set of tests is obtained. In each round, we use the conditional RF to classify a test set of simulated time series not considered in tree growing. By evaluating the means and standard deviations of the misclassification rates in each round, we can assess the performance of the respective set of tests and eventually identify the “optimal” set.

5 Results

Table 1 reports the rejection rates for the non-seasonal unit root tests, using a nominal 1% significance level.⁵ The ERS and GM tests have acceptable size properties but the former have relatively low power. The ADF, KPSS and PP tests display severe size distortions but compensate those with a relatively high power (above 70% except for the ADF test without deterministic terms). The ZA test has the worst size properties but the highest power amongst all tests considered. Overall, size issues seem to increase with the length of series for most tests, whereas the power tends to improve. Similar results are reported in [8, 9, 13, 42]. In particular, [8] do not recommend using unit root tests for series with less than 100 observations.

⁴ For computational reasons and parsimony, the tests have been performed with minimal lag length. Seasonal differences and the seasonal status of the series are not considered. The finite-sample critical values that we obtained from our own simulations in R (version 3.5.1) are very similar to the ones tabulated in the original papers (details are available upon request).

⁵ Similar results are obtained for a nominal 5% significance level.

Table 1 Rejection rates for non-seasonal unit root tests (%), $\alpha = 0.01$

Test	Model (3)	$d = 0$				$d = 1$			
		All	5-year	10-year	20-year	All	5-year	10-year	20-year
ADF _{τ}	$\mu = 0, \beta = 0$	25.5	17.2	23.2	36.2	19.0	17.0	19.1	20.9
	$\mu \neq 0, \beta = 0$	76.6	73.7	77.5	78.6	23.4	21.2	24.4	24.6
	$\mu \neq 0, \beta \neq 0$	88.9	82.1	92.8	91.9	32.7	27.8	35.4	34.9
ERS _{τ}	$\mu \neq 0, \beta = 0$	31.6	19.3	32.5	43.1	3.3	2.3	3.5	4.1
	$\mu \neq 0, \beta \neq 0$	40.4	17.3	44.2	59.7	8.1	5.6	8.8	9.8
ERS _{ρ}	$\mu \neq 0, \beta = 0$	49.4	41.7	49.4	57.0	6.2	8.2	6.0	4.5
	$\mu \neq 0, \beta \neq 0$	46.5	36.7	45.6	57.3	9.3	11.1	8.7	8.1
GM	$\mu \neq 0, \beta = 0$	77.9	75.9	78.2	79.6	9.6	17.0	8.7	3.2
PP _{ρ}	$\mu \neq 0, \beta = 0$	85.0	87.4	84.9	82.8	15.3	15.7	16.0	14.4
	$\mu \neq 0, \beta \neq 0$	97.0	96.5	98.0	96.6	27.5	26.1	29.1	27.2
PP _{τ}	$\mu \neq 0, \beta = 0$	85.9	87.9	85.8	84.1	28.5	27.8	29.4	28.3
	$\mu \neq 0, \beta \neq 0$	97.3	96.7	98.3	97.0	38.6	36.6	40.5	38.7
ZA	$\mu \neq 0, \beta = 0$	97.2	94.9	98.2	98.5	42.7	36.1	44.8	47.3
	$\mu = 0, \beta \neq 0$	97.4	94.7	98.3	99.0	45.6	37.6	47.2	52.0
	$\mu \neq 0, \beta \neq 0$	97.2	94.7	98.0	98.8	45.7	37.6	47.3	52.3
KPSS	$\mu \neq 0, \beta = 0$	23.4	22.0	22.4	25.9	83.4	76.4	82.5	91.3
	$\mu \neq 0, \beta \neq 0$	11.9	7.5	12.4	15.8	70.0	50.5	70.1	89.3

Table 2 contains the rejection rates for the seasonal unit root tests, where we restrict ourselves to the use of the joint F -test for all roots for the HEGY tests and to the t -statistic for the seasonal unit root for the OCSB tests. All tests show severe size distortions except for the GM test when being applied to the raw time series. The HEGY tests are especially biased as they reject a true unit root null hypothesis in almost all cases, in particular for the longer series. On the contrary, almost all tests display an acceptable power of more than 80%, especially for the 10- and 20-year series, with the exception of the ADHF test with seasonal dummies when being applied to the raw data.

Table 3 shows the overall misclassification rates of the non-seasonal and seasonal unit root tests. Whilst especially the PP _{ρ} and the GM tests perform well in comparison to some of the other tests, they still misclassify about 15% of the time series. This clearly leaves room for improvement by combination of tests, even more so as no individual test dominates the others. For detecting the absence or presence of seasonal unit roots, the GM test—in both variants—clearly outperforms the other tests considered here.

We now seek to find a combination of unit root tests that has lower misclassification rates than any individual test. To this end, we run the CRFE algorithm with 50 conditional RFs in each round. The use of conditional RFs is suggested by the empirical cross-correlations between the p -values of the individual tests, as those range from -0.80 to 0.91 for the non-seasonal unit root tests and from -0.40 to 0.97 for the seasonal unit root tests. To reduce the computational burden,

Table 2 Rejection rates for seasonal unit root tests (%; $\alpha = 0.01$, $\Delta = \text{differenced series}$)

Test	Model (8)	$D = 0$			$D = 1$				
		All	5-year	10-year	20-year	All	5-year	10-year	20-year
ADHF τ	$\mu = 0, \beta = 0, \gamma = \mathbf{0}$	85.2	82.2	86.0	86.7	58.0	54.3	59.7	59.9
	$\mu = 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	96.0	94.5	96.8	96.8	72.0	62.9	74.4	78.6
	$\mu = 0, \beta = 0, \gamma \neq \mathbf{0}$	64.2	50.4	69.6	72.7	45.7	36.6	49.6	51.1
	$\mu = 0, \beta = 0, \gamma \neq \mathbf{0}, \Delta$	82.6	60.8	91.9	95.1	68.2	51.4	76.9	76.2
	$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}$	95.6	96.4	96.2	95.1	9.9	23.4	5.2	0.9
	$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	96.2	96.8	96.2	95.7	19.3	36.7	17.6	3.5
HEGY	$\mu \neq 0, \beta \neq 0, \gamma \neq \mathbf{0}$	93.8	81.6	99.8	99.9	89.1	72.8	96.6	98.0
	$\mu \neq 0, \beta \neq 0, \gamma \neq \mathbf{0}, \Delta$	91.8	76.1	99.5	99.9	87.2	74.2	92.9	94.3
	$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}$	99.4	98.2	99.9	100.0	83.9	69.7	89.0	93.1
	$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	98.5	95.8	99.9	100.0	87.9	77.0	91.6	95.0
	$\mu \neq 0, \beta = 0, \gamma \neq \mathbf{0}$	95.0	85.2	99.8	100.0	91.6	88.7	97.6	98.3
	$\mu \neq 0, \beta = 0, \gamma \neq \mathbf{0}, \Delta$	92.8	78.8	99.6	99.9	87.8	76.3	92.9	94.3
OCSB	$\mu \neq 0, \beta \neq 0, \gamma = \mathbf{0}$	99.0	97.1	99.9	99.9	82.4	64.3	87.9	94.9
	$\mu \neq 0, \beta \neq 0, \gamma = \mathbf{0}, \Delta$	99.0	97.1	99.9	99.9	82.4	64.3	87.9	94.9
	$\mu = 0, \beta = 0, \gamma = \mathbf{0}$	92.2	85.3	94.5	96.9	60.1	46.1	62.5	71.6
	$\mu = 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	84.6	71.9	87.5	94.3	46.6	32.5	47.8	59.5

Table 3 Misclassification rates (MCR) of unit root tests (%; $\alpha = 0.01$)

Non-seasonal unit root tests			Seasonal unit root tests		
Test	Model (3)	MCR	Test	Model (8)	MCR
ADF _{τ}	$\mu = 0, \beta = 0$	46.7	ADHF _{τ}	$\mu = 0, \beta = 0, \gamma = \mathbf{0}$	36.4
	$\mu \neq 0, \beta = 0$	23.4		$\mu = 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	38.0
	$\mu \neq 0, \beta \neq 0$	21.9		$\mu = 0, \beta = 0, \gamma \neq \mathbf{0}$	40.8
ERS _{τ}	$\mu \neq 0, \beta = 0$	35.8	GM	$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}$	42.8
	$\mu \neq 0, \beta \neq 0$	33.8		$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	7.1
ERS _{ρ}	$\mu \neq 0, \beta = 0$	28.4	HEGY	$\mu \neq 0, \beta = 0, \gamma \neq \mathbf{0}$	11.5
	$\mu \neq 0, \beta \neq 0$	31.4		$\mu \neq 0, \beta \neq 0, \gamma \neq \mathbf{0}$	47.7
GM	$\mu \neq 0, \beta = 0$	15.8		$\mu \neq 0, \beta \neq 0, \gamma \neq \mathbf{0}, \Delta$	47.7
PP _{ρ}	$\mu \neq 0, \beta = 0$	15.2		$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}$	42.3
	$\mu \neq 0, \beta \neq 0$	15.2		$\mu \neq 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	44.7
PP _{τ}	$\mu \neq 0, \beta = 0$	21.3		$\mu \neq 0, \beta = 0, \gamma \neq \mathbf{0}$	48.3
	$\mu \neq 0, \beta \neq 0$	20.7		$\mu \neq 0, \beta = 0, \gamma \neq \mathbf{0}, \Delta$	47.5
ZA	$\mu \neq 0, \beta = 0$	22.8	OCSB	$\mu \neq 0, \beta \neq 0, \gamma = \mathbf{0}$	41.7
	$\mu = 0, \beta \neq 0$	24.1		$\mu \neq 0, \beta \neq 0, \gamma = \mathbf{0}, \Delta$	41.7
	$\mu \neq 0, \beta \neq 0$	24.3		$\mu = 0, \beta = 0, \gamma = \mathbf{0}$	33.9
KPSS	$\mu \neq 0, \beta = 0$	20.0		$\mu = 0, \beta = 0, \gamma = \mathbf{0}, \Delta$	31.0
	$\mu \neq 0, \beta \neq 0$	21.0			

the initial set of tests only includes tests with misclassification rates less than 50% for $d = 0$ and $d = 1$ in the non-seasonal case and for $D = 0$ and $D = 1$ in the seasonal case, resulting in a set of 12 non-seasonal and 4 seasonal unit root tests. Also, the conditional RF in the i -th round is grown on N_i time series, where $N_i = \max\{1000 - (k_i - 1) \times 100, 200\}$ with k_i being the number of tests considered in the i -th round.

In general, due to the random selection of predictors, unit root tests that do not contribute to, or even worsen, the performance are eliminated during the early rounds. The overall misclassification rate is reduced by leaving out these tests, and, in most rounds, the standard deviation of these rates, calculated over the 50 conditional RFs, is decreased. Eventually, the most important tests remain in the set and eliminating more tests increases the overall misclassification rate. As this lower bound is reached, the optimal set of tests is identified.

Table 4 reports the CRFE results for non-seasonal unit root tests, indicating that a minimum misclassification rate of almost 10% is achieved when four tests remain in the set: the PP _{ρ} test with a trend, the GM algorithm and the KPSS tests with a constant and a trend. Thus, the classification performance is increased by more than 5 percentage points compared to the best individual non-seasonal unit root test (PP _{ρ} tests and GM algorithm; see Table 3). Table 5 reports the CRFE results for the seasonal unit root tests. The initial set of tests already achieves a misclassification rate of almost 6%. The CRFE approach does not improve this initial benchmark, but

Table 4 Recursive feature elimination for non-seasonal unit root tests

Round i	Elimination after round i		Misclassification rates		N_i
	Test	Model (3)	Mean	SD	
1	ADF $_{\tau}$	$\mu \neq 0, \beta \neq 0$	11.1	0.8	200
2	PP $_{\tau}$	$\mu \neq 0, \beta = 0$	11.3	0.8	200
3	ZA	$\mu \neq 0, \beta \neq 0$	11.1	0.7	200
4	ZA	$\mu = 0, \beta \neq 0$	11.3	0.9	200
5	ADF $_{\tau}$	$\mu \neq 0, \beta = 0$	10.7	0.7	300
6	ZA	$\mu \neq 0, \beta = 0$	10.4	0.5	400
7	PP $_{\rho}$	$\mu \neq 0, \beta = 0$	10.4	0.6	500
8	PP $_{\tau}$	$\mu \neq 0, \beta \neq 0$	10.1	0.6	600
9	KPSS	$\mu \neq 0, \beta = 0$	10.1	0.5	700
10	KPSS	$\mu \neq 0, \beta \neq 0$	10.4	0.5	800
11	GM	$\mu \neq 0, \beta = 0$	12.3	1.2	900
12	PP $_{\rho}$	$\mu \neq 0, \beta \neq 0$	12.3	0.5	1000

Table 5 Recursive feature elimination for seasonal unit root tests

Round i	Elimination after round i		Misclassification rates		N_i
	Test	Model (8)	Mean	SD	
1	ADHF $_{\tau}$	$\mu = 0, \beta = 0, \gamma \neq 0$	6.4	0.3	700
2	OCSB	$\mu = 0, \beta = 0, \gamma = 0, \Delta$	6.5	0.4	800
3	GM	$\mu \neq 0, \beta = 0, \gamma = 0, \Delta$	6.6	0.4	900
4	GM	$\mu \neq 0, \beta = 0, \gamma = 0$	7.3	0.2	1000

still we are able to improve about 1 percentage point over the best individual test by combining only four tests.

6 Summary

We set out to evaluate the size and power of a large set of unit root tests and their capabilities of correctly identifying non-seasonal and seasonal unit roots. To this end, we simulate time series that are representative of the economic time series in the M4 competition. Furthermore, we employ a conditional random forest-based elimination algorithm to assess which combination of tests decreases the misclassification rates the most.

The best individual unit root tests misclassifies the absence or presence of a non-seasonal unit root in more than 15% of all cases. By combining the p -values of not more than eight unit root tests, the misclassification rate can be reduced to almost 10%. In the case of seasonal unit roots, the reduction is less pronounced. Still, a combination of tests slightly improves the performance of the best individual test.

For future research, we aim to devise an overall unit root test, possibly in combination with a random forest-based overall seasonality test [46]. In this regard, we could also consider a more nuanced lag-length selection for the augmented tests studied here as well as additional unit root tests in order to improve the design of our approach. This strategy could then be compared with recent bootstrap approaches to unit root testing [21, 33, 34, 39, 45]. Future research could also aim to gain further insights into the theoretical properties of our approach. Some results on the consistency of random forests are already available [3], but, to the best of our knowledge, those have not been demonstrated for conditional random forests and the CRFE algorithm so far.

References

1. Ayat, L., Burridge, P.: Unit root tests in the presence of uncertainty about the non-stochastic trend. *J. Econometrics* **95**, 71–96 (2000)
2. Beaulieu, J.J., Miron, J.A.: Seasonal unit roots in aggregate U.S. data. *J. Econometrics* **55**, 305–328 (1993)
3. Biau, G., Devroye, L., Lugosi, G.: Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9**, 2015–2033 (2008)
4. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
5. Cario, M.C., Nelson, B.L.: Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois, 1–19 (1997)
6. Clements, M.P., Hendry, D.F.: Forecasting with difference-stationary and trend-stationary models. *Economet. J.* **4**, S1–S19 (2001)
7. Cochrane, J.H.: A critique of the application of unit root tests. *J. Econ. Dyn. Control* **15**, 275–284 (1991)
8. DeJong, D.N., Nankervis, J.C., Savin, N.E., Whiteman, C.H.: Integration versus trend stationarity in time series. *Econometrica* **60**, 423–433 (1992)
9. DeJong, D.N., Nankervis, J.C., Savin, N.E., Whiteman, C.H.: The power problems in time series with autoregressive errors. *J. Econometrics* **53**, 323–343 (1992)
10. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**, 427–431 (1979)
11. Dickey, D.A., Fuller, W.A.: Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49**, 1057–1072 (1981)
12. Dickey, D.A., Hasza, D.P., Fuller, W.A.: Testing for unit roots in seasonal time series. *J. Am. Stat. Assoc.* **79**, 355–367 (1984)
13. Diebold, F.X., Rudebusch, G.D.: On the Dickey-Fuller tests against fractional alternatives. *Econ. Lett.* **35**, 155–160 (1991)
14. Elliott, G., Müller, U.K.: Minimizing the impact of the initial condition on testing for unit roots. *J. Econometrics* **135**, 285–310 (2006)
15. Elliott, G., Rothenberg, T.J., Stock, J.H.: Efficient tests for an autoregressive unit root. *Econometrica* **64**, 813–836 (1996)
16. Ernst, P.A., Shepp, L.A., Wyner, A.J.: Yule’s “nonsense correlation” solved!. *Ann. Stat.* **45**, 1789–1809 (2017)
17. Franses, P.H.: Seasonality, non-stationarity and the forecasting of monthly time series. *Int. J. Forecasting* **7**, 199–208 (1991)
18. Fuller, W.A.: Introduction to Statistical Time Series. Wiley, New York (1976)

19. Gómez, V., Maravall, A.: Automatic Modeling Methods for Univariate Series. In: Peña, D., Tiao, G.C., Tsay, R.S. (eds.) *A Course in Time Series Analysis*, pp. 171–201. Wiley, New York (2001)
20. Granger, C.W.J., Newbold, P.: Spurious regressions in econometrics. *J. Econometrics* **2**, 111–120 (1974)
21. Hansen, B.E., Racine, J.S.: Bootstrap model averaging unit root inference. McMaster University, Department of Economics Working Paper No. 2018-09 (2018)
22. Harvey, D.I., Leybourne, S.J., Taylor, A.M.R.: Unit root testing in practice: dealing with uncertainty over the trend and initial condition. *Economet. Theor.* **25**, 587–636 (2009)
23. Hassler, U., Hosseinkouchack, M.: Understanding nonsense correlation between (independent) random walks in finite samples. *Stat. Pap.* **63**, 181–195 (2022)
24. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* **15**, 651–674 (2006)
25. Hylleberg, S., Engle, R.F., Granger, C.W.J., Yoo, B.S.: Seasonal integration and cointegration. *J. Econometrics* **44**, 215–238 (1990)
26. Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root. *J. Econometrics* **54**, 159–178 (1992)
27. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecasting* **34**, 802–808 (2018)
28. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecasting* **36**, 54–74 (2020)
29. Müller, U.K., Elliott, G.: Tests for unit roots and the initial condition. *Econometrica* **71**, 1269–1286 (2003)
30. Ng, S., Perron, P.: Lag length selection and the construction of unit root tests with good size and power. *Econometrica* **69**, 1519–1554 (2001)
31. Ollech, D., Webel, K.: A random forest-based approach to combining and ranking seasonality tests. *J. Economet. Method.* forthcoming (2022)
32. Osborn, D.R., Chui, A.P.L., Smith, J.P., Birchenhall, C.R.: Seasonality and the order of intergration for consumption. *Oxford B. Econ. Stat.* **50**, 361–377 (1988)
33. Palm, F.C., Smeekes, S., Urbain, J.-P.: Bootstrap unit-root tests: comparison and extensions. *J. Time Ser. Anal.* **29**, 371–401 (2008)
34. Park, J.Y.: Bootstrap unit root tests. *Econometrica* **71**, 1845–1895 (2003)
35. Perron, P.: Trends and random walks in macroeconomic time series – Further evidence from a new approach. *J. Econ. Dyn. Control* **12**, 297–332 (1988)
36. Phillips, P.C.B.: Understanding spurious regressions in econometrics. *J. Econometrics* **33**, 311–340 (1986)
37. Phillips, P.C.B.: Time series regression with a unit root. *Econometrica* **55**, 277–301 (1987)
38. Phillips, P.C.B., Perron, P.: Testing for a unit root in time series regression. *Biometrika* **75**, 335–346 (1988)
39. Psaradakis, Z.: Bootstrap tests for an autoregressive unit root in the presence of weakly dependent errors. *J. Time Ser. Anal.* **22**, 577–594 (2001)
40. Rodrigues, P.M.M., Osborn, D.R.: Performance of seasonal unit root tests for monthly data. *J. Appl. Stat.* **26**, 985–1004 (1999)
41. Said, S.E., Dickey, D.A.: Testing for unit roots in autoregressive moving average models of unknown order. *Biometrika* **71**, 599–607 (1984)
42. Schwert, G.W.: Tests for unit roots: a Monte Carlo investigation. *J. Bus. Econ. Stat.* **7**, 147–159 (1989)
43. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. *BMC Bioinformatics* **9**, Article 307 (2008)
44. Stone, C.J., Hansen, M.H., Kooperberg, C., Truong, Y.K.: Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *Ann. Stat.* **25**, 1371–1470 (1997)
45. Swensen, A.R.: Bootstrapping unit root tests for integrated processes. *J. Time Ser. Anal.* **24**, 99–126 (2003)

46. Webel, K., Ollech, D.: An overall seasonality test based on recursive feature elimination in conditional random forests. *Proc. Int. Conf. Time Ser. Forecasting*, 20–31 (2018)
47. West, K.D.: A note on the power of least squares tests for a unit root. *Econ. Lett.* **24**, 249–252 (1987)
48. Wolters, J., Hassler, U.: Unit root testing. *Allg. Stat. Arch.* **90**, 43–58 (2006)
49. Yule, G.U.: Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *J. R. Stat. Soc.* **89**, 1–63 (1926)
50. Zivot, E., Andrews, D.W.K.: Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *J. Bus. Econ. Stat.* **10**, 251–270 (1992)

Probabilistic Forecasting of Seasonal Time Series



Combining Clustering and Classification for Forecasting

Colin Leverger, Thomas Guyet, Simon Malinowski, Vincent Lemaire, Alexis Bondu, Laurence Rozé, Alexandre Termier, and Régis Marguerie

Abstract In this article, we propose a framework for seasonal time series probabilistic forecasting. It aims at forecasting (in a probabilistic way) the whole next season of a time series, rather than only the next value. Probabilistic forecasting consists in forecasting a probability distribution function for each future position. The proposed framework is implemented combining several machine learning techniques (1) to identify typical seasons and (2) to forecast a probability distribution of the next season. This framework is evaluated using a wide range of real seasonal time series. On the one side, we intensively study the alternative combinations of the algorithms composing our framework (clustering, classification), and on the other side, we evaluate the framework forecasting accuracy. As demonstrated by our experiences, the proposed framework outperforms competing approaches by achieving lower forecasting errors.

Keywords Time series · Probabilistic forecasting · Seasonality

C. Leverger
IRISA, Rennes Cedex, France

Orange Labs, Rennes, France

T. Guyet (✉)
Inria Centre, Grenoble, France
e-mail: thomas.guyet@irisa.fr

S. Malinowski · A. Termier
Université Rennes 1/Inria/IRISA, Rennes, France

V. Lemaire · A. Bondu · R. Marguerie
Orange Labs, Rennes, France

L. Rozé
INSA/Inria/IRISA, Rennes, France

1 Introduction

Forecasting the evolution of a temporal process is a critical research topic, with many challenging applications. In this work, we focus on time series forecasting and on data-driven forecasting models. A time series is a timestamped sequence of numerical values, and the goal of forecasting is, at a given point of time, to predict the next values of the time series based on previously observed values and possibly on other linked exogenous variables. Data-driven algorithms are used to predict future time series values from past data, with models that are able to adapt automatically to any type of incoming data. The data science challenge is to learn accurate and reliable forecasting models with as few human interventions as possible. Time series forecasting has many applications in medicine (for instance, to forecast blood glucose of a patient [1]), in economics (for instance, to forecast macroeconomic variable changes [2]), in the financial domain (forecasting financial time series [3]), in electricity load [4] or in industry (for instance, to forecast the server load [5, 6]).

Time series forecasting algorithms provide information about possible situations in the future and can be used to anticipate crucial decisions. Taking correct decisions requires anticipation and accurate forecasts. Unfortunately, these objectives are often contradictory. Indeed, the larger the forecasting horizon, the wider the range of expectable situations. In such case, a probabilistic forecasting algorithm is a powerful decision support tool, because it handles the uncertainty of the predictions. Probabilistic or density forecasting is a class of forecasting that provides intervals or probability distributions as outcomes of the forecasting. It is claimed in [7] that, in recent years, probabilistic forecasts have become widely used. For instance, fan charts [8], highest density regions [9] or functional data analysis [10] enables to forecast ranges for possible values of future data.

We are particularly interested in time series that have some periodic regularities in their values. This kind of time series is said to be seasonal. For instance, time series related to human activities or natural phenomena are often seasonal, because they often exhibit daily regularities (also known as the circadian cycle). Knowing that a time series is seasonal is a valuable information that can help for forecasting. More specifically, learning the seasonal structures can help to generate longer-term predictions as it provides information about several seasons ahead.

Furthermore, seasonality of a time series gives a natural midterm forecasting horizon. Classical forecasting models (e.g. SARIMA [11]) predict the future values of a given time series stepwise. The predicted values are used by further steps. At each step, there is then a risk of the error to be accumulated due to the recursive nature of the forecasts. The prediction of a whole season at once aims at spreading the forecasting error all along the season. Thus, we expect to forecast more accurately the salient part of a season that may lie in the middle of the season. More practically, the prediction of a whole season at once allows applications where such prediction is required to plan actions (e.g. to plan electricity production a day ahead, it is necessary to predict the consumption for the next 24 hours).

A second limitation of usual seasonal forecasting methods is the assumption that the seasons have the same shape, i.e. the values evolve in the same way over the season. The differences with each other are due to noise and an additive constant. Nevertheless, most of the real seasonal time series often contains more than just one periodic pattern. For instance, daily connections to a given website exhibit different patterns for a weekday or for a Sunday, for instance. This kind of structure cannot be well captured by classical forecasting methods.

In this article, we propose a generic framework called P-F2C (which stands for “Probabilistic Forecasting with Clustering and Classification”) for seasonal time series forecasting. This approach extends the F2C framework [6] (which stands for “Forecasting with Clustering and Classification”). P-F2C predicts future values for a complete season ahead at once, and this in a probabilistic manner. The P-F2C predictions may be used for supporting decision-making about the next season, handling the uncertainty in the future through the probabilistic presentation of the result.

2 Probabilistic Seasonal Time Series Forecasting

In this section, we introduce the notations and the problem of seasonal time series forecasting.

2.1 Seasonal Time Series

A time series Y is an ordered sequence of values $y_{0:n} = y_0, \dots, y_{n-1}$, where $\forall i \in [0, n-1]$, $y_i \in \mathbb{R}$ (univariate time series). n denotes the length of the observed time series.

Y is said to be (ideally) *seasonal* with season length s if there exists $\mathcal{S} = \{S^1, \dots, S^p\}$ a finite collection of p sub-series (of length s) called *typical seasons* such that

$$\forall i \in [0, m-1], y_{(s \times i):s \times (i+1)} = \sum_{j=1}^p \sigma_{i,j} S^j + \boldsymbol{\varepsilon}_i \quad (1)$$

where m is the number of seasons in the time series, $\boldsymbol{\varepsilon}_i \in \mathbb{R}^s$ represents a white noise and $\sum_j \sigma_{i,j} = 1$ for all j . In other words, it means that for a seasonal time series Y , every season in Y is a weighted linear combination of typical seasons. Intuitively, this modelling of a typical season corresponds to additive measurements (e.g. consumption or traffic) for which the observed measure at time t is the sum of individual behaviours. In this case, a typical season corresponds to a typical

behaviour of individuals, and the $\sigma_{\cdot,j}$ represents the proportion of individuals of type j contributing to the observed measure.

In the following, $\mathbf{y}_i = y_{(s \times i):(s \times (i+1))} \in \mathbb{R}^s$ denotes the i -th season of Y .

2.2 Seasonal Probabilistic Forecasting

Let $Y = y_0, \dots, y_{n-1}$ be a seasonal time series and s be its season length. Note that the season length of a time series (s) is estimated using Fisher's g -statistics [12]. Without loss of generality, we assume that the length of a time series is a multiple of the season length, i.e. $n = m \times s$. m denotes the number of seasons in the observed time series. The goal of seasonal probabilistic forecasting is to estimate

$$\Pr(y_{n:n+s}^* \mid y_{(n-\gamma \times s):n}) = \Pr(\mathbf{y}_m^* \mid \mathbf{y}_{(m-\gamma):m}) \quad (2)$$

where $\mathbf{y}_m^* = y_{n:n+s}^*$ are the forecasts of the s next values (next season) of the observed time series and $\mathbf{y}_{(m-\gamma):m} = y_{(n-\gamma \times s):n}$ are the observed values of the last γ seasons. γ is a parameter given by the user.

We now propose an equivalent formulation of this problem considering our hypothesis on seasonal time series and we denote $\mathcal{S} = \{S^1, \dots, S^p\}$ the set of p typical seasons. Thus, Eq. 2 can be rewritten as follows:

$$\Pr(\mathbf{y}_m^* \mid \mathbf{y}_{(m-\gamma):m}) = \sum_{S \in \mathcal{S}} \Pr(\mathbf{y}_m^* \mid S) \cdot \Pr(S \mid \mathbf{y}_{(m-\gamma):m}) \quad (3)$$

where $\Pr(\mathbf{y}_m^* \mid S)$ is the probability of having \mathbf{y}_m^* given the type of the next season and $\Pr(S \mid \mathbf{y}_{(m-\gamma):m})$ is the probability that the next season is of type S given past observations.

The problem formulation given by Eq. 3 turns the difficult problem of Eq. 2 into two well-known tasks in time series analysis:

- Estimating the first term, $\Pr(\mathbf{y}_m^* \mid S)$ leads to a problem of time series clustering. The problem is both to define the typical seasons, \mathcal{S} , and to estimate the distributions of the season values. A clustering of the seasons $(\mathbf{y}_i)_{i=0:m}$ of the observed time series identifies the typical seasons (clusters) and gives the required empirical distributions $\hat{\Pr}(\mathbf{y}, S)$.
- Estimating the second $\Pr(S \mid \mathbf{y}_{(m-\gamma):m})$ is a probabilistic time series classification problem. This distribution can be empirically learnt from the past observations $(\mathbf{y}_{i-\gamma:i}, S_i^*)_{i=\gamma:m}$ where S_i^* denotes the empirical type of the i -th season obtained from the clustering assignment above.

This problem formulation and remarks sketch the principles of a probabilistic seasonal time series forecasting. P-F2C is an implementation of these principles with a specific time series clustering.

3 The P-F2C Forecaster

P-F2C is composed of a clusterer that models the latent typical seasons and a classifier that predicts the next season type given the recent data. The forecaster is fit on the historical data of a time series. Then, the forecaster can be applied on the time series to predict the next season(s).

P-F2C clusterer is based on a probabilistic co-clustering model that is presented in the next section. In Sect. 3.2, we present how to use classical classifiers to predict the next seasons.

3.1 Co-clustering of Time Series: A Probabilistic Model

Co-clustering is a particular type of unsupervised algorithm which differs from regular clustering approaches by creating co-clusters. The co-clustering approach consists in simultaneously partitioning the lines and the columns of an input data table. Thus, a co-cluster is defined as a set of examples belonging to both a group of rows and a group of columns. In [13], Boullé proposed an extension of co-clustering to tri-clustering in order to cluster time series. In this approach, a time series with an identifier C is seen as a set of couples (T, V) , where T is a timestamp and V a value of a measurement. Thus, the whole set of time series is a large set of points represented by triples (C, T, V) . The tri-clustering approach handles the three variables (C is categorical and T, V are numerical) to create homogeneous groups. A co-cluster gathers time series (group of identifiers) that have similar values during a certain interval of time. Contrary to the classical clustering approaches (e.g. KMeans, K-shape, GAK) [14] that are based on the entire time series, the co-clustering approach uses a local criterion. This difference is illustrated in Fig. 1: A distance-based clustering (on the left) evaluates the distance between whole time

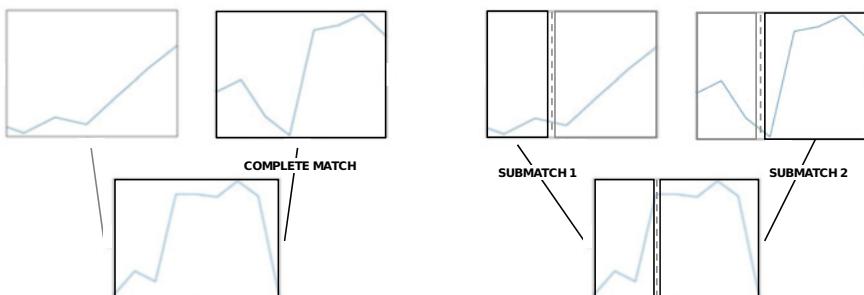


Fig. 1 Difference between clustering (on the left), which matches the entire time series, with co-clustering (on the right), which is able to match subintervals of the time series of various other time series

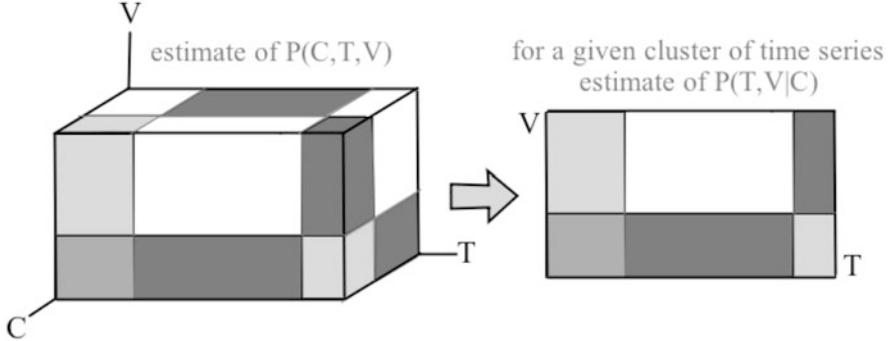


Fig. 2 Illustration of a trivariate co-clustering model where a slice referred to forecasting “grid” is extracted

series; in the co-clustering approaches, the distance is based on subintervals of the seasons. This enables to identify which parts of the season are the most discriminant. Besides, tri-clustering is robust to missing values in time series.

The tri-clustering approach of Boullé is based on the MODL framework [10]. The MODL framework makes a constant piecewise assumption to estimate the joint distribution $\Pr(C, T, V)$ by jointly discretising the variables T , V and grouping the time series identifiers of the variable C . The resulting model consists of the Cartesian product¹ of the three partitions of the variables C , T , V . This model can be represented as a 3D grid (see Fig. 2, on the left). In this 3D grid, if one considers a given group of time series (i.e. a given group of C), the model provides a bivariate discretisation which estimates $\Pr(T, V | C) = \frac{\Pr(C, T, V)}{\Pr(C)}$ as a 2D grid (see Fig. 2, on the right). This 2D grid gives the probability to have a given range of values during a given interval of time. Therefore, knowing that a time series belongs to a given cluster, the corresponding 2D grid may then be used for crafting forecasts (see next section).

In the MODL approach, finding the most probable tri-clustering model is turning into a model selection problem. To do so, a Bayesian approach called maximum a posteriori (MAP) is used to select the most probable model given the data. Details about how this 3D grid model is learned may be found in [13, 15]. The main idea could be summarised as finding the grid which maximises the contrast compared to a grid based on the assumption that T , V and C are independent (i.e. $\Pr(V, T, C) = \Pr(V)\Pr(T)\Pr(C)$). Therefore, the estimation of this MAP model outputs: (i) v intervals of values $V_i = [v_i^l, v_i^u]$ for $i = 1, \dots, v$, (ii) τ intervals of times $T_i = [t_i^l, t_i^u]$ for $i = 1, \dots, \tau$, (iii) groups of time series. These groups of time series correspond to the typical seasons, denoted \mathcal{S} in the above model. $|\mathcal{S}|$ is

¹ The Cartesian product of the three partitions is used as a constant piecewise estimator—i.e. a 3D histogram.

the number of clusters at the finer level that is optimal in the sense of the MODL framework.

In the time series forecasting approach proposed in this paper, the right number of (tri-)clusters is optimised regarding to the forecasting task. More precisely, this number is optimised according to the performance of the model at prediction time, using the validation ensemble. This value could differ from $|\mathcal{S}|$. Therefore, the MODL co-clustering approach allows applying a hierarchical clustering to the finest level to have a coarse level with a lower number of clusters called C^* , $C^* < |\mathcal{S}|$. A grid search selects the C^* value based on the forecast accuracy on the valid dataset.

Let us now come back to the formalisation of probabilistic time series forecasting: $\hat{\Pr}(\mathbf{y}_m^* \mid S)$ is estimated by the MODL model from the conditional probabilities $\Pr(V, T \mid C = S)$ where S denotes one of the time series groups, i.e. a typical season. In practice, the grid is used to estimate the distribution of values at each time point of a season. With MODL, the distribution is modelled by a piecewise constant function. It is worth noting that MODL is a non-parametric approach.

3.2 Predict the Next Type of Seasons

The problem is here to estimate empirically $\Pr(S_{i+1} \mid \mathbf{y}_{(i-\gamma):i})$ the probability of having a type of season $S_{i+1} \in \mathcal{S}$ for the $(i + 1)$ -th season given the observations over the γ past seasons. We consider two different sets of features to represent the γ previous seasons. The first approach consists in having only the time series values $\mathbf{y}_{(i-\gamma):i}$ as features. The second approach uses the time series values and the types of the previous seasons as features.

Then, the next season prediction problem consists in learning a probabilistic classifier (naive Bayes classifier, logistic regression, decision tree or random forests) or a time series classifiers (TSForest [16], Rocket [17]). Note that time series classifiers can use only the time series values.

3.3 Select the Best Parameters (Portfolio)

The P-F2C forecaster is parameterised by the number of seasons in the past (γ) used for learning next season type, a maximum number of typical seasons to detect in a non-supervised way and the type of classifier. The γ parameter is introduced in the problem definition and its choice is left to the user who specifies what is the forecasting task. On the other hand, the other parameters may be difficult to be set by the user, and we do not think that one of the classifiers will outperform the others for all the time series. For these reasons, the portfolio approach (denoted PP-F2C) implements a grid search for the best parameters by splitting the dataset into a training (75%) and a validation dataset (25%) to identify the best value of the parameters. Once the best values have been set, the clusterer and the classifier are fitted on the entire dataset.

4 Illustration on a Synthetic Dataset

This section shows results with synthetic data. The goal is to illustrate the probabilistic grid used in P-F2C method and to give intuitions behind probabilistic forecasting that are provided by P-F2C. We compare the output of P-F2C against the output of DeepAR [18], a state-of-the-art probabilistic time series forecaster.

4.1 The Data Generated

Generating data is a good strategy for checking assumptions before launching experiments at scale. Indeed, the shape of the generated data is often simpler, and completely controlled. Experiments may be executed with various parameters, to plot understandable results and to validate basic expectations.

The seasonal data generated for this section follows some well-established seasonal sequences. Three different time series patterns are defined for three different latent types of season of length 10. In Fig. 3, one type of season (s_1 in dashed-line orange) with always increasing values is observed, one type of season (s_2 in dotted-line green) with two peaks is observed, etc. Those three different types of season are then repeated 50 times in a defined order ($s_1, s_1, s_2, s_0, s_1, s_1, s_0, s_2$, as observed in Fig. 3, on the right, which shows the entire sequence that is being repeated), and noise is added to the final time series to make the forecasting process less straightforward.

4.2 Grid Probabilistic Forecasts

Once trained, we apply the P-F2C forecaster after the penultimate season of the time series illustrated in Fig. 3 (at time 70). Knowing the sequence of patterns, we can guess that a season of type s_2 is coming ahead. Indeed, the last three patterns seems

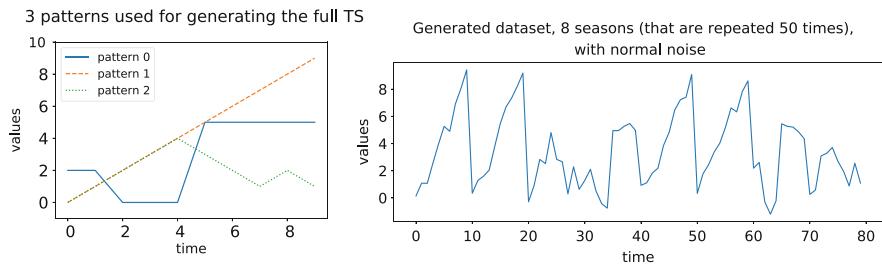


Fig. 3 On the left: typical seasons of length 10 used for generating the time series. On the right: examples of generated time series with white noise (8 seasons)

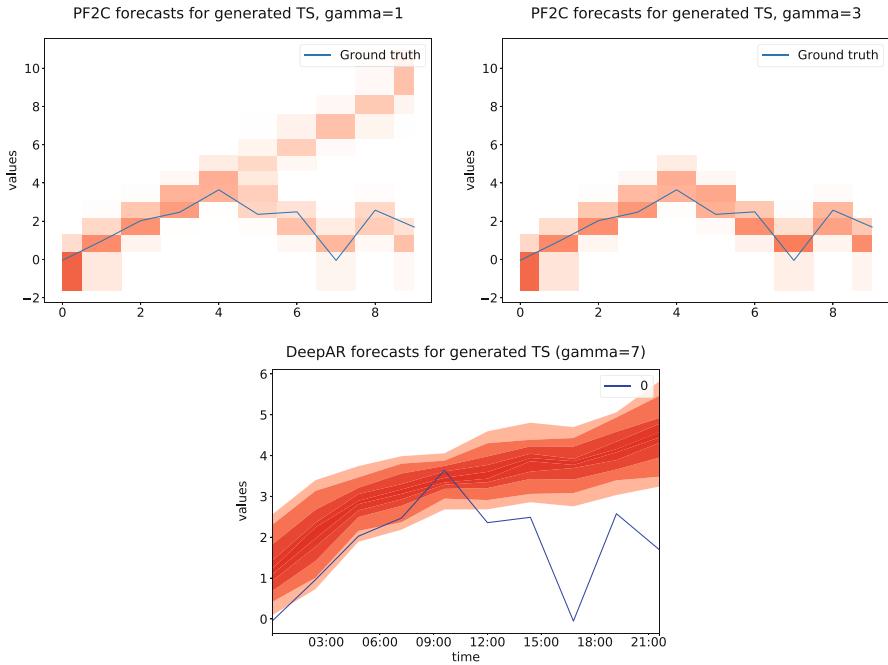


Fig. 4 One season ahead grid forecasts for the generated time series with $\gamma = 1$ at the top left and $\gamma = 3$ at the top right, and DeepAR at the bottom

to follow the sequence $[s_1, s_1, s_0]$; thus, the next type of season can be deduced from the known patterns.

Figure 4 shows two examples of forecasts with different values of γ .

The real values of the predicted time series are in blue (noisy version of the s_2 pattern). The probabilistic forecasts are shown in a red overlay. It is a set of rectangles that visualise the homogeneous regions that have been identified by MODL co-clustering. The darker the red, the more probable next season ahead lay in this (T, V) interval.

Figure 4 top left is the forecast obtained with $\gamma = 1$. It illustrates a probabilistic forecast with uncertainty. Indeed, light red cells are observed in the figure where the data are predicted to lay (with a low probability). In this case, the classifier is unable to predict accurately the next type of season. With $\gamma = 1$ the classifier has only the information of the preceding season (of type s_0). In this case, the forecaster encountered two types of season after a s_0 season: s_1 or s_2 with the same probability. Then, the predicted grid is a mixture of the two types of grids. For the first half of the season, the forecast is confident in predicting the linear increase of the value (darker red cells), but for the second half, the forecast suggests two possible behaviours: continue the linear increase (s_1) or a decrease (s_2). Note that the grids of all typical seasons share the same squaring. MODL necessarily creates the same cuttings of a dimension (V or T) along the others (C).

Figure 4 top right is the forecast obtained with $\gamma = 3$. It illustrates a good probabilistic forecast. The real values (in blue) often appear in the red boxes where the red is very dark. It means that the season type was both well described by MODL and well predicted by the classifier. In this case, a larger memory of the forecaster disentangles the two possible choices it had above. After a $[s_1, s_1, s_0]$, the forecaster always observed seasons of type s_2 . Thus, the grid of this pattern is predicted.

It is worth noting that, for $\gamma = 1$, the use of the MODL probabilistic grid suggests two distinct possible evolutions of the time series, but there is an uncertainty on which evolution will actually occur. In the classical probabilistic forecasts, probabilities are distributed around a mean time series. This is illustrated in Fig. 4 at the bottom with DeepAR using the seven seasons in the past to predict the next season. On the second half of the season, the predicted probabilistic distribution suggests a behaviour in between s_1 and s_2 with a larger uncertainty. Such model makes confusion between uncertainty of behaviour and imprecise forecast. In the case of seasonal time series with different types of season, the mean time series has no meaning for an analyst.

5 Experiments

This section presents experiments to assess the accuracy of P-F2C. We start by introducing the experimental settings, then we investigate some parameters of our model and finally we present the result of an intensive comparison of P-F2C to competitors.

5.1 Experimental Protocol

The framework has been developed in Python 3.5. The MODL co-clustering is performed by the Khiops tool [19]. The classification algorithms are borrowed from the *sklearn* library [20].

In our experiments, we used a dataset made of 36 time series (see Annexes), from various sources and nature: technical devices, human activities, electricity consumption, natural processes, etc. All these datasets have been selected because seasonality was identified and validated with a Fisher g -test [12]. Each time series is normalised using a z -normalisation prior to data splitting, in order to have comparable results. For the experiments, 90% of the time series are used to train the forecaster (this train test is internally split in training and valid datasets), and 10% of the original time series are used to evaluate the accuracy.

P-F2C and PP-F2C are compared with classical deterministic time series forecasters (ARIMA, SARIMA, HoltWinters), with LSTM [21], with Prophet [22] and with the F2C method [6] which uses the principles as P-F2C but with K-means clustering algorithm and random forest classifiers to learn the structure in

the season sequence. P-F2C being a probabilistic methodology, we also compare it with DeepAR [18].

We use mean absolute error (MAE) and continuous ranked probability score (CRPS) to compare the forecasts to the real time series. The MAE is dedicated to deterministic forecasts, while CRPS is to probabilistic ones. It is worth noting that the CRPS is analogous to MAE for deterministic forecasts. Therefore, comparing MAE measure for deterministic forecasts against CRPS values for probabilistic forecasts is technically sound [23]. The CRPS is used for DeepAR and P-F2C. All the other approaches forecast crisp time series and their accuracy is evaluated through MAE. For each experiment, we illustrate the results with critical difference diagrams. A critical difference diagram represents the mean rank of the methods that have been obtained on the set of the 36 times series. The lower the better. In addition, the representation shows horizontal bars that group some methods. In a same group, the methods are not statistically different according to the Nemenyi test.

5.2 Parameters' Sensitivity

In this section, an analysis of the alternative settings of the P-F2C methodology is conducted. We investigate the effect of two choices: the choice of the γ value, i.e. the number of seasons to consider in the history, and the choice of the classifier to predict the next type of season in case we do not use the portfolio optimisation.

Figure 5 on the left shows a critical diagram that compares the ranking of P-F2C with different values of γ (1, 2 or 3). For this experiment, the classifier is the RandomForestClassifier (and we had the same results with the other classifiers). We notice that the larger γ , the lower the error. Indeed, as seen in Sect. 4, larger γ improves the accuracy of the forecast of the next season type. Nonetheless, we observed that for some time series, lower γ may be better. We explain this counter-intuitive results by the small length of some of the time series. In these cases, the number of seasons in the training set is too small to fit the numerous parameters of a classifier with $\gamma \times s$ features.

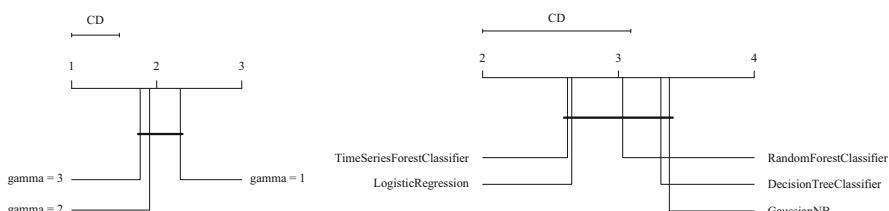


Fig. 5 Critical diagrams used to find the best parameters for the P-F2C implementation

Figure 5 on the right shows a critical diagram that compares the classifiers used to predict the next type of season. It shows that time series forest classifier [16] is on average in first position. This classifier has been designed specifically for time series classification; it explains why it outperforms the other approaches. Nonetheless, the differences with logistic regression and random forest are not statistically significant. Their capability to use extra information, such as the type of seasons, may be an interesting advantage to improve performances.

5.3 P-F2C and PP-F2C vs Opponents

The critical diagram of Fig. 6 compares the performances of the methods. P-F2C denotes our approach configured with the best parameters on average found in Sect. 5.2. PP-F2C denotes P-F2C that is optimised on the valid test for each dataset (portfolio). It shows that rank-wise, the seasonal forecaster F2C, P-F2C and PP-F2C are performing better than the others. We can first notice that the portfolio actually improves the performances of P-F2C. Nonetheless, the non-probabilistic approach outperforms PP-F2C.

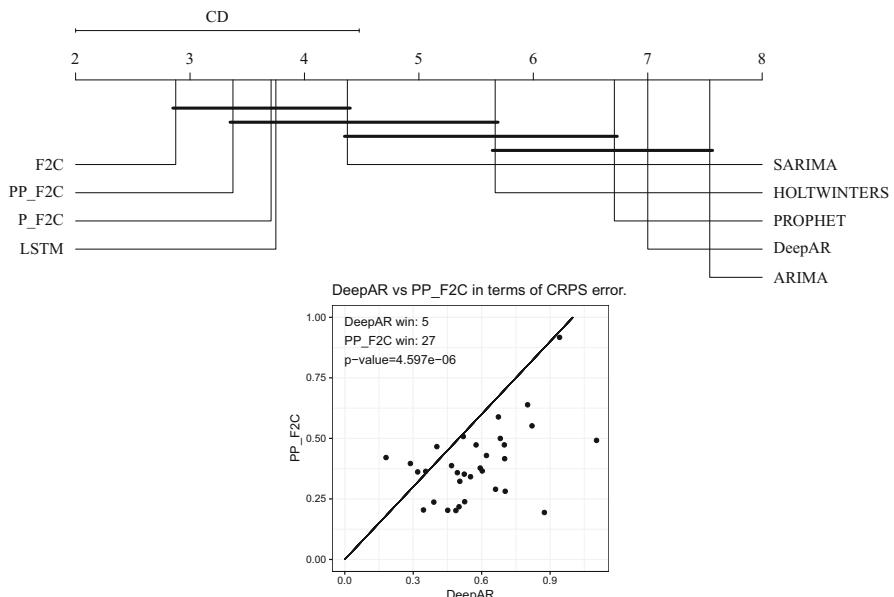


Fig. 6 At the top: critical diagram of the comparison between different prediction approaches (acronyms of method are detailed in the text). At the bottom: win/tie/lose graph between PP-F2C and DeepAR

We also notice that F2C outperforms PP-F2C. Even if a PP-F2C forecast fits the time series (see Fig. 4), the piece-wise approximation generates a spread of the probabilistic distribution that penalises the CRPS. Nonetheless, it is worth noting that the rank difference with F2C is not statistically significant and that a probabilistic forecast provides meaningful information which increase the trustworthy in the forecasts.

Then, we compared PP-F2C with another probabilistic forecaster (DeepAR). The critical diagram of Fig. 6 shows that PP-F2C outperforms DeepAR significantly ($p < 10^{-6}$). The win/tie/lose graph at the bottom shows how many times PF2C won against DeepAR (points below the diagonal) and the relative values of CRPS. The point positions illustrate that PP-F2C outperforms DeepAR significantly on most of the datasets.

6 Conclusion

P-F2C is a probabilistic forecaster for seasonal time series. It assumes that seasons are a mixture of typical seasons to transform the forecasting problem into both a clustering and a classification of time series. The P-F2C applies parameterless co-clustering approach that generates grid forecasts, each typical grid being a typical seasonal behaviour. In addition, we proposed PP-F2C that adjusts P-F2C parameters for each time series. PP-F2C outperforms on average the competitors except F2C on various seasonal time series. F2C is based on the same principle as PP-F2C but is not probabilistic and parameterless. Nonetheless, we have shown the value of probabilistic grid forecasting to give information about uncertain distinct mean behaviours. Indeed, the probabilistic grid mixture is more interpretable than combining probabilistic distribution around a mean.

Annexes

Table 1 gives the detailed description of the dataset used in the experiments.

Table 1 Datasets used for experimentations

Dataset	Origin	Acquisition freq.	No. pts/seas	No. seas
Monthly beer production Australia megalitres	kaggle (https://www.kaggle.com/mpwolke/australian-monthly-beer-production)	1 month	12	39
Electricity production	kaggle (https://www.kaggle.com/robikscube/hourly-energy-consumption)	1 h	12	32
Daily maximum temperatures in Melbourne, Australia, 1981–1990	hdrcede (https://pkg.robjhyndman.com/hdrcede/reference/maxtemp.html)	1 day	7	470
Internet traffic data I from Jun. 7, 2005 to Jul. 31, 2005	tsdl [24]	1 h	24	51
Internet traffic data II from Nov. 19, 2004 to Jan. 27, 2005	tsdl [24]	1 h	24	69
Monthly sunspot Zuerich	R dataset [25]	1 month	12	235
Mon pax web	Adelaide Airport Aircraft Movements (source: Australian Bureau of Infrastructure, Transport and Regional Economics: (https://data.gov.au/data/dataset/airport-traffic-data))	Monthly	12	114
Currency	kaggle (https://www.kaggle.com/kashnitsky/topic-9-part-1-time-series-analysis-in-python)	1 day	3	100
Weather Canada	kaggle (https://www.kaggle.com/selfishgene/historical-hourly-weather-data?select=temperature.csv)	1 h	24	57
Enedis	Electricity consumption (source: Enedis (https://data.enedis.fr))	30 min	48	143
Traffic	New York Traffic Volume (source: New York Metropolitan Transportation Council (https://opendata.cityofnewyork.us/))	1 hour	24	106
Bidmc	Electrocardiogram (ECG) (source: Physionet, https://physionet.org/content/bidmc/1.0.0/)	125 Hz	45	69
Rossman Sales	kaggle (https://www.kaggle.com/c/rossmann-store-sales)	1 day	7	82
311SF	Number of calls for ‘Graffitis’ cases to the San Francisco call center (source: SF Open data, https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6)	1 h	24	75

(continued)

Table 1 (continued)

Dataset	Origin	Acquisition freq.	No. pts/seas	No. seas
Tide	San Francisco sea level (source: NOAA https://coastwatch.pfeg.noaa.gov/erddap/)	6 min	124	112
Pedestrian Counting System	City of Melbourne [26]	1 h	24	1490
Orange Hits per quarter	Orange	15 min	96	512
CO.GT	Air quality indicators from Mar. 10, 2004 to Apr. 04 2005 in an Italian city [27]	1 h	12	250
PT08.S1.CO	Air quality indicators [27]	1 h	12	250
C6H6.GT	Air quality indicators [27]	1 h	12	250
PT08.S2.NMHC	Air quality indicators [27]	1 h	12	250
NOx.GT	Air quality indicators [27]	1 h	12	250
RH	Air quality indicators [27]	1 h	24	125
Global horizontal radiation	Solar radiation monitoring from Apr. 25, 2016 to Aug. 25, 2016 [28]	1 h	14	214
Direct normal radiation	Solar radiation [28]	1 h	14	214
Diffuse horizontal radiation	Solar radiation [28]	1 h	14	214
Amial	Porto water consumption from different locations in the city of Porto from Nov. 11, 2015 to Jan. 11, 2016 [28]	30 min	48	62
Preciosa mar	water consumption [28]	30 min	48	62
Humidity	Bike sharing from Jan. 1, 2011 [28]	1 h	23	58
No. of Births in Quebec from Jan. 1, 1977 to Dec. 31, 1990	tsdl [24]	1 day	7	428
Electricity total load	Hospital energy loads from Jan. 1, 2016 to Mar. 25, 2016 [28]	1 h	24	125
Electricity Total demand	Hospital energy loads [28]	minutes	48	59
Equipment load	Hospital energy loads [28]	1 h	24	125
Gas energy	Hospital energy loads [28]	1 h	24	125
Gas heat energy	Hospital energy loads [28]	1 h	24	125
Water heater Energy	Hospital energy loads [28]	1 h	24	125

References

1. Liu, C., Vehí, J., Avari, P., Reddy, M., Oliver, N., Georgiou, P., Herrero, P.: Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal. *Sensors* **19**(19), 4338 (2019)
2. Li, J., Chen, W.: Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *Int. J. Forecast.* **30**(4), 996–1015 (2014)
3. Tay, F., Cao, L.: Application of support vector machines in financial time series forecasting. *Omega* **29**(4), 309–317 (2001)
4. Laurinec, P., Lóderer, M., Lucká, M., Rozinajová, V.: Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption. *J. Intell. Inf. Syst.* **53**(2), 219–239 (2019)
5. Bodik, P.: Automating Datacenter Operations Using Machine Learning. PhD thesis, UC Berkeley (2010)
6. Leverger, C., Malinowski, S., Guyet, T., Lemaire, V., Bondu, A., Termier, A.: Toward a framework for seasonal time series forecasting using clustering. In: Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, pp. 328–340 (2019)
7. De Gooijer, J., Hyndman, R.: 25 years of time series forecasting. *Int. J. Forecast.* **22**(3), 443–473 (2006)
8. Wallis, K.F.: Asymmetric density forecasts of inflation and the bank of england’s fan chart. *Natl. Inst. Econ. Rev.* **167**(1), 106–112 (1999)
9. Hyndman, R.: Highest-density forecast regions for nonlinear and non-normal time series models. *J. Forecast.* **14**(5), 431–441 (1995)
10. Boullé, M.: Data grid models for preparation and modeling in supervised learning. *Hands On Pattern Recognit. Chall. Mach. Learn.* **1**, 99–130 (2011)
11. Kareem, Y., Majeed, A.R.: Monthly peak-load demand forecasting for sulaimany governorate using SARIMA. In: Proceedings of the International Conference on Transmission & Distribution Conference and Exposition, pp. 1–5 (2006)
12. Wichert, S., Fokianos, K., Strimmer, K.: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* **20**(1), 5–20 (2004)
13. Boullé, M.: Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition* **45**(12), 4389–4401 (2012)
14. Paparrizos, J., Gravano, L.: Fast and accurate time-series clustering. *ACM Trans. Database Syst. (TODS)* **42**(2), 1–49 (2017)
15. Bondu, A., Boullé, M., Cornuéjols, A.: Symbolic representation of time series: A hierarchical coclustering formalization. In: International Workshop on Advanced Analysis and Learning on Temporal Data, pp. 3–16. Springer (2015)
16. Deng, H., Runger, G., Tuv, E., Vladimir, M.: A time series forest for classification and feature extraction. *Information Sciences* **239**, 142–153 (2013)
17. Dempster, A., Petitjean, F., Webb, G.I.: Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels (2019). arXiv:1910.13051
18. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: Deepar: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36**(3), 1181–1191 (2020)
19. Boullé, M.: Khiops: Outil d’apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables. In: Actes de la conférence Extraction et Gestion des Connaissances, pp. 505–510 (2016)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
21. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. In: Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN), pp. 850–855 (1999)

22. Taylor, S., Letham, B.: Forecasting at scale. *Am. Stat.* **72**(1), 37–45 (2018)
23. Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**(5), 559–570 (2000)
24. Hyndman, R.: Time Series Data Library (TSDL) (2011)
25. Andrews, D.F., Herzberg, A.M.: *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer Science & Business Media (2012)
26. Melbourne, C.O.: Pedestrian Counting System (2016)
27. Asuncion, A., Newman, D.: Uci Machine Learning Repository (2007)
28. Cerqueira, V., Torgo, L., Pinto, F., Soares, C.: Arbitrated ensemble for time series forecasting. In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 478–494 (2017)

Nonstatistical Methods for Analysis, Forecasting, and Mining Time Series



Vilém Novák and Irina Perfilieva

Abstract This is an overview paper, in which we briefly present results obtained over several years in the analysis, forecasting, and mining information from time series using methods that predominantly have nonstatistical character. Our main goal is to show the readers from the area of probability theory and statistics that nonstatistical methods can be pretty successful in time series processing. Besides the standard tasks such as estimation of trend/trend-cycle and forecasting, our methods are also powerful in providing additional information that can hardly be obtained using the statistical methods, namely, evaluation of the local course, finding perceptually important points, identification of structural breaks, finding periods of monotonous behavior including its evaluation, or summarization of information about large sets of time series. Our goal is not to beat statistical methods, but vice versa—to benefit from the synergy of both.

Keywords Time series · Fuzzy transform · Evaluative linguistic expressions · Fuzzy natural logic · Mining information from time series

1 Introduction

The main goal of this paper is to provide overview of the results that have been obtained over several years in analyzing, forecasting, and mining information from time series using methods that predominantly have nonstatistical character. Since

The work was supported from ERDF/ESF by the project “Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region” No. CZ.02.1.01/0.0/0.0/17-049/0008414.

V. Novák (✉) · I. Perfilieva

University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, Ostrava, Czech Republic

e-mail: Vilem.Novak@osu.cz; Irina.Perfilieva@osu.cz

our methods are not sufficiently known among statisticians, we want to fill in this gap. We argue that our methods can be successful in time series processing and demonstrate this on examples. On the one hand, they provide similar results as statistical methods but give a different point of view on them. On the other hand, they also give an explanation of the obtained results. Moreover, our methods can provide additional information that can hardly be obtained using statistical methods. We will discuss this feature of our methods in the sequel. However, let us emphasize that our goal is not to beat statistical methods but rather to extend the power of time series processing methods and benefit from the mutual synergy. Let us remark that all methods described below have been developed by the authors of this paper and their collaborators.

Recall that by a fuzzy set A on the universe U , we understand a function $A : U \rightarrow [0, 1]$.¹ We often write $A \subseteq U$ or $A \in \mathcal{F}(U)$ where $\mathcal{F}(U)$ is the set of all fuzzy sets on U .

It is important to note that there is an essential difference between the probabilistic and fuzzy techniques, which we extensively argued in [32] and elsewhere. Recall from the latter that the probability theory provides a model of the uncertainty characterized by a lack of information about possible results of some event (experiment). Fuzzy sets, on the other hand, provide a mathematical model of the vagueness phenomenon. The latter rises if we want to form a class of all objects with vague property (e.g., “warm,” “strong,” “steep,” etc.). No random event (a result of some experiment) occurred in this case and so, it cannot be considered.

Both phenomena are occurring in reality and should be treated using different mathematical principles. While probability theory is based on the properties of measure and a key notion is that of independence of events, fuzzy set theory (and fuzzy logic) is based on the properties of ordered structures. Thus, both probability and fuzzy set theory are complementary rather than competitive.²

In this paper, we present special techniques of fuzzy modeling suitable for applications in time series processing. The first one is the *fuzzy transform* (F-transform) and the second one are a few selected methods of *fuzzy natural logic* (FNL). These techniques are in detail described in various papers. A comprehensive explanation including applications can be found in the book [34].

Our methods can be applied to processing of *classical time series*. There is also a branch focusing on elaboration of the so-called fuzzy time series [36, 42] which are sequences $\{A(t) \mid t \in \mathbb{T}\}$ where $A(t)$ are fuzzy sets. We do not deal with this approach in this paper.

This paper is structured as follows. In the first two sections, we introduce the basic concepts of fuzzy transform and fuzzy natural logic. In Sect. 4, we introduce the decomposition of time series into components that are later elaborated

¹ The interval $[0, 1]$ is a set of *truth* values where 0 means falsity, 1 truth, and the other values express partial truth. This interval can be replaced by a suitable bounded lattice.

² It has a good sense to speak about the probability of fuzzy events. For example, what is the probability that in the next few minutes, we will meet a tall woman.

separately. Section 5 describes basic principles of forecasting of time series. In Sect. 6, we overview three main kinds of information that can be mined from time series. Its last subsection is a brief overview of other applications of our methods in time series processing.

2 Fuzzy Transform

This is a universal technique introduced by I. Perfilieva in [37, 39] that is widely applied in many areas. Its fundamental idea is to transform a real bounded continuous function $f : [a, b] \rightarrow [c, d]$, where $[a, b], [c, d] \subset \mathbb{R}$, to a finite vector of components and then transform it back. The former is called a *direct F-transform* and the latter an *inverse one*. The result of the inverse F-transform is a function $\hat{f} : [a, b] \rightarrow \mathbb{R}$ that *approximates* the original function f . We can set the parameters so that the approximating function \hat{f} has the desired properties.

The F-transform has several strengths: excellent approximation abilities, ability to filter out high frequencies, ability to reduce noise [18, 19, 35], and ability to estimate values of the first and second derivatives in an approximately specified area (cf. [11]).

The first step of the F-transform procedure is to form a *fuzzy partition* of the domain $[a, b]$ which consists of a finite set of fuzzy sets on $[a, b]$

$$\mathcal{A}_h = \{A_0, \dots, A_n\}, \quad n \geq 2, \quad (1)$$

defined over the set of nodes $a = c_0, \dots, c_n = b$ such that $c_{k+1} = c_k + h$ where $h > 0$. Each fuzzy set A_k has a support defined over three nodes c_{k-1}, c_k, c_{k+1} where $A(c_k) = 1$ and $A(c_{k-1}) = A(c_{k+1}) = 0$. The fuzzy sets A_k are often called *basic functions*. The properties of basic functions are defined axiomatically; for the details, see [37] and elsewhere.

If the fuzzy partition is given, then an $(n + 1)$ -tuple

$$\mathbf{F}^m[f] = (F_0^m[f], \dots, F_n^m[f])$$

is called m -th degree *direct fuzzy transform* of f if

$$F_k^m[f](x) = \beta_k^0[f] + \beta_k^1[f](x - c_k) + \dots + \beta_k^m[f](x - c_k)^2, \quad (2)$$

for all $k = 0, \dots, n$. We call $F_k^m[f]$ in (2) *components* of the fuzzy transform. Precise computation of the components (2) is in detail described in [20] and elsewhere.

The F-transform is linear, namely, if $f = \alpha_1 g_1 + \alpha_2 g_2$ where α_1, α_2 are numbers and g_1, g_2 real bounded functions on the same domain, then

$$F_k^m[f] = \alpha_1 F_k^m[g_1] + \alpha_2 F_k^m[g_2]$$

for all $k = 1, \dots, n$.

The *inverse F-transform* is

$$\hat{f}_h^m(x) = \sum_{k=0}^n F_k^m[f] \cdot A_k(x), \quad x \in [a, b]. \quad (3)$$

It can be proved that the function \hat{f}_h^m approximates the original function f with arbitrary precision depending on the choice of h when forming the fuzzy partition \mathcal{A}_h . The computational complexity of fuzzy transform is linear.

The following holds for the coefficients β_k^j in (2) (see [39]):

$$\beta_k^0[f] = f(c_k) + O(h^2), \quad (4)$$

$$\beta_k^1[f] = f'(c_k) + O(h^2), \quad (5)$$

$$\beta_k^2[f] = \frac{f''(c_k)}{2} + O(h^2). \quad (6)$$

Hence, each coefficient β_k^j provides a weighted average of values as well as of derivatives of the function f over the area characterized by the fuzzy set $A_k \in \mathcal{A}_h$.

3 Fuzzy Natural Logic

This is a class of special theories of mathematical fuzzy logic whose goal is to model the reasoning of people based on using natural language. So far, it consists of the following theories:

- (a) A formal theory of *evaluative linguistic expressions* explained in detail in [24] (see also [23, 34]) that are expressions of natural language such as *small, medium, big, very short, more or less deep, quite roughly strong, extremely high, etc.*
- (b) A formal theory of *fuzzy/linguistic IF-THEN rules* and approximate reasoning [22, 30, 33, 34]. The basic concept here is that of a *linguistic description*, that is, a finite set of fuzzy/linguistic IF-THEN rules:

$$\begin{aligned} \mathcal{R}_1 &= \text{IF } X \text{ is } \mathcal{A}_1 \text{ THEN } Y \text{ is } \mathcal{B}_1, \\ &\dots \\ \mathcal{R}_m &= \text{IF } X \text{ is } \mathcal{A}_m \text{ THEN } Y \text{ is } \mathcal{B}_m \end{aligned} \quad (7)$$

where “ X is \mathcal{A}_j ,” “ Y is \mathcal{B}_j ,” $j = 1, \dots, m$ are *evaluative linguistic predictions* (e.g., “trend is very steep, difference is small, trend-cycle is stagnating,” etc.). The linguistic description can be *learned from data*.

To find a proper conclusion on the basis of linguistic description, it is necessary to use a special reasoning method called *perception-based logical deduction* (PbLD). This method is based on the mathematical model of the used evaluative predications. To find conclusion, it acts locally so that it mimics the way how people make their reasoning on the basis of linguistic information. More detailed description of PbLD can be found in [34].

- (c) A formal theory of *intermediate and generalized fuzzy quantifiers* [6, 15, 25] and elsewhere. These are expressions of natural language such as *most, many, a lot of, a few, several*, etc.

Theory (b) is applied in forecasting. Theories (a) and (c) are applied in mining information from time series described in Sect. 6. For more details about FLN, see the cited literature. Less informal explanation of methods of FLN including description of applications can be found in [34].

4 Analysis of Time Series

Application of techniques of fuzzy modeling in time series analysis is based on the assumption that time series can be decomposed as follows: let $\mathbb{T} = \{1, \dots, p\}$ be a set of natural numbers interpreted as time moments. Then a time series is a set $X = \{X(t, \omega) \mid t \in \mathbb{T}, \omega \in \Omega\}$ where

$$X(t, \omega) = TC(t) + S(t) + R(t, \omega), \quad t \in \mathbb{T}, \omega \in \Omega. \quad (8)$$

The $TC(t)$ is a *trend-cycle* that can be further decomposed into *trend* and *cycle*, i.e., $TC(t) = Tr(t) + C(t)$. The $S(t)$ is a *seasonal* component that is a mixture of r periodic functions:

$$S(t) = \sum_{j=1}^r P_j e^{i\lambda_j t} \quad (9)$$

where $\lambda_1, \dots, \lambda_r$ are frequencies and $P_j, j = 1, \dots, r$ are constants. Without loss of generality, we assume that the frequencies are ordered $\lambda_1 < \dots < \lambda_r$ (this corresponds to ordering of periodicities $T_1 >, \dots, T_r$).

Note that TC and S are ordinary non-stochastic functions. Only R is a random *noise* and we assume that it is a stationary stochastic process with the mean $\mathbf{E}(R(t, \omega)) = 0$ and variance $\mathbf{Var}(R(t, \omega)) < \sigma, t \in \mathbb{T}$.

In practice, we always have only one realization of time series at disposal, which is obtained by fixing $\omega \in \Omega$. Then

$$X = \{X(t) \mid t \in \mathbb{T}\} \quad (10)$$

is an ordinary real (or complex) valued function.

Let us now choose a fuzzy partition \mathcal{A}_h for some $h > 0$ and apply the F-transform to X in (10). The result of the inverse F-transform is

$$\hat{X}(t) = \hat{T}C + \hat{S}(t) + \hat{R}(t), \quad t \in \mathbb{T}. \quad (11)$$

Then the following can be proved:

Theorem 1 ([16, 17, 35])

- (a) If we set $h = \bar{d}\bar{T}$ where $d > 0$ and \bar{T} is the longest periodicity occurring in S , then $\lim_{d \rightarrow \infty} |\hat{S}(t)| = 0$.
- (b) $\lim_{h \rightarrow \infty} \text{Var}(\hat{R}(t)) = 0$.
- (c) There is a number $D(m, h)$, $m \in \mathbb{N}$, such that

$$|\hat{X}(t) - TC(t)| \leq 2\omega(h, TC) + D(m, h), \quad t \in [c_1, c_{n-1}] \quad (12)$$

where $\lim_{h \rightarrow \infty} D(m, h) = 0$ and $\omega(h, TC)$ is a modulus of continuity w.r.t. h and TC .

It follows from this theorem that, by proper setting of h , the F-transform makes it possible to “wipe out” part or the whole of the seasonal component S of the time series and significantly reduce its noise. To set h , we follow the general OECD specification: *Trend (tendency)* is a component of a time series that represents variations of low-frequency, high-frequency, and medium-frequency fluctuations having been filtered out. *Trend-cycle* is a component that represents variations of low and medium frequency in a time series, the high-frequency fluctuations having been filtered out.

Hence, we proceed as follows:

- (i) Find periodicities:

$$T_1 > \dots > T_s \quad (13)$$

using *periodogram* (see [1, 2, 8] and elsewhere). Choose a proper periodicity T from the list (13) and due to Theorem 1(a), set $h = dT$ for some d (we usually put $d \in \{1, 2\}$).

- (ii) Form a fuzzy partition (1) and compute F^m -transform components:

$$\mathbf{F}^m[X] = (F_0^m[X], \dots, F_{n-1}^m[X]) \quad (14)$$

for $m \in \{0, 1\}$.

- (iii) Compute estimation of trend or trend-cycle using the inverse F-transform (3). Taking into account the equality (11) and Theorem 1, we can estimate *trend-cycle* as (by \approx we denote approximate equality)

$$TC \approx \hat{X}_{h_{TC}}$$

and *trend* as

$$T(t) \approx \hat{X}_{h_T}(t)$$

where h_{TC} is set according to a periodicity T that is chosen from the middle of the list (13) and for h_T it is chosen from the left part of it.

Note that (3) provides also analytic form of the estimation. As a consequence of Theorem 1, we conclude that the F-transform makes it possible to estimate trend or trend-cycle with high fidelity. A convincing demonstration of this statement is presented in [35].

5 Forecasting Time Series

Recall that the direct F-transform provides estimation of the trend-cycle TC (or trend) in the form of the vector of components (14). Then, using the special learning method developed in FNL (see [34]), we can learn the linguistic description which characterizes the principles of the behavior of TC . Then, using the PbLD method, we can forecast k future F-transform components:

$$F_n^m[X], \dots, F_{n+k-1}^m[X]. \quad (15)$$

Finally, from (15) we compute the forecast of the trend-cycle as the inverse F-transform $\hat{X}(t)$ for $t = p + 1, \dots, p + K$ where K is the forecast horizon. In our case, it is k -times the width of the basic functions, i.e., $K = 2kh$. The idea of the forecast is depicted in Fig. 1.

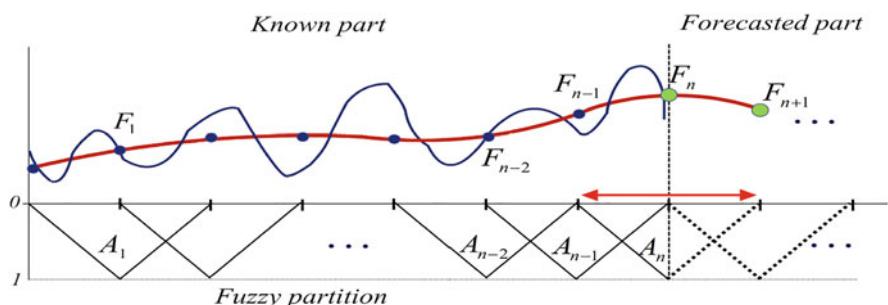


Fig. 1 Scheme of the forecasting idea: The component F_{n+1} is forecasted using PbLD method on the basis of the learned linguistic description



Fig. 2 Example of the time series forecasting. The left part is validation part; on the right is testing part (never used in computation); The dotted lines are real and forecasted trend-cycle; the full lines are real and forecasted values of time series

Example of such a learned linguistic description is

$$\begin{aligned}
 \mathcal{R}_1 = & \text{ IF } F_i[X] \text{ is extremely big AND } \Delta F_i[X] \text{ is rather medium} \\
 & \text{ THEN } F_{i+1}[X] \text{ is roughly big,} \\
 & \dots \\
 \mathcal{R}_m = & \text{ IF } F_i[X] \text{ is almost zero AND } \Delta F_i[X] \text{ is very small} \\
 & \text{ THEN } F_{i+1}[X] \text{ is significantly small}
 \end{aligned} \tag{16}$$

Let us emphasize that the learned linguistic description (16) explains in natural language the way how the forecast was obtained. This can be interesting information for the user that he/she can use in further decision or strategy of behavior (Fig. 2).

The seasonal component is forecast separately. The forecast of the whole time series is obtained by summing predictions of *TC* and *S* (cf. (8)). Demonstration of our forecasting method on real time series is presented in [21, 44] and elsewhere.

6 Mining Information from Time Series

One of the essential characteristics of our methods is the possibility to characterize various features of time series using expressions of natural language. Hence, our methods have a big potential in the area of mining information from time series since they can provide information that cannot be obtained using statistical methods. Below we will mention some of them. A concise overview of methods for mining information from time series is given in [7].

Linguistic Evaluation of Local Trend An exciting question is what trend (tendency) of the time series can be recognized in a specific time interval. Surprisingly, recognizing the trend is by no means a trivial task even when people watch the time series graph. Moreover, it can be essentially influenced by a subjective opinion. Therefore, objective and independent tool for this task is welcome. A convenient

one is the F^1 -transform since it makes it possible to estimate the average slope (tangent) over an imprecisely determined area, and using methods of FNL, it can be characterized in natural language. For example, we can say “*clear decrease (or huge increase)* of trend,” “the trend is *negligibly increasing (or stagnating)*,” etc. Such linguistic expressions characterize trend (tendency) of the time series in an area specified by the user. The ability to generate such linguistic evaluations is a quite important achievement of the fuzzy techniques. The method is based on the theoretical results in fuzzy natural logic and is described in more detail in [26, 27].

Evaluation of the steepness of the slope in natural language is determined using the function of *local perception*:³

$$\mathcal{A} = \text{LPerc}(\beta^1, w_{tg}) \quad (17)$$

which assigns a proper evaluative expression \mathcal{A} to the value β^1 w.r.t. the context w_{tg} . To determine it, we must first specify what does it mean “extreme (utmost) increase (decrease)” in a given context. It can be determined as the largest acceptable difference of time series values in relation to a given (basic) time interval (e.g., 12 months, 31 days), i.e., a minimal and maximal tangent. In practice, we set only the maximal tangent v_R , while the smallest one is usually $v_L = 0$. The typically medium value v_S is determined analogously as v_R . The result is the context $w_{tg} = \langle v_L, v_S, v_R \rangle$ that determines the interval $[v_L, v_S] \cup [v_S, v_R]$. Demonstration of the evaluation of the slope is in Fig. 3.

A related task is to find intervals in which the time series has a *monotonous trend* (see [32]). This means that we decompose the time domain \mathbb{T} into a set of intervals:

$$\mathcal{T} = \{\mathbb{T}_i \mid i = 1, \dots, s\}, \quad \bigcup \mathcal{T} = \mathbb{T} \quad (18)$$

so that the time series $X|\mathbb{T}_i$ (restriction of X to the interval \mathbb{T}_i) has a monotonous trend and each two adjacent time intervals $\mathbb{T}_i, \mathbb{T}_{i+1}$ have a common time point. Each \mathbb{T}_i is the largest interval that is evaluated using the same evaluative expression \mathcal{A} . For example, it is the largest interval, in which the trend is *stagnating (sharply increasing/decreasing, etc.)*, while the interval \mathbb{T}_{i+1} has a different slope.

Finding Perceptually Important Points Finding perceptually important points is another task successfully solved using our methods. According to [7], these are points where the time series essentially changes its course. In this paper, however, the authors have in mind just isolated points in the time series. However, its character can be quite complicated, various frequencies and noise are present, and,

³ Such a function is implemented in the experimental software LFL Forecaster (see http://irafm.osu.cz/en/c110_lfl-forecaster/) developed in the Inst. for Research and Applications of Fuzzy Modeling of the University of Ostrava, Czech Republic, which implements the described methods. Its author is Viktor Pavliska. The results demonstrated in this paper were obtained using the mentioned software.

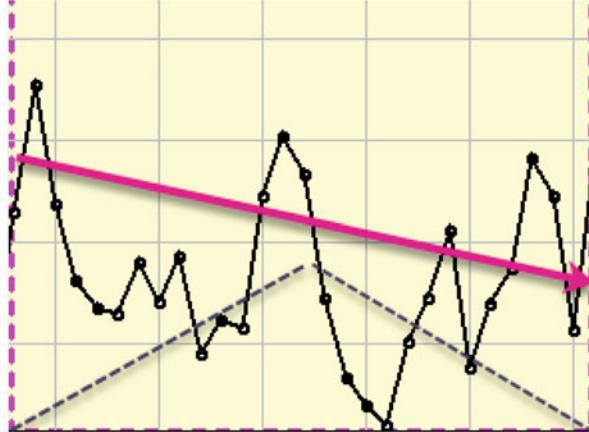


Fig. 3 Demonstration of the slope determined by the value of β^1 computed over an area characterized by a triangular basic function (depicted above the x -axis). It can be characterized linguistically w.r.t. context w_{tg} of the tangent which is determined by the ratio of the largest difference between values of the time series and the basic time interval (day, week, month, etc.). The evaluation in this picture is “slightly decreasing.” Note that human eye does not immediately see this slope from the course of the time series

therefore, we cannot expect that perceptually important point is just one isolated time point, but it is better an area that cannot be precisely determined. Therefore, we suggest a method based on the higher-degree F-transform because it makes it possible to estimate the first and second derivatives (5) and (6) of a function with complicated course in a vaguely specified area. The perceptually important points can be recognized in areas $A_k \in \mathcal{A}_h$ for $k \in \{1, \dots, k\}$ in which estimation of the slope (2) is close to zero, i.e., $\beta_k^1 \approx 0$. The method is in more detail explained in [29].

Recently, a new promising method for finding perceptually important points in time series has been presented in [38]. It is based on construction of a special Laplacian with kernels producing fuzzy partition used in fuzzy transform. The method can further be used to register similar time series or in a new algorithm for their forecasting.

Structural Breaks Structural breaks are sudden, considerable changes in the ordinary course of the time series X . In statistics, there are many methods suggested to solve this task [4, 5, 40].

In [28], we suggested a method for their detection which is similar to finding intervals of monotonous behavior described above. We check the slope of time series within two subsequent intervals determined by two adjacent fuzzy sets $A_i, A_{i+1} \in \mathcal{A}_h$ for a particular fuzzy partition \mathcal{A}_h with shorter h (in practice, we set $h \in \{4, \dots, 7\}$). The main difference lays in searching intervals \mathbb{T}_i in which the slope of $X|\mathbb{T}_i$ is *largely of hugely increasing/decreasing* and the slope $X|\mathbb{T}_{i+1}$ in

the adjacent interval \mathbb{T}_{i+1} is much less increasing/decreasing or even *stagnating*. Examples demonstrating our method in identification of structural breaks are in [44] (this volume). Let us remark we have also developed a method for detection of structural breaks in time series volatility.

Other Applications On the basis of our theories, we also developed the following methods:

- (a) Detection of “bull and bear” phases of financial time series—see [21].
- (b) Measures of similarity between time series — see [12, 31]. We suggested two indexes that measure similarity (and, potentially, dependence) between two time series. Both indexes are based on the F-transform and give convincing results.
- (c) Automatic summarization of knowledge about one or several time series. This task is addressed by various authors (see, e.g., [3, 7, 9, 10, 13, 41]). The theory of fuzzy natural logic contains a sophisticated formal theory of *intermediate quantifiers* that are expressions of natural language such as “many, almost all, most, a few,” etc. Using them, it is possible to derive statements that provide summarization of knowledge about time series. A typical example of such summarizing statement is

The trend during the past three years is in almost all tracked time series is clearly increasing.

The theory enables also humanlike syllogistic reasoning on the basis of the formal model of *generalized Aristotle’s syllogisms*. For more details, see [14, 26].

- (d) It is well-known that there are many methods for time series forecasting but none of them outperforms all the other ones. The reason is that each method is well suited to time series having specific features that, however, may not be fulfilled by the other ones and so, the given method fails. This suggests the idea to form a linear combination of several forecasting methods using weights that express a certain degree of successfulness of each method. However, it is difficult to set the weights. Our idea is to find a linguistic description (7) based on specific features of time series, for example, trend, seasonality, stationarity, and other ones. The linguistic description is learned using a method for mining linguistic associations. Our approach is described in [43] where also experimental justification is provided.

All our methods are robust and very fast because of the linear time complexity of the F-transform.

7 Conclusion

In this paper, we gave an overview of a few nonstatistical methods for analyzing and forecasting time series and mining information from them. The theoretical background of our methods is the theory of fuzzy transform and the theory of fuzzy natural logic. The former enables to estimate trend or trend-cycle with high fidelity and to reduce noise. Moreover, the F-transform also provides an analytic form of the latter. Using selected methods of FNL, we can accompany these results by explanation in natural language.

Moreover, a combination of the latter and the F-transform provides a forecast of time series. Further applications of our methods are in the area of mining information from them. They include finding intervals of monotonous behavior completed by its linguistic evaluation, detection of perceptually important points and structural breaks, summarization, measuring of similarity, and other applications.

References

1. Anděl, J.: Statistical Analysis of Time Series. SNTL, Praha (1976 (in Czech))
2. Bovas, A., Ledolter, J.: Statistical Methods for Forecasting. Wiley, New York (2003)
3. Castillo-Ortega, R., Marín, N., Sánchez, D.: A fuzzy approach to the linguistic summarization of time series. *Multiple Val. Logic Soft Comput.* **17**(2-3), 157–182 (2011)
4. De Wachter, S., Tzavalis, D.: Detection of structural breaks in linear dynamic panel data models. *Computat. Stat. Data Anal.* **56**(11), 3020–3034 (2012)
5. Doerr, B., Fischer, P., Hilbert, A., Witt, C.: Detecting structural breaks in time series via genetic algorithms. *Soft Computing* **21**(16), 4707–4720 (2017)
6. Dvořák, A., Holčapek, M.: L-fuzzy quantifiers of the type (1) determined by measures. *Fuzzy Sets Syst.* **160**, 3425–3452 (2009)
7. Fu, T.C.: A review on time series data mining. *Eng. Appl. Artif. Intell.* **24**, 164–181 (2011)
8. Hamilton, J.: Time Series Analysis. Princeton, Princeton University Press (1994)
9. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets Syst.* **159**, 1485–1499 (2008)
10. Kacprzyk, J., Wilbik, A., Zadrożny, S.: An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Int. J. Intell. Syst.* **25**, 411–439 (2010)
11. Kreinovich, V., Perfilieva, I.: Fuzzy transforms of higher order approximate derivatives: A theorem. *Fuzzy Sets Syst.* **180**, 55–68 (2011)
12. Mirshahi, S., Novák, V.: A fuzzy method for evaluating similar behaviour between assets. *Soft Computing* **25**, 7813–7823 (2021)
13. Moyse, G., Lesot, M.: Linguistic summaries of locally periodic time series. *Fuzzy Sets Syst.* **285**, 94–117 (2016)
14. Murinová, P., Novák, V.: A formal theory of generalized intermediate syllogisms. *Fuzzy Sets Syst.* **186**, 47–80 (2012)
15. Murinová, P., Novák, V.: The structure of generalized intermediate syllogisms. *Fuzzy Sets Syst.* **247**, 18–37 (2014)
16. Nguyen, L., Holčapek, M.: Suppression of high frequencies in time series using fuzzy transform of higher degree. In: Carvalho, J., et al. (eds.) *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 16th International Conference, IPMU 2016*, vol. 2, pp. 705–716. Springer (2016)

17. Nguyen, L., Holčapek, M.: Higher degree fuzzy transform: Application to stationary processes and noise reduction. In: Kacprzyk, J., et al. (eds.) *Advances in Fuzzy Logic and Technology 2017*, vol. 3, pp. 1–12. Springer (2018)
18. Nguyen, L., Novák, V.: Filtering out high frequencies in time series using F-transform with respect to raised cosine generalized uniform fuzzy partition. In: Proc. Int. Conference FUZZ-IEEE 2015. IEEE Computer Society, CPS, Istanbul (2015)
19. Nguyen, L., Novák, V.: Trend-cycle forecasting based on new fuzzy techniques. In: Proc. Int. Conference FUZZ-IEEE 2017, pp. 1–6. Naples, Italy (2017)
20. Nguyen, L., Holčapek, M., Novák, V.: Multivariate fuzzy transform of complex-valued functions determined by monomial basis. *Soft computing*, 3641–3658 (2017)
21. Nguyen, L., Mirshahi, S., Novák, V.: Trend-cycle estimation using fuzzy transform and its application for identifying of bull and bear phases in markets. *Intell. Syst. Account. Finance Manag.* **27**, 111–124 (2020). <https://doi.org/10.1002/isaf.1473>
22. Novák, V.: Perception-based logical deduction. In: Reusch, B. (ed.) *Computational Intelligence, Theory and Applications*, pp. 237–250. Springer, Berlin (2005)
23. Novák, V.: Mathematical fuzzy logic in modeling of natural language semantics. In: Wang, P., Ruan, D., Kerre, E. (eds.) *Fuzzy Logic – A Spectrum of Theoretical & Practical Issues*, pp. 145–182. Elsevier, Berlin (2007)
24. Novák, V.: A comprehensive theory of trichotomous evaluative linguistic expressions. *Fuzzy Sets Syst.* **159**(22), 2939–2969 (2008)
25. Novák, V.: A formal theory of intermediate quantifiers. *Fuzzy Sets Syst.* **159**(10), 1229–1246 (2008)
26. Novák, V.: Linguistic characterization of time series. *Fuzzy Sets Syst.* **285**, 52–72 (2016)
27. Novák, V.: Mining information from time series in the form of sentences of natural language. *Int. J. Approx. Reason.* **78**, 192–209 (2016)
28. Novák, V.: Detection of structural breaks in time series using fuzzy techniques. *Int. J. Fuzzy Logic Intell. Syst.* **18**(1), 1–12 (2018)
29. Novák, V.: Fuzzy vs. probabilistic techniques in time series analysis. In: Anh, L., Dong, L., Kreinovich, V., Thach, N. (eds.) *Econometrics for Financial Applications*, pp. 213–234. Springer, Berlin (2018)
30. Novák, V., Lehmkne, S.: Logical structure of fuzzy IF-THEN rules. *Fuzzy Sets Syst.* **157**, 2003–2029 (2006)
31. Novák, V., Mirshahi, S.: On the similarity and dependence of time series. *MDPI Mathematics* **9**(5), 550–563 (2021). <https://doi.org/10.3390/math9050550>. <http://www.mdpi.com/2227-7390/9/5/550>
32. Novák, V., Pavliska, V.: Time series: how unusual local behavior can be recognized using fuzzy modeling methods. In: Kreinovich, V. (ed.) *Statistical and Fuzzy Approaches to Data Processing, with Applications to Econometrics and Other Areas*, pp. 157–177. Springer, Berlin (2021)
33. Novák, V., Perfilieva, I.: On the semantics of perception-based fuzzy logic deduction. *Int. J. Intell. Syst.* **19**, 1007–1031 (2004)
34. Novák, V., Perfilieva, I., Dvořák, A.: *Insight into Fuzzy Modeling*. Wiley, Hoboken, NJ (2016)
35. Novák, V., Perfilieva, I., Holčapek, M., Kreinovich, V.: Filtering out high frequencies in time series using F-transform. *Information Sciences* **274**, 192–209 (2014)
36. Panigrahi, S., Behera, H.: Fuzzy time series forecasting: A survey. In: Behera, H., Nayak, J., Naik, B., Pelusi, D. (eds.) *Computational Intelligence in Data Mining*. pp. 641–651. Springer Singapore, Singapore (2020)
37. Perfilieva, I.: Fuzzy transforms: theory and applications. *Fuzzy Sets Syst.* **157**, 993–1023 (2006)
38. Perfilieva, I., Adamczyk, D.: Features as keypoints and how fuzzy transforms retrieve them. In: Rojas, I., Joya, G., Català, A. (eds.) *Advances in Computational Intelligence, IWANN 2021*, vol. 12862. Springer, Cham (2021)
39. Perfilieva, I., Daňková, M., Bede, B.: Towards a higher degree F-transform. *Fuzzy Sets Syst.* **180**, 3–19 (2011)

40. Preuss, P., Puchstein, R., Detter, H.: Detection of multiple structural breaks in multivariate time series. *J. Am. Stat. Assoc.* **110**, 654–668 (2015)
41. Said, A., Taskaya-Temizel, T., Khurshid, A.: Summarizing time series: Learning patterns in ‘Volatile’ series. In: Yang, Z., Yin, H., Everson, R. (eds.) *Intelligent Data Engineering and Automated Learning? IDEAL 2004*, pp. 523–532. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg (2004)
42. Song, Q., Chisom, B.: Forecasting enrollments with fuzzy time series - Part I. *Fuzzy Sets Syst.* **54**, 1–9 (1993)
43. Štěpnička, M., Burda, M., Štěpničková, L.: Fuzzy rule base ensemble generated from data by linguistic associations mining. *Fuzzy Sets Syst.* **285**, 140–161 (2016)
44. Truong, P., Novák, V.: An improved forecasting and detection of structural breaks in time series using fuzzy techniques. In: Rojas, I. (ed.) *Contribution to Statistics*. Springer (2022)

PMF Forecasting for Count Processes: A Comprehensive Performance Analysis



Annika Homburg, Christian H. Weiß, Layth C. Alwan, Gabriel Frahm,
and Rainer Göb

Abstract Coherent forecasting techniques account for the discrete nature of count processes. Besides point and interval forecasts, a third way for achieving coherent forecasts is to consider the full predictive probability mass function (PMF) as the actual forecast value. For a large variety of count processes, the performance of PMF forecasting under estimation uncertainty is analyzed. Furthermore, also Gaussian approximate PMF forecasting is investigated. Different approaches for performance evaluation are taken into consideration, with the main focus on mean squared errors computed for either the full PMF or its lower and upper tails, respectively. A real-world example from finance is presented for illustration.

Keywords Coherent forecasting · Count time series · Estimation error · Forecast distribution · Mean squared error

1 Introduction

In many real-world situations, we are concerned with count time series x_1, \dots, x_T , $T \in \mathbb{N} = \{1, 2, \dots\}$, where the observations x_t are nonnegative integer values, $x_t \in \mathbb{N}_0 = \{0, 1, \dots\}$ [12]. Examples include the numbers of transactions of financial products per trading day (see also the data example below), yearly numbers of natural disasters, monthly counts of major strikes, or daily numbers of new infections with a certain disease. The forecasting of the underlying count process $(X_t)_{t \in \mathbb{N}}$ should be done with different approaches than those used if we would be

A. Homburg · C. H. Weiß (✉) · G. Frahm

Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

L. C. Alwan

Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

R. Göb

Institute of Mathematics, Department of Statistics, University of Würzburg, Würzburg, Germany

concerned with a real-valued process, as the computed forecasts should account for the discrete nature of the counts. This task is referred to as *coherent forecasting* by [2]. Coherent forecasting is achieved by first deriving the (discrete) h -step-ahead conditional distribution of X_{T+h} (with forecast horizon $h \in \mathbb{N}$, given the past x_T, \dots, x_1), and by then computing

- The median or mode of X_{T+h} given x_T, \dots, x_1 as a central point forecast (PF)
- An extreme quantile of X_{T+h} given x_T, \dots, x_1 as a noncentral PF
- A finite subset of \mathbb{N}_0 satisfying the coverage requirement as a discrete type of prediction interval (PI) for X_{T+h} given x_T, \dots, x_1

Such types of coherent forecasts have been investigated by many researchers, including [2, 5–7]. PFs generally suffer from the fact that the observation X_{T+h} rarely agrees with the PF value (for real-valued processes, the agreement probability is even 0, whereas it is truly positive for discrete count processes). PIs for discrete count processes, in turn, often have a true coverage probability being much larger than the given coverage requirement (while in the real-valued case, an exact match is possible). Therefore, if forecasting a discrete-valued process, some authors suggest the full predictive probability mass function (PMF) itself as the actual forecast value; see [1, 8–11, 14]. Then, the user can judge which value will occur with which probability.

Two real-world examples (about transaction counts) of such PMF forecasts (PMFFs) are shown in Fig. 1. There, inspired by [13], the PMFF at time t is plotted as a vertical band of gray levels, where the intensity is proportional to the respective probability (white = zero probability). For comparison, also the actual observations are plotted in Fig. 1 such that we can judge their plausibility. Further details on these data, and on how to compute the PMFFs, are discussed in Sect. 5.

In this article, we provide a comprehensive analysis on the performance of coherent PMFFs in the presence of estimation uncertainty, where the latter results from fitting a model to the available time series data x_1, \dots, x_T . We consider a broad variety of data-generating processes (DGPs) to cover the practically relevant cases of unbounded counts (i.e., having full \mathbb{N}_0 as their range) and bounded counts (i.e., having range $\{0, \dots, n\}$ with some $n \in \mathbb{N}$), of different marginal

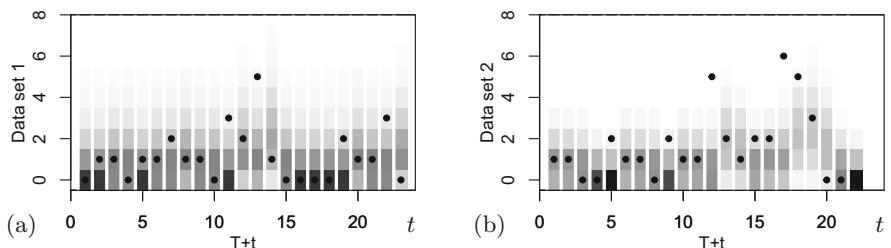


Fig. 1 PMFFs (expressed by gray levels) for two data sets of transaction counts, and actual observations plotted as dots against time t ; see (a) and (b), respectively

features such as equi-/overdispersion (variance equal/larger than the mean) or zero inflation (excessive number of zeros), and of different dependence structures (different autoregressive schemes of different orders). The detailed definitions and descriptions of the considered DGPs are provided by Supplement S.1 in the file “[PmfPredCountTS_suppl.pdf](#)”.

Although coherent approaches are commonly recommended for the forecasting of count processes, practitioners often use approximate forecasts instead (derived from, e.g., a fitted Gaussian autoregressive moving-average (ARMA) model). Typical reasons are an insufficient communication of count models and coherent approaches, as well as the ease of implementation because of readily available software solutions for Gaussian ARMA forecasting; see [6] for further details. Thus, as the second main research question, we compare the performance of PMFFs obtained from a Gaussian approximation (approximate PMFFs) with that of the coherent PMFFs. Both research questions are analyzed by means of a comprehensive simulation study.

Because of the strict page limit, the main manuscript focuses on a discussion of these simulation results. We refer the reader to the supplemental material for the full simulation results (file “[PmfPredCountTS_suppl.zip](#)”, found at <https://www.hsu-hh.de/mathstat/en/research/projects/forecastingrisk>). The outline of the article is as follows. Section 2 presents details on the definition and computation of the different types of PMFF. Section 3 provides a literature review on ways of evaluating the performance of PMFFs, and it concludes with a description of the approach considered in this research. Section 4, where we discuss our comprehensive performance analyses, constitutes the main part of our research. From the simulation results being provided there (and the full results in the supplemental material), it becomes clear that approximate PMFFs perform considerably worse than coherent PMFFs. The real-world application introduced in Fig. 1 is continued in Sect. 5, and the article concludes in Sect. 6.

2 Coherent and Approximate PMF Forecasting

In this section, we introduce the relevant terminology and notations regarding PMFFs, and we provide a clear description how coherent and approximate PMFFs are computed. For the count DGP (X_t), let the model be determined by some parameter vector θ , covering all distributional and dependence parameters. We consider three types of h -step-ahead PMFF, namely, $\hat{p}_{T+h}(\theta)$, $\hat{p}_{T+h}(\hat{\theta})$, and $\hat{p}_{T+h,a}(\hat{\theta})$, which are defined as follows.

Having observed the time series x_1, \dots, x_T , the conditional PMF of X_{T+h} given x_T, \dots, x_1 is the *true* h -step-ahead PMFF value. We denote this forecast value by $\hat{p}_{T+h}(\theta)$, where the x th component equals $\hat{p}_{T+h,x}(\theta) = P(X_{T+h} = x | x_T, \dots, x_1)$ for $x \in \mathbb{N}_0$. Equivalently, we may consider the cumulative distribution function (CDF) instead. In that case, we use the letters f and F instead

of \hat{p} and p , respectively, where $\hat{f}_{T+h,x}(\theta) = P(X_{T+h} \leq x | x_T, \dots, x_1)$. In practice, the model parameters are usually not known, so θ has to be estimated based on x_1, \dots, x_T . Then, the *coherent PMFF* is computed by using the parameter estimate $\hat{\theta}$ instead of θ , leading to the forecast value $\hat{p}_{T+h}(\hat{\theta})$. Because of estimation errors, $\hat{p}_{T+h}(\hat{\theta})$ will usually deviate from $\hat{p}_{T+h}(\theta)$. Ways of evaluating the forecast inaccuracy, i.e., the discrepancy between $\hat{p}_{T+h}(\hat{\theta})$ and $\hat{p}_{T+h}(\theta)$, are discussed in Sect. 3.

If forecasting is based on a Gaussian ARMA approximation, then first the Gaussian ARMA model is fitted to x_1, \dots, x_T , leading to the estimate $\hat{\vartheta}$ for the ARMA parameters ϑ . Then, the *approximate PMFF* is derived from the conditional normal distribution of this Gaussian model, say $N(\hat{\mu}_{T+h}, \hat{\sigma}_{T+h}^2)$, leading to $\hat{p}_{T+h,a}(\hat{\vartheta})$, where the subscript “a” abbreviates “approximate.” Now, the forecast inaccuracy, i.e., the deviation between $\hat{p}_{T+h,a}(\hat{\vartheta})$ and $\hat{p}_{T+h}(\hat{\theta})$, is caused by both approximation and estimation error. As discussed by [4], there are two common ways of deriving a Gaussian approximation $\hat{p}_{T+h,a}(\hat{\vartheta})$ from $N(\hat{\mu}_{T+h}, \hat{\sigma}_{T+h}^2)$. Let Φ denote the CDF of the standard normal distribution, $N(0, 1)$, then:

- The *simple normal approximation* implies to define $\hat{f}_{T+h,a,x}(\hat{\vartheta}) := \Phi((x - \hat{\mu}_{T+h})/\hat{\sigma}_{T+h})$
- Whereas the *continuity-corrected normal approximation* would lead to define $\hat{f}_{T+h,a,x}(\hat{\vartheta}) := \Phi((x - \hat{\mu}_{T+h} + 0.5)/\hat{\sigma}_{T+h})$

In any case, the PMFF $\hat{p}_{T+h,a}(\hat{\vartheta})$ is computed as discrete differences of $\hat{f}_{T+h,a}(\hat{\vartheta})$.

3 Performance Evaluation: A Critical Literature Review

To judge the performance of the different forecast approaches, we evaluate the forecast inaccuracy of the considered PMFF \hat{p} with respect to the true PMFF \hat{p}_0 . In our case, we take $\hat{p} \in \{\hat{p}_{T+h}(\hat{\theta}), \hat{p}_{T+h,a}(\hat{\vartheta})\}$ and $\hat{p}_0 = \hat{p}_{T+h}(\theta)$. Different solutions have been proposed in the literature yet.

In [4], it is distinguished between global and local inaccuracy measures. A *global* inaccuracy measure (or a visual tool used for global comparison) compares the PMFFs across the full support \mathbb{N}_0 , e.g., by using one of the probability metrics (divergence measures) reviewed by [3]. By contrast, *local* inaccuracy measures are restricted to certain properties of the PMFF that are judged as being particularly important, such as certain moments or quantiles of the PMFF. For example, if solely focusing on the PMFF’s median, then we essentially end up in evaluating the resulting central PF’s performance, as it was comprehensively investigated by [5]. In what follows, we take a global view at the PMFFs’ performance.

In [14], a χ^2 -distance between \hat{p} and \hat{p}_0 is used, i.e., the distance is proportional to $\sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2 / \hat{p}_{0,x}$; also see [3]. So we are concerned with a type of weighted *mean squared error* (MSE) between \hat{p} and \hat{p}_0 , with the weights being $1/\hat{p}_{0,x}$ for

$x \in \mathbb{N}_0$. This is similar to [10], who consider an unweighted MSE between $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}_0$ as one of their inaccuracy measures, i. e., $\|\hat{\mathbf{p}} - \hat{\mathbf{p}}_0\|^2 = \sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2$, while [4] used the Kullback-Leibler divergence, the Kolmogorov metric, and Raff's maximum error for this purpose. It should be noted, however, that the actual conclusions from the different global inaccuracy measures were quite similar, i. e., the overall evaluation of the different forecast competitors in [4] was the same for all metrics.

In [1], such global inaccuracy measures are criticized as being misleading in some applications, because overall goodness-of-fit may not exclude a rather poor performance in the upper tail, for example. Hence, [1] suggest to quantify the inaccuracy in terms of deviations between a set of upper quantiles; also see [4]. [10] did not only use the global MSE between $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}_0$ for performance evaluation, but also two MSEs referring to the lower and upper tail, respectively: one MSE is computed for the probabilities referring to the lower 25%-tail of $\hat{\mathbf{p}}_0$, and another one for the upper 10%-tail of $\hat{\mathbf{p}}_0$. With \hat{f} , \hat{f}_0 denoting the CDFs corresponding to $\hat{\mathbf{p}}$, $\hat{\mathbf{p}}_0$, we compute the local MSEs as $\sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2 \mathbb{1}(\hat{f}_{0,x} \leq 0.25)$ and $\sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2 \mathbb{1}(\hat{f}_{0,x} \geq 0.90)$, respectively, where the indicator function $\mathbb{1}(A)$ equals 1 (0) if A is true (false).

Finally, [8, 11] used different types of scoring rules for assessing the predictive performance. [8] recommends to use the quadratic score (Brier score), $s_{qs}(\hat{\mathbf{p}}, x) = -2 \hat{p}_x + \|\hat{\mathbf{p}}\|^2$. It should be noted, however, that considering the increase in the expected score by using $\hat{\mathbf{p}}$ instead of $\hat{\mathbf{p}}_0$, we end up with

$$\begin{aligned} & E \left[s_{qs}(\hat{\mathbf{p}}, X) - s_{qs}(\hat{\mathbf{p}}_0, X) \mid X \sim \hat{\mathbf{p}}_0 \right] \\ &= \sum_{x=0}^{\infty} \left(\|\hat{\mathbf{p}}\|^2 - \|\hat{\mathbf{p}}_0\|^2 + 2(\hat{p}_{0,x} - \hat{p}_x) \right) \hat{p}_{0,x} \\ &= \sum_{x=0}^{\infty} \left(\hat{p}_x^2 - 2\hat{p}_x \hat{p}_{0,x} + \hat{p}_{0,x}^2 \right) = \|\hat{\mathbf{p}} - \hat{\mathbf{p}}_0\|^2, \end{aligned} \quad (1)$$

which is one of the MSE measures used by [10]. Similarly, if using the ranked probability score $s_{rps}(\hat{f}, x) = \sum_{k=0}^{\infty} (\hat{f}_k - \mathbb{1}(x \leq k))^2 = \sum_{k=0}^{\infty} (\hat{f}_k^2 + (1 - 2\hat{f}_k) \mathbb{1}(x \leq k))$ based on the CDFs [8, 11], then

$$\begin{aligned} & E \left[s_{rps}(\hat{f}, X) - s_{rps}(\hat{f}_0, X) \mid X \sim \hat{\mathbf{p}}_0 \right] \\ &= \sum_{x,k=0}^{\infty} \left(\hat{f}_k^2 - \hat{f}_{0,k}^2 + 2(\hat{f}_{0,k} - \hat{f}_k) \mathbb{1}(x \leq k) \right) \hat{p}_{0,x} \\ &= \sum_{k=0}^{\infty} \left(\hat{f}_k^2 - \hat{f}_{0,k}^2 + 2(\hat{f}_{0,k} - \hat{f}_k) \hat{f}_{0,k} \right) = \|\hat{f} - \hat{f}_0\|^2. \end{aligned} \quad (2)$$

So again, we end up with an MSE as the inaccuracy measure, now relying on the CDFs instead of the PMFs.

To sum up, in view of the practice in evaluating PMFFs, we decided to use MSE-based inaccuracy measures. To avoid possibly misleading results as pointed out by

[1], we consider both the global measures computed for the full support \mathbb{N}_0 and the local measures restricted to the lower-25% tail and the upper-10% tail of \hat{p}_0 , respectively, where the latter choices are in accordance with [10]. Further details on performance evaluation are provided in the subsequent Sect. 4.

4 Results from a Comprehensive Simulation Study

For the DGPs described in Supplement S.1 and for each corresponding scenario according to Table 1, we simulated 1000 time series and fitted the respective model to the data. Here, we used the method of moments together with the moment formulae provided by Supplement S.1. Then, the PMF forecasts (or CDF forecasts, respectively) were computed according to the formulae for the transition probabilities in Supplement S.1. These PMF or CDF forecasts were used to compute the different types of MSE described in Sect. 3:

- Global MSEs $\|\hat{p} - \hat{p}_0\|^2 = \sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2$ (coherent PMF), $\|\hat{f} - \hat{f}_0\|^2$ (coherent CDF), $\|\hat{p}_a - \hat{p}_0\|^2$ (approximate PMF), and $\|\hat{f}_a - \hat{f}_0\|^2$ (approximate CDF)
- Local MSEs $\sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2 \mathbb{1}(\hat{f}_{0,x} \leq 0.25)$ (lower-25% tail MSE for coherent PMF) and $\sum_{x=0}^{\infty} (\hat{p}_x - \hat{p}_{0,x})^2 \mathbb{1}(\hat{f}_{0,x} \geq 0.90)$ (upper-10% tail MSE for coherent PMF), and the respective tail versions for approximate and CDF forecasts

If solely analyzing the coherent PMFFs' performance, we investigate the MSE values themselves. If comparing the approximate PMFFs' performance to the coherent ones, we focus on the differences computed between the approximate and the coherent MSEs, such as $\|\hat{p}_a - \hat{p}_0\|^2 - \|\hat{p} - \hat{p}_0\|^2$ or the respective tail and CDF versions. Here, a value > 0 implies that the approximate MSE is larger than the coherent one.

Either the 1000 MSE values per scenario themselves or the 1000 MSE differences were analyzed by using a “lean type of boxplot”:

- The median of the MSE (difference) values is plotted as a black dot.

Table 1 Scenarios for different DGPs of simulation study, with 1000 replications each

Means $\mu \in \{1, 1.075, \dots, 9.925, 10\}$ for unbounded counts,
upper bounds $n \in \{10, \dots, 130\}$, and probability $\pi \in \{0.15, 0.45\}$ for bounded counts
Dispersion ratios $I \in \{1.4, 2.4\}$ if considering overdispersion
Dependence parameter α in $\{0.33, 0.55, 0.8\}$ (ACF at lag 1),
and $\alpha_2 \in \{0.25, 0.35, 0.45\}$ as well as $\alpha_1 = \alpha(1 - \alpha_2)$ for AR(2)-like models
Sample sizes $T \in \{75, 250, 2500\}$

- The quartiles are connected by a thick gray line (as a substitute of the boxplot's box).
- The 10%- and 90%-quantiles are connected by a thin black line (as a substitute of the whiskers).

These boxplots are then plotted (closely together) against increasing mean μ . The full set of plots is provided in the supplemental material (but a few illustrative graphs are also shown below). There, we distinguish between boxplots for PMF and CDF forecasting, between the simple normal approximation (i. e., without continuity correction) and the continuity-corrected one, and between boxplots for the MSE values and for the MSE differences.

4.1 General Results

Before discussing the specific DGPs in some more detail (see Supplement S.1 for their definition), let us present some general conclusions drawn from the obtained simulation results. First, we compared the performance evaluation in terms of the PMF-based MSE (1) with the CDF-based MSE (2). In most cases, there was not much difference between both approaches. Although the actual MSE values might be different, the drawn conclusions regarding, e. g., the effects of overdispersion, or the performance difference between coherent and approximate PMFF forecasts, are the same. Thus, we focus on the PMF-based MSE values in the sequel, like it is done in [10].

Second, comparing the simple normal approximation to the continuity-corrected one, recalling Sect. 2, it turns out that the simple approximation does by far worse. In some cases, the MSE values of the simple approximation are increased by a factor between 5 and 10, both regarding the global MSE and the tail MSEs. Thus, the simple normal approximation is not further considered in the remaining discussion. If referring to approximate PMFFs, from now on, it is always assumed that the continuity correction described in Sect. 2 is used.

4.2 Performance of Coherent Forecasting

Let us start our discussion with the Poi-INAR(1) DGP (see Supplement S.1). The coherent PMFFs are generally close to the true PMFFs, with decreasing MSE values (of all types) for increasing mean μ and sample size T . Increases of the dependence parameter α , by contrast, lead to increased MSE values. Furthermore, see the upper panel in Fig. 2 for illustration; the lower-tail MSE is slightly larger than the upper-tail MSE, i. e., the estimation error becomes more apparent there. Note that in Fig. 2 (also later in Fig. 3), the “lean boxplots” defined on page 84 are shown, which should not be confused with the PMFFs of Fig. 1. If now considering

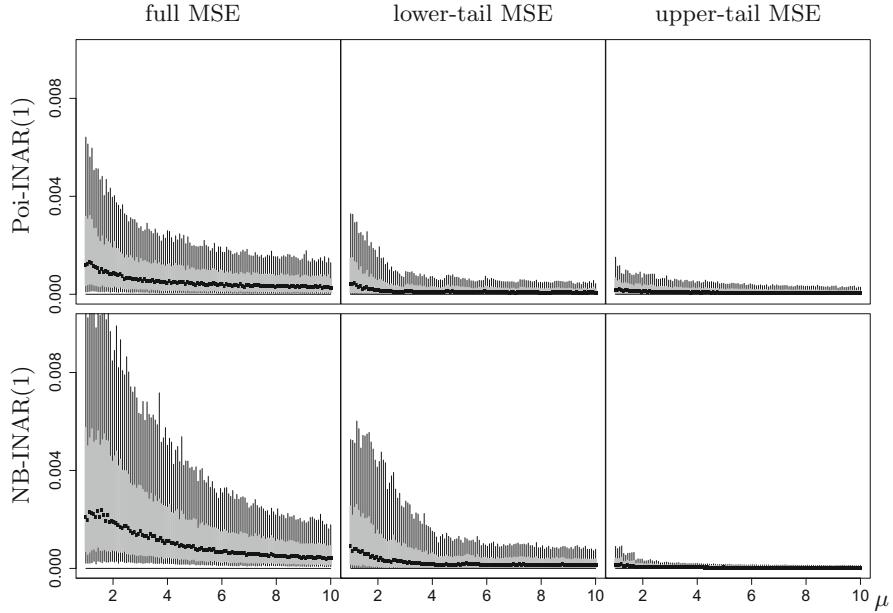


Fig. 2 Different types of MSE for coherent PMFFs with $\alpha = 0.55$, $T = 250$, and $h = 1$, plotted against mean μ . Upper panel, Poi-INAR(1) DGP; lower panel, NB-INAR(1) DGP with dispersion ratio $I = 2.4$

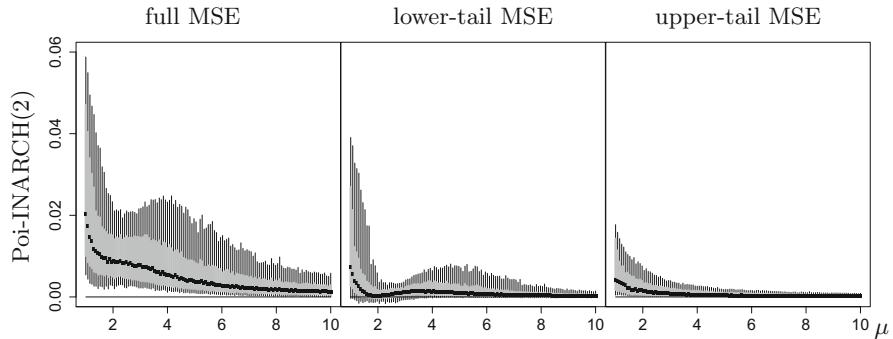


Fig. 3 Different types of MSE differences. PMFFs for Poi-INARCH(2) DGP with $\alpha = 0.55$, $\alpha_2 = 0.45$, $T = 250$, and $h = 1$, plotted against mean μ

additional overdispersion caused by the NB- or ZIP-INAR(1) model, see also the lower panel in Fig. 2; we observe increasing MSE values. This happens mainly for low μ (say $\mu \leq 4$) and large α (such as $\alpha = 0.8$), and especially for low T (such as $T = 75$). The increase is stronger for the lower- than for the upper-tail MSE. Also, the increase is slightly more pronounced for the ZIP model, i. e., where the overdispersion is caused by a single point mass in zero (“zero inflation”).

Next, let us investigate for a possible effect of the dependence structure, caused either by a different AR-type count DGP (Poi-INARCH vs. INAR family) or by an increased model order. In view of the effect of α already noted before, it is plausible that a further increase of the model order also further increases the MSEs. However, this happens again mainly for low T and large α as before. But it is interesting to note that the MSEs are generally larger for the INAR-type DGPs than for the INARCH-type ones. This seems to be due to the different types of sample paths generated by these models: highly dependent INAR DGPs lead to long constant segments (“runs,” i.e., low conditional variance), whereas INARCH sample paths always exhibit a lot of fluctuation; also see Section 4 in [12].

Finally, let us turn to the bounded counts generated by either the BinAR(1) or BinARCH(1) model. For the low value 0.15 of the normalized mean $\pi = \mu/n$, the MSE values are very similar to those of the respective (unbounded) Poi-INAR(CH) model. This is plausible in view of the Poisson limit theorem. Like before, the MSEs increase with increasing α , and BinARCH usually leads to smaller MSEs than BinAR. A notable difference is observed for $\pi = 0.45$, i.e., if the PMFFs are nearly symmetric distributions. Then, the MSE values get by far smaller (also those of the tails), and they are again smaller for the BinARCH than for the BinAR case.

To sum up, coherent PMFFs are affected by estimation error mainly for low sample size and large serial dependence. For sample size $T \geq 250$ or lag-1 ACF $\alpha \leq 0.55$, there is hardly any MSE left. PMFFs perform slightly better for INARCH- than INAR-type DGPs, and they further improve for bounded counts with nearly symmetric distribution. Overdispersion, by contrast, leads to a slight deterioration of forecast performance.

4.3 Performance of Approximate Forecasting

To compare the performance of the approximate PMFFs to those of the coherent ones, we analyzed the MSE differences; see Fig. 3 for illustration. In the large majority of all simulation runs, the approximate PMFFs produce clearly larger MSEs. This is particularly clear if T increases, because then, the coherent PMFFs’ performance notably improves (recall the Sect. 4.2), while this does not necessarily happen for the approximate PMFFs. The discrepancy between approximate and coherent PMFFs is particularly large for low means (say, $\mu \leq 6$), where the counts exhibit a rather asymmetric distribution. In view of this, it is also plausible that the approximate PMFFs perform rather well only for bounded counts with $\pi = 0.45$ (nearly symmetric distribution) and upper bound $n \geq 20$.

The discrepancy also intensifies for increasing overdispersion or increasing dependence. Regarding higher-order dependence, it can be recognized that the approximation especially increases the lower-tail MSE; also see Fig. 3. While the approximation error is stronger for Poi-INAR(2) than for Poi-INARCH(2), it is the other way round for the first-order models. In the presence of overdispersion, the approximation leads to increases in the global and the lower-tail MSE, while the

upper-tail MSE increases mainly if α is large (and then more for the ZIP- than for the NB-DGP).

All in all, we must clearly advise against the use of the approximate PMFFs. In a few situations, such as nearly symmetric bounded counts, the approximation does slightly worse, only. But in most of the considered scenarios (especially with increasing serial dependence or overdispersion), we observe a strong deterioration in forecast performance, independent of the sample size.

5 Application: PMF Forecasting of Transaction Counts

Let us pick up the data example of Fig. 1. We analyzed count time series about transaction numbers per trading day, referring to structured products from on-market and off-market trading as offered by the Cascade-Turnoverdata¹ of Deutsche Börse AG. Two exemplary time series are shown in Figs. 1 and 4. The first data set consists of $T_1 = 381$ counts (February 2017–July 2018) used for model fitting (see Fig. 4a), while the 23 counts from August 2018 are left for out-of-sample forecasting in Fig. 1a. The second data set from Fig. 4b includes one additional year of data ($T_2 = 636$ counts for February 2017–July 2019) for model fitting, and the 22 counts from August 2019 for forecasting in Fig. 1b. Figure 4 also shows the sample PACFs for both examples, which indicate that data set 1 might be well described by an AR(1)-like model, and data set 2 by an AR(2)-like model. Furthermore, data set 2 has a low mean (≈ 0.719) and a dispersion ratio close to one (≈ 0.913) such that a Poi-model seems appropriate, while data set 1 has the mean ≈ 1.493 and exhibits overdispersion (dispersion ratio ≈ 1.518). After thorough investigation, we decided to fit an NB-INAR(1) model to data set 1, and a Poi-INAR(2) to data set 2. Then, we applied these model fits to compute the 1-step-ahead PMFFs shown in Fig. 1.

A possible application of the obtained PMFFs is their integration into a “risk alert” system. The achieved transaction counts could be compared with their respective PMFFs to judge their plausibility, which is similar to a control chart application in statistical process control [12]. A visual inspection of Fig. 1, for example, suggests a possibly “unusual order book behavior” for data set 2, namely, for the counts at $t = 12, 17$. This could give rise to inform the traders on these days. In fact, such real-time risk alerts are a relevant topic for market infrastructure providers such as Deutsche Börse AG.² Certainly, if generating risk alerts based on PMFFs, also the estimation uncertainty due to model fitting should be taken into account. The PMFFs for data set 2 rely on $T_2 = 636$ observations, while those for data set 1 only use $T_1 = 381$. In view of our simulations, we do not expect a

¹ Retrieved December 15, 2020, from <https://datashop.deutsche-boerse.com/reference-data>.

² Retrieved December 15, 2020, from <https://www.deutsche-boerse.com/dbg-en/media/press-releases/Deutsche-Borse-supports-traders-with-real-time-risk-alerts-for-most-liquid-Eurex-futures-683428>.

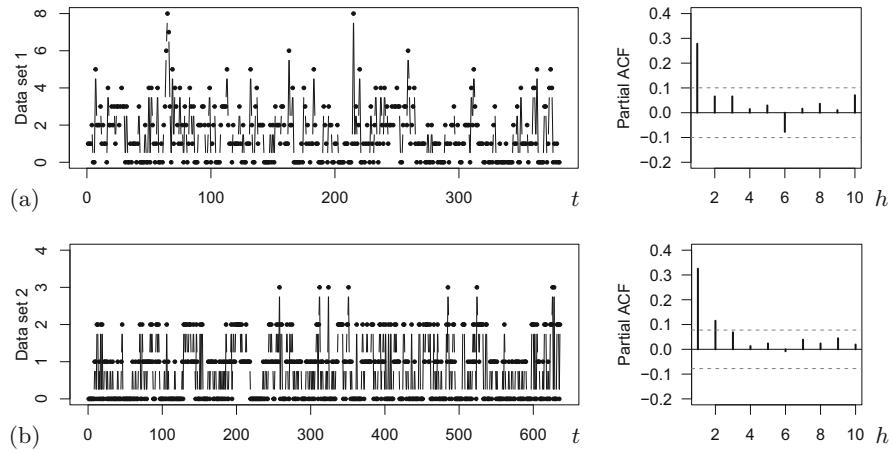


Fig. 4 Time series plot and sample PACF for two data sets of transaction counts; see (a) and (b), respectively

notable effect of parameter estimation, especially for data set 2. But since risk alerts are typically generated based on the tails of the PMFF, a careful investigation is recommended anyway. A possible solution could be to use a parametric bootstrap approach in analogy to [13].

6 Conclusions

For PMF forecasting, the full predictive PMF is taken as the forecast value, which is much more informative than a simple PF value, and which is more flexible than a PI with fixed coverage requirement. We did a comprehensive performance analysis of PMFFs for count processes, namely, by computing MSEs for the full predictive PMF as well as for its lower and upper tail, respectively, for a large variety of count processes. For coherent PMFFs, the effect of estimation error is generally rather low: deteriorations happen mainly for low sample size and strong dependence, and they are more pronounced for low means and in the presence of overdispersion. Thus, for the real-world examples on transaction counts in Sect. 5, we do not expect a notable effect of parameter estimation on PMFF performance.

The situation is quite different if PMFFs rely on a Gaussian approximation rather than on a coherent count model. Then, we observe a strong deterioration in forecast performance, even if a continuity correction is used for computing the approximate PMFFs. Also an increased sample size does not guarantee an improved performance. Thus, although such approximations may seem tempting in terms of implementation benefits, their use in practice is strongly discouraged.

Acknowledgments The authors thank the referees for their useful comments on an earlier draft of this article. The transaction count data of Sect. 5 were kindly made available to the authors by Deutsche Börse AG. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 394832307.

References

1. Boylan, J., Syntetos, A.: Accuracy and accuracy-implication metrics for intermittent demand. *Foresight* **4**, 39–42 (2006)
2. Freeland, R.K., McCabe, B.P.M.: Forecasting discrete valued low count time series. *Int. J. Forecast.* **20**(3), 427–434 (2004)
3. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. *Int. Stat. Rev.* **70**(3), 419–435 (2002)
4. Homburg, A.: Criteria for evaluating approximations of count distributions. *Comm. Stat. Simul. Comput.* **49**(12), 3152–3170 (2020)
5. Homburg, A., Weiß, C.H., Alwan, L.C., Frahm, G., Göb, R.: Evaluating approximate point forecasting of count processes. *Econometrics* **7**(3), 30 (2019)
6. Homburg, A., Weiß, C.H., Alwan, L.C., Frahm, G., Göb, R.: A performance analysis of prediction intervals for count time series. *J. Forecast.* **40**(4), 603–625 (2021)
7. Jung, R.C., Tremayne, A.R.: Coherent forecasting in integer time series models. *Int. J. Forecast.* **22**(2), 223–238 (2006)
8. Kolassa, S.: Evaluating predictive count data distributions in retail sales forecasting. *Int. J. Forecast.* **32**(3), 788–803 (2016)
9. McCabe, B.P.M., Martin, G.M.: Bayesian predictions of low count time series. *Int. J. Forecast.* **21**(2), 315–330 (2005)
10. McCabe, B.P.M., Martin, G.M., Harris, D.: Efficient probabilistic forecasts for counts. *J. Roy. Stat. Soc. Ser. B* **73**(2), 253–272 (2011)
11. Snyder, R.D., Ord, J.K., Beaumont, A.: Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *Int. J. Forecast.* **28**(2), 485–496 (2012)
12. Weiß, C.H.: An Introduction to Discrete-Valued Time Series. Wiley, Chichester (2018).
13. Weiß, C.H., Homburg, A., Alwan, L.C., Frahm, G., Göb, R.: Efficient accounting for estimation uncertainty in coherent forecasting of count processes. *J. Appl. Stat.* **49**(8), 1957–1978 (2022)
14. Willemain, T.R.: Forecast-accuracy metrics for intermittent demands: Look at the entire distribution of demand. *Foresight* **4**, 36–38 (2006)

A Novel First-Order Autoregressive Moving Average Model to Analyze Discrete-Time Series Irregularly Observed



César Ojeda, Wilfredo Palma, Susana Eyheramendy, and Felipe Elorrieta

Abstract A novel first-order autoregressive moving average model for analyzing discrete-time series observed at irregularly spaced times is introduced. Under Gaussianity, it is established that the model is strictly stationary and ergodic. In the general case, it is shown that the model is weakly stationary. The lowest dimension of the state-space representation is given along with the one-step linear predictors and their mean squared errors. The maximum likelihood estimation procedure is discussed, and their finite-sample behavior is assessed through Monte Carlo experiments. These experiments show that the bias, the root mean square error, and the coefficient of variation are smaller when the length of the series increases. Further, the method provides good estimations for the standard errors, even with relatively small sample sizes. Also, the irregularly spaced times seem to increase the estimation variability. The application of the proposed model is made through two real-life examples. The first is concerned with medical data, whereas the second describes an astronomical data set analysis.

This article is based on Chapter 3 of the first author's doctoral thesis [28]. Supported by CONICYT PFCHA/2015-21151457 and the ANID Millennium Science Initiative ICN12_009, awarded to the Millennium Institute of Astrophysics.

C. Ojeda (✉)

Escuela de Estadística, Universidad del Valle, Cali, Colombia

e-mail: cesar.ojeda@correounivalle.edu.co

W. Palma

Millennium Institute of Astrophysics, Santiago, Chile

S. Eyheramendy

Faculty of Engineering and Sciences, Universidad Adolfo Ibañez, Santiago, Chile

e-mail: susana.eyheramendy@uai.cl

F. Elorrieta

Department of Mathematics and Computer Science, Universidad de Santiago, Santiago, Chile

e-mail: felipe.elorrieta@usach.cl

Keywords State-space representation · Maximum likelihood · Prediction · General backward continued fraction

1 Introduction

In statistics, time series analysis establishes a principal tool for studying time-ordered observations that are naturally dependent. Nowadays, to study discrete-time series, many methods assume that time series are regularly observed; that is, the interval between observations is constant over time [5, 6, 15]. However, there are several fields as diverse as astronomy, climatology, economics, finance, medical sciences, and geophysics, where time series are observed at irregularly spaced intervals [2, 3, 8, 10, 12, 13, 17, 21, 23–26, 37]. For example, author [26] mentions that conventional time series analysis largely ignored irregularly spaced structures that climate time series has to consider.

The statistical analysis of irregular structures in time series poses several difficulties. First, the overwhelming majority of the available time series methods assume regularly observed data, as mentioned above. Second, when this assumption is dropped, several technical problems arise including the issue of formulating appropriate methodologies for carrying out statistical inferences. Third, most of the currently available numerical algorithms for computing estimators and forecasts are based on the regularity of the data collection process.

According to the paper [19], irregularly spaced time series can occur in two different ways. On the one hand, data can be regularly spaced with missing observations. On the other hand, data can be truly irregularly spaced with no underlying sampling interval. Techniques considering discrete-time series in the presence of missing data have been studied, for instance, by [9, 18, 30, 32]. Nevertheless, these techniques cannot be applied if data are really irregularly spaced. When data are irregularly observed, it has been treated through two approaches. First, it could be transformed irregularly spaced time series into regularly spaced time series through interpolation to use traditional techniques. In paper [1] can be found a summary of such transformations frequently used to analyze astronomical data. However, these interpolation methods typically produce bias (for instance, over smoothing), changing the dynamic of the process. Second, irregularly spaced time series can be treated as discrete realizations of a continuous stochastic process [31, 33, 35]. Nevertheless, continuous time series models tend to be computationally demanding and complicated (mostly due to the difficulty of estimating and evaluating them from discretely sampled data). To analyze discrete-time series observed at irregularly spaced times directly, authors [13] propose a first-order autoregressive model, while authors [29] propose a first-order moving average model. Consequently, a novel model is proposed in this paper which allows for the treatment of moving averages and autoregressive structures with irregularly spaced discrete-times.

The remainder of the paper is organized as follows. Section 2 introduces the construction of the model. The model definition and its properties are given in

Sect. 3. Also, this section provides the state-space representation of the model along with one-step linear predictors and their mean squared errors. The maximum likelihood estimation method is introduced in Sect. 4. The finite-sample behavior of this estimator is studied via Monte Carlo in Sect. 5. Two real-life data applications are discussed in Sect. 6, while conclusions are given in Sect. 7.

2 Model Formulation

This section describes a stationary stochastic process with an autoregressive moving average structure that allows to consider irregularly spaced times. The pattern of irregular spacing is assumed to be independent of the stochastic process properties. Also, it is assumed that all joint moments up to order two are finite.

Let $\mathbb{T} = \{t_n\}_{n \in \mathbb{N}^+}$ be a set of given times such that its consecutive differences, $\Delta_{n+1} = t_{n+1} - t_n$, are such that there is $\Delta_L > 0$ such that $\Delta_L \leq \Delta_{n+1}$ for all n . Without loss of generality, it is assumed that $\Delta_L = 1$. Otherwise, each t_n can be rescaled by Δ_L . These conditions are compatible with any physical measurement and determine \mathbb{T} as a discrete and therefore countable subset of \mathbb{R} .

Let $\{\zeta_{t_n}\}_{t_n \in \mathbb{T}}$ be a sequence of uncorrelated-standardized random variables and define the following sequence of real-valued random variables:

$$X_{t_1} = v_1^{1/2} \zeta_{t_1}, \quad X_{t_{n+1}} = \phi^{\Delta_{n+1}} X_{t_n} + v_{n+1}^{1/2} \zeta_{t_{n+1}} + \varpi_n v_n^{1/2} \zeta_{t_n},$$

where $0 \leq \phi < 1$; $\{v_n\}_{n \in \mathbb{N}^+}$ and $\{\varpi_n\}_{n \in \mathbb{N}^+}$ are time-varying sequences that characterize the moments of the process. Thus, for all n , $E(X_{t_n}) = 0$, $\text{Var}(X_{t_1}) = v_1$, $\text{Var}(X_{t_{n+1}}) = \phi^{2\Delta_{n+1}} \text{Var}(X_{t_n}) + v_{n+1} + \varpi_n^2 v_n + 2\phi^{\Delta_{n+1}} \varpi_n v_n$, and

$$\text{Cov}(X_{t_n}, X_{t_{n+k}}) = \begin{cases} \phi^{\Delta_{n+1}} \text{Var}(X_{t_n}) + \varpi_n v_n, & k = 1, \\ \phi^{\Delta_{n+k}} \text{Cov}(X_{t_n}, X_{t_{n+k-1}}), & k \geq 2. \end{cases} \quad (1)$$

By successive substitutions, for $k \geq 2$,

$$\text{Cov}(X_{t_n}, X_{t_{n+k}}) = \phi^{t_{n+k} - t_{n+1}} \text{Cov}(X_{t_n}, X_{t_{n+1}}).$$

To obtain a stationary process, it is required that, for all n , $\text{Var}(X_{t_n}) = \gamma_0$ and $\text{Cov}(X_{t_n}, X_{t_{n+1}}) = \gamma_{1,\Delta_{n+1}}$ with γ_0 time-independent and $\gamma_{1,\Delta_{n+1}}$ a function of Δ_{n+1} and not of the times itself. Thus,

$$\phi^{2\Delta_{n+1}} \gamma_0 + v_{n+1} + \varpi_n^2 v_n + 2\phi^{\Delta_{n+1}} \varpi_n v_n = v_1 = \gamma_0, \text{ and} \quad (2)$$

$$\phi^{\Delta_{n+1}} \gamma_0 + \varpi_n v_n = \gamma_{1,\Delta_{n+1}}. \quad (3)$$

From (3),

$$\varpi_n = \frac{\gamma_{1,\Delta_{n+1}} - \phi^{\Delta_{n+1}}\gamma_0}{v_n}. \quad (4)$$

Replacing (4) into (2),

$$v_{n+1} = \gamma_0 + \phi^{2\Delta_{n+1}}\gamma_0 - 2\phi^{\Delta_{n+1}}\gamma_{1,\Delta_{n+1}} - \frac{(\gamma_{1,\Delta_{n+1}} - \phi^{\Delta_{n+1}}\gamma_0)^2}{v_n}, \quad \text{with } v_1 = \gamma_0.$$

Also, since the process must be real-valued (i.e., without complex components), it is necessary that $v_n > 0$, for all n . Thus, particular forms can be specified to γ_0 and $\gamma_{1,\Delta_{n+1}}$ that satisfy this condition to get the desired model. In this case, these forms are chosen to obtain the traditional ARMA(1,1) model when times are regularly observed. Consequently, consider $\gamma_0 = \sigma^2(1+2\phi\theta+\theta^2)/(1-\phi^2)$ and $\gamma_{1,\Delta_{n+1}} = \phi^{\Delta_{n+1}}\gamma_0 + \sigma^2\theta^{\Delta_{n+1}}$ with $\sigma^2 > 0$ and $0 \leq \phi, \theta < 1$. Thence, $\varpi_n = \sigma^2\theta^{\Delta_{n+1}}/v_n$ and

$$v_{n+1} = \sigma^2 \left(\frac{(1+2\theta\phi+\theta^2)}{(1-\phi^2)}(1-\phi^{2\Delta_{n+1}}) - 2\phi^{\Delta_{n+1}}\theta^{\Delta_{n+1}} - \frac{\sigma^2\theta^{2\Delta_{n+1}}}{v_n} \right)$$

with $v_1 = \gamma_0$. To show that $v_n > 0$, for all n , define

$$c_{n+1}(\phi, \theta) = c_1(\phi, \theta)(1 - \phi^{2\Delta_{n+1}}) - 2\phi^{\Delta_{n+1}}\theta^{\Delta_{n+1}} - \frac{\theta^{2\Delta_{n+1}}}{c_n(\phi, \theta)}$$

with $c_1(\phi, \theta) = (1+2\phi\theta+\theta^2)/(1-\phi^2)$. Hence, $v_1 = \sigma^2 c_1(\phi, \theta)$, $v_{n+1} = \sigma^2 c_{n+1}(\phi, \theta)$, and it would only be necessary to show that $c_n(\phi, \theta) > 0$ for all n . Since $0 \leq \phi, \theta < 1$, then $c_1(\phi, \theta) \geq 1 + \theta^2 / (1 - \phi^2) \geq 1 + \theta^2 = c_1(\theta) > 0$. Also, since $1 \leq \Delta_{n+1}$ for all n , then $\phi\theta \geq \phi^{\Delta_{n+1}}\theta^{\Delta_{n+1}}$ for all n . Thus,

$$c_{n+1}(\phi, \theta) \geq 1 + \theta^2 - \frac{\theta^{2\Delta_{n+1}}}{c_n(\phi, \theta)} = c_{n+1}(\theta).$$

Here, $c_n(\phi, \theta) = c_n(\theta)$ since it is only a function of θ . So, it suffices to show that $c_n(\theta) > 0$ for all n with $c_1(\theta) = 1 + \theta^2$ and $c_{n+1}(\theta) = c_1(\theta) - \theta^{2\Delta_{n+1}}/c_n(\theta)$. From [20], the sequence $\{c_n(\theta)\}_{n \in \mathbb{N}^+}$ is known as a general backward continued fraction. In [29], it is shown that assuming $1 \leq \Delta_{n+1}$ for all n and $0 \leq \theta < 1$, this sequence is strictly positive. Thus, $v_n > 0$ for all n , and the desired model has been obtained.

3 An Irregular Observed First-Order Autoregressive Moving Average Model

A novel stationary stochastic process with an autoregressive moving average structure that allows considering irregularly observed times is defined. It is called irregularly observed first-order autoregressive moving average (iARMA) model.

Definition 1 (iARMA Model) Let $\{\varepsilon_{t_n}\}_{t_n \in \mathbb{T}}$ be a sequence of uncorrelated random variables with mean 0 and variance $\sigma^2 c_n(\phi, \theta)$ with $\sigma^2 > 0$, $0 \leq \phi, \theta < 1$, $c_1(\phi, \theta) = \frac{1+2\theta\phi+\theta^2}{1-\phi^2}$, and

$$c_{n+1}(\phi, \theta) = c_1(\phi, \theta)(1 - \phi^{2\Delta_{n+1}}) - 2\phi^{\Delta_{n+1}}\theta^{\Delta_{n+1}} - \frac{\theta^{2\Delta_{n+1}}}{c_n(\phi, \theta)}.$$

The process $\{X_{t_n}\}_{t_n \in \mathbb{T}}$ is said to be an iARMA process if $X_{t_1} = \varepsilon_{t_1}$, and

$$X_{t_{n+1}} = \phi^{\Delta_{n+1}} X_{t_n} + \varepsilon_{t_{n+1}} + \frac{\theta^{\Delta_{n+1}}}{c_n(\phi, \theta)} \varepsilon_{t_n}. \quad (5)$$

It is said that $\{X_{t_n}\}_{t_n \in \mathbb{T}}$ is an iARMA process with mean μ if $\{X_{t_n} - \mu\}_{t_n \in \mathbb{T}}$ is an iARMA process.

In the iARMA model, when $\phi = 0$, it is obtained the so-called iMA process [29], while when $\theta = 0$, the so-called iAR process [13] is obtained. Also, when $\Delta_{n+1} = 1$ for all n , it is obtained the traditional ARMA(1,1) process.

3.1 Properties

For the iARMA process, the mean and the autocovariance functions are

$$\mathbb{E}(X_{t_n}) = 0, \quad \text{and} \quad \text{Cov}(X_{t_n}, X_{t_{n+k}}) = \begin{cases} \sigma^2 c_1(\phi, \theta), & k = 0, \\ \gamma_{1, \Delta_{n+1}}, & k = 1, \\ \phi^{t_{n+k}-t_{n+1}} \gamma_{1, \Delta_{n+1}}, & k \geq 2, \end{cases}$$

for all n , where $\gamma_{1, \Delta_{n+1}} = \sigma^2 [\phi^{\Delta_{n+1}} c_1(\phi, \theta) + \theta^{\Delta_{n+1}}]$. The autocorrelation function is

$$\text{Cor}(X_{t_n}, X_{t_{n+k}}) = \begin{cases} 1, & k = 0, \\ \rho_{1, \Delta_{n+1}}, & k = 1, \\ \phi^{t_{n+k}-t_{n+1}} \rho_{1, \Delta_{n+1}}, & k \geq 2, \end{cases}$$

for all n , where $\rho_{1,\Delta_{n+1}} = \phi^{\Delta_{n+1}} + \theta^{\Delta_{n+1}}/c_1(\phi, \theta)$. Since the process has a constant mean and a covariance function that depends only on the time differences, the process is weakly stationary. In particular, if $\{\varepsilon_{t_n}\}_{t_n \in \mathbb{T}}$ are independent random variables each $N(0, \sigma^2 c_n(\phi, \theta))$, then the iARMA process would be a weakly stationary Gaussian process, and therefore strictly stationary.

Now, from (5), consider $Y_{t_{n+1}} = \varepsilon_{t_{n+1}} + [\theta^{\Delta_{n+1}}/c_n(\phi, \theta)]\varepsilon_{t_n}$ with $\text{Var}(Y_{t_{n+1}}) = \sigma^2[c_1(\phi, \theta)(1 - \phi^{2\Delta_{n+1}}) - 2\phi^{\Delta_{n+1}}\theta^{\Delta_{n+1}}]$. Hence, $X_{t_{n+1}} = \phi^{\Delta_{n+1}}X_{t_n} + Y_{t_{n+1}}$ for all n , with $X_{t_1} = \varepsilon_{t_1}$ and $\text{Cov}(X_{t_n}, Y_{t_{n+1}}) = \sigma^2\theta^{\Delta_{n+1}}$. By successive substitutions,

$$X_{t_{n+1}} = \phi^{t_{n+1}-t_1}\varepsilon_{t_1} + \sum_{j=1}^n \phi^{t_{n+1}-t_{j+1}}Y_{t_{j+1}}. \quad (6)$$

Consequently, for larger n , the initial condition effect vanishes. Thus, the process “forgets” its initial starting value. Also, from (6), X_{t_n} can be expressed as a function of $\{\varepsilon_{t_j}\}_{j=1}^n$, for each n . Then, under independence between these errors, X_{t_n} is ergodic [34].

3.2 State-Space Representation

From Definition 1, it is presented a state-space representation of the model (5). It enables the application of the Kalman filter for prediction and allows the maximum likelihood estimation; see [16]. This representation has the lowest dimension of the state vector and is given by

$$X_{t_n} = \alpha_{t_n} + \varepsilon_{t_n}, \quad \alpha_{t_1} = 0, \quad \alpha_{t_{n+1}} = \phi^{\Delta_{n+1}}\alpha_{t_n} + \left(\phi^{\Delta_{n+1}} + \frac{\theta^{\Delta_{n+1}}}{c_n(\phi, \theta)}\right)\varepsilon_{t_n}.$$

In this representation, measurement and transition equation disturbances are correlated. From put off [16], these equations can be transformed into a new system with disturbances uncorrelated, which are

$$X_{t_n} = \alpha_{t_n} + \varepsilon_{t_n}, \quad \alpha_{t_1} = 0, \quad \alpha_{t_{n+1}} = \left(\phi^{\Delta_{n+1}} + \frac{\theta^{\Delta_{n+1}}}{c_n(\phi, \theta)}\right)X_{t_n} - \frac{\theta^{\Delta_{n+1}}}{c_n(\phi, \theta)}\alpha_{t_n}. \quad (7)$$

The inclusion of X_{t_n} in (7) does not affect the Kalman filter, as X_{t_n} is known at time t_n .

3.3 Prediction

Using the innovations algorithm [6], the one-step linear predictors for the iARMA model are $\hat{X}_{t_1}(\phi, \theta) = 0$, with mean squared error $E\{(X_{t_1} - \hat{X}_{t_1}(\phi, \theta))^2\} = \sigma^2 c_1(\phi, \theta)$, and

$$\hat{X}_{t_{n+1}}(\phi, \theta) = \phi^{\Delta_{n+1}} X_{t_n} + \frac{\theta^{\Delta_{n+1}}}{c_n(\phi, \theta)} (X_{t_n} - \hat{X}_{t_n}(\phi, \theta)), \quad n \geq 1,$$

with mean squared errors $E\{(X_{t_{n+1}} - \hat{X}_{t_{n+1}}(\phi, \theta))^2\} = \sigma^2 c_{n+1}(\phi, \theta)$.

4 Maximum Likelihood Estimation

Let X_t be observed at points t_1, \dots, t_N . The log-likelihood under Gaussianity is

$$-\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2} \sum_{n=1}^N \ln c_n(\phi, \theta) - \frac{1}{2} \sum_{n=1}^N \frac{(X_{t_n} - \hat{X}_{t_n}(\phi, \theta))^2}{\sigma^2 c_n(\phi, \theta)},$$

where ϕ , θ , and σ^2 are any admissible parameter values. Now, optimizing it for σ^2 , replacing the optimum into the log-likelihood, and organizing terms, it is obtained the reduced likelihood $q_N(\phi, \theta) = \ln \hat{\sigma}_N^2(\phi, \theta) + 1/N \sum_{n=1}^N \ln c_n(\phi, \theta)$ with $\hat{\sigma}_N^2(\phi, \theta) = 1/N \sum_{n=1}^N (X_{t_n} - \hat{X}_{t_n}(\phi, \theta))^2 / c_n(\phi, \theta)$. The maximum likelihood estimates of ϕ and θ , denoted as $\hat{\phi}_N$ and $\hat{\theta}_N$, respectively, are the values minimizing $q_N(\phi, \theta)$. The estimate of σ^2 is $\hat{\sigma}_N^2 = \sigma_N^2(\hat{\phi}_N, \hat{\theta}_N)$. The optimization can be done through the method proposed by [7], which allows general box constraints. Specifically, $q_N(\phi, \theta)$ can be minimized under the constraint $0 \leq \phi, \theta < 1$. Also, this method allows for finding the numerically differentiated Hessian matrix at the solution given. Solving it, and according to [15], estimated standard errors can be obtained.

5 Monte Carlo Experiments

This section provides a Monte Carlo study that assesses the finite-sample performance of the maximum likelihood (ML) estimator. The simulation considers $\sigma^2 = 1$, $\phi \in \{0.5\}$, $\theta \in \{0.1, 0.5, 0.9\}$, and $N \in \{100, 500, 1500\}$, where N represents the length of the series. Furthermore, $M = 1000$ trajectories are simulated, and for

each, ϕ and θ are estimated. It is regarded as regular ($\Delta_n = 1$ for $n = 2, \dots, N$) as well as irregular spaced times, where $\Delta_n \stackrel{\text{ind}}{\sim} 1 + \exp(\lambda = 1)$, for $n = 2, \dots, N$. Now, let $\hat{\phi}_m$ and $\hat{\theta}_m$ be the ML estimations for the m -th trajectory with $\widehat{\text{se}}(\hat{\phi}_m)$ and $\widehat{\text{se}}(\hat{\theta}_m)$ their estimated standard errors. These standard errors are estimated through the curvature of the likelihood surface at $\hat{\phi}_m$ and $\hat{\theta}_m$ (see Sect. 4). As a summary of these quantities, the mean value of the M maximum likelihood estimations are computed. For example, for the moving average parameter, $\hat{\theta} = 1/M \sum_{m=1}^M \hat{\theta}_m$ and $\widehat{\text{se}}(\hat{\theta}) = 1/M \sum_{m=1}^M \widehat{\text{se}}(\hat{\theta}_m)$.

5.1 Performance Measures

As a measure of estimator performance, root mean square error (RMSE) and coefficient of variation (CV) are considered. For example, for the ML estimator for θ , $\text{RMSE}_{\hat{\theta}} = (\widehat{\text{se}}(\hat{\theta})^2 + \text{bias}_{\hat{\theta}}^2)^{1/2}$, and $\text{CV}_{\hat{\theta}} = \widehat{\text{se}}(\hat{\theta})/|\hat{\theta}|$, where $\text{bias}_{\hat{\theta}} = \hat{\theta} - \theta$. Furthermore, as an approximate variance of the estimator, $\widehat{\text{se}}^2(\hat{\theta}) = 1/M-1 \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta})^2$ is used. Finally, according to [22], the Monte Carlo error (MCE) is estimated for every simulation via asymptotic theory through $\widehat{\text{se}}(\hat{\theta})/\sqrt{M}$. Remember that the MCE is a estimation of the standard deviation of the Monte Carlo estimator, taken across repetitions of the simulation, where each simulation is based on the same design and consists of M replications.

5.2 Simulation Results

Table 1 shows the performance measures of the estimator for maximum likelihood method. Bias, RMSE, and CV are smaller when N increases as expected. Also, the method provides good estimations for the standard error, even with relatively small sample sizes. Furthermore, although it is not shown, comparing these results with the one obtained assuming regularly spaced times (the conventional first-order ARMA model), the irregularly spaced times seem to increase the estimation variability.

Table 1 Monte Carlo results for the irregularly spaced time case. The maximum MCE estimated (in all simulations) is 0.008. When $\phi = 0.5$, we use $\theta = 0.5$

N	θ	$\hat{\theta}$	$\widehat{se}(\hat{\theta})$	$\widetilde{se}(\hat{\theta})$	bias $_{\hat{\theta}}$	RMSE $_{\hat{\theta}}$	CV $_{\hat{\theta}}$
100	0.1	0.294	0.245	0.245	0.194	0.312	0.835
	0.5	0.500	0.252	0.263	0.000	0.252	0.505
	0.9	0.796	0.232	0.228	-0.104	0.255	0.292
500	0.1	0.192	0.158	0.179	0.092	0.183	0.827
	0.5	0.501	0.149	0.160	0.001	0.149	0.298
	0.9	0.885	0.090	0.094	-0.015	0.091	0.102
1500	0.1	0.131	0.102	0.116	0.031	0.106	0.780
	0.5	0.499	0.094	0.098	-0.001	0.094	0.188
	0.9	0.895	0.050	0.049	-0.005	0.050	0.056
N	ϕ	$\hat{\phi}$	$\widehat{se}(\hat{\phi})$	$\widetilde{se}(\hat{\phi})$	bias $_{\hat{\phi}}$	RMSE $_{\hat{\phi}}$	CV $_{\hat{\phi}}$
100	0.5	0.448	0.155	0.167	-0.052	0.163	0.346
500		0.488	0.076	0.079	-0.012	0.077	0.156
1500		0.497	0.046	0.048	-0.003	0.046	0.092

6 Applications

This section illustrates the application of the proposed time series model to two real-life data sets. The first example is concerned with medical data, whereas the second application describes the analysis of an astronomical data set.

6.1 Lung Function of an Asthma Patient

In put off [3], measurements of the lung function of an asthma patient are analyzed. The observations are collected mostly at 2-hour time intervals but with irregular gaps (see the unequal spaced of tick marks in Fig. 1). However, as it was shown in [36], the trend component (obtained by decomposing original time series into trend, seasonal, and irregular components via the Kalman smoother) exhibits structural changes after 100th observation. Thus, the first 100 observations are considered here to analyze such a phenomenon. Below, the ML estimates are reported along with their respective estimated standard errors. Here, the autoregressive estimate is not significant (not shown), but the other estimates are significant at the 5% significance level suggesting an iMA model:

$$\hat{\theta} = 0.853 \quad \widehat{se}(\hat{\theta}) = 0.069 \quad \hat{\sigma}^2 = 258.286 \quad \widehat{se}(\hat{\sigma}^2) = 36.537$$

From Fig. 1, the fit seems adequate. Also, the standardized residuals seem to follow a standard normal distribution. Furthermore, this figure shows the ACF estimated and the results from a Ljung-Box test for the standardized residuals.

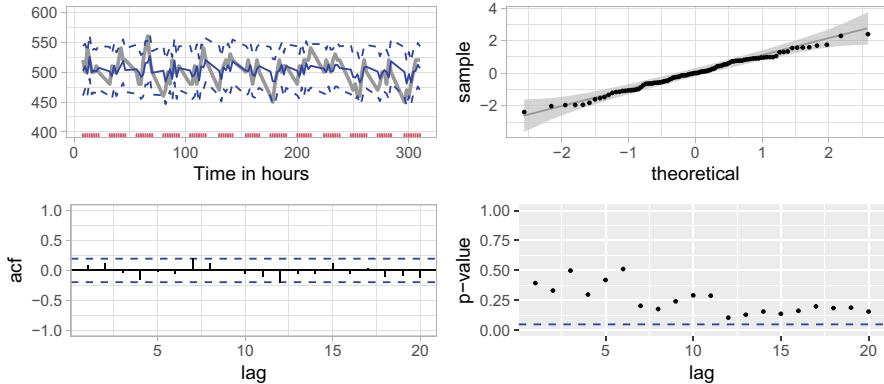


Fig. 1 On the left-top, the lung function of an asthma patient with the predicted values and their respective variability bands. For the standardized residuals: on the right-top, the quantile-quantile plot with normality reference bands [27]; on the bottom-left, the autocorrelation function estimated; on the bottom-right, the Ljung-Box test for randomness

Observe that the residuals satisfy the white noise test at the 5% significance level. Note that, since the standardized residuals are assumed to be realizations of a random sample, its correlation structure does not depend on the irregularly spaced between observations. Thus, unlike the original time series, the ACF and the Ljung-Box test can be applied to the standardized residuals.

6.2 Light Curve of an Astronomical Object

In astronomy, the study of the temporal behavior of the brightness of different objects is a matter of interest (see, for instance, [11]). The time series of the brightness of an astronomical object is called as light curve. Light curves are commonly measured at irregular times. In this work, it is also assessed the performance of the iARMA model in a light curve of an astronomical object. The light curve that it is used was observed with the Zwicky Transient Facility (ZTF) (see [4]) and belongs to a Blazar astronomical object coded as “ZTF18aabxyhf.” The time series data of this Blazar were processed by the ALerCE broker [14]. The light curve of this object has 65 measurements of the brightness of this object in a range of approximately 584 days. The average gap of the observations of this light curve is 9.13 days. The iARMA model parameters were estimated via maximum likelihood method in this light curve yielding the following results:

$$\begin{aligned}\hat{\phi} &= 0.702 \quad \hat{s.e}(\hat{\phi}) = 0.112 \quad \hat{\theta} = 0.682 \quad \hat{s.e}(\hat{\theta}) = 0.366 \\ \hat{\sigma}^2 &= 0.209 \quad \hat{s.e}(\hat{\sigma}^2) = 0.064.\end{aligned}$$

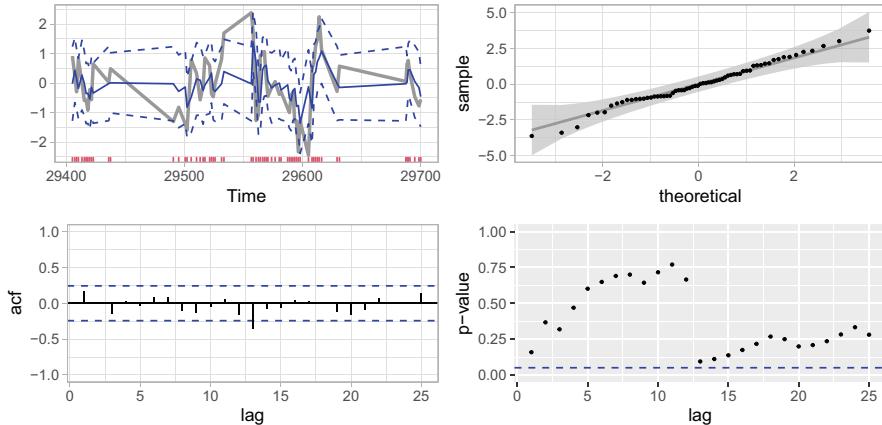


Fig. 2 On the left-top, the light curve of the Blazar object with the predicted values and their respective variability bands. For the standardized residuals: on the right-top, the quantile-quantile plot with normality reference bands [27]; on the bottom-left, the autocorrelation function estimated; on the bottom-right, the Ljung-Box test for randomness

According to this results, both the ϕ and θ parameters are significative at 10% level. Furthermore, in Fig. 2, it is shown that the residuals of the iARMA model do not hold an autocorrelation structure. In other words, the iARMA explains all the time dependence of the observed light curve. Also, the standardized residuals seem to follow a standard normal distribution.

7 Conclusions

An irregularly observed first-order autoregressive moving average model was proposed that allows treating first-order autoregressive moving averages structures with irregularly spaced times. It is established that, under Gaussianity, the model is strictly stationary and ergodic. The lowest dimension of the state-space representation along with the one-step linear predictors and its mean squared errors were given. Through the Monte Carlo study, for the ML estimation method, it is shown that bias, RMSE, and CV are smaller when N increases. Also, the method provides good estimations for the standard errors, even with relatively small sample sizes. Furthermore, the irregularly spaced times seem to increase the estimation variability. It should be noted that, despite not being presented here, the same Monte Carlo study was done for a proposed bootstrap estimation method. It showed a consistent behavior similar to what was found for the ML method. Finally, the practical application of the proposed methodology was illustrated by means of two real-life data examples involving medical and astronomical time series.

References

1. Adorf, H.M.: Interpolation of irregularly sampled data series—a survey. In: Shaw, R.A., Payne, H.E., Hayes, J.J.E. (eds.) *Astronomical Data Analysis Software and Systems IV*, ASP Conference Series, vol. 77, pp. 460–463. Astronomical Society of the Pacific (1995)
2. Babu, G.J., Mahabal, A.: Skysurveys, light curves and statistical challenges. *Int. Stat. Rev.* **84**(3), 506–527 (2016). <https://doi.org/10.1111/insr.12118>
3. Belcher, J., Hampton, J.S., Tunnicliffe Wilson, G.: Parametrization of continuous time autoregressive models for irregularly sampled time series data. *J. R. Stat. Soc. Ser. B (Methodological)* **56**(1), 141–155 (1994)
4. Bellm, E.C.: The zwicky transient facility: System overview, performance, and first results. *Publ. Astron. Soc. Pacific* **131**(995), 018002 (2018). <https://doi.org/10.1088/1538-3873/aaecbe>
5. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: *Time series analysis: forecasting and control*, 5th edn. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ (2016)
6. Brockwell, P.J., Davis, R.A.: *Time series: theory and methods*, 2nd edn. Springer Series in Statistics. Springer Science +Business Media, LLC, New York, USA (1991)
7. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16**(5), 1190–1208 (1995)
8. Corduas, M., Piccolo, D.: Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal.* **52**(4), 1860–1872 (2008). <https://doi.org/10.1016/j.csda.2007.06.001>
9. Dunsmuir, W.: A central limit theorem for estimation in gaussian stationary time series observed at unequally spaced times. *Stochast. Process. Appl.* **14**, 279–295 (1983)
10. Edelmann, D., Fokianos, K., Pitsillou, M.: An updated literature review of distance correlation and its applications to time series. *Int. Stat. Rev.* **87**(2), 237–262 (2019). <https://doi.org/10.1111/insr.12294>
11. Elorrieta, F.: Classification and modeling of time series of astronomical data. Ph.D., Pontificia Universidad Católica de Chile, Santiago de Chile (2018). <https://repositorio.uc.cl/handle/11534/22162>
12. Elorrieta, F., Eyheramendy, S., Palma, W.: Discrete-time autoregressive model for unequally spaced time-series observations. *A&A* **627**, A120 (2019). <https://doi.org/10.1051/0004-6361/201935560>
13. Eyheramendy, S., Elorrieta, F., Palma, W.: An irregular discrete time series model to identify residuals with autocorrelation in astronomical light curves. *Month. Not. R. Astron. Soc.* **481**(4), 4311–4322 (Dec 2018)
14. Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., Estévez, P.A., Sánchez-Sáez, P., Arredondo, J., Bauer, F.E., Carrasco-Davis, R., Catelan, M., Elorrieta, F., Eyheramendy, S., Huijse, P., Pignata, G., Reyes, E., Reyes, I., Rodríguez-Mancini, D., Ruz-Mieres, D., Valenzuela, C., Álvarez-Maldonado, I., Astorga, N., Borissova, J., Clocchiatti, A., Cicco, D.D., Donoso-Oliva, C., Hernández-García, L., Graham, M.J., Jordán, A., Kurtev, R., Mahabal, A., Maureira, J.C., Muñoz-Arancibia, A., Molina-Ferreiro, R., Moya, A., Palma, W., Pérez-Carrasco, M., Protopapas, P., Romero, M., Sabatini-Gacitua, L., Sánchez, A., Martín, J.S., Sepúlveda-Cobo, C., Vera, E., Vergara, J.R.: The automatic learning for the rapid classification of events (ALeRCE) alert broker. *Astron. J.* **161**(5), 242 (Apr 2021). <https://doi.org/10.3847/1538-3881/abe9bc>
15. Hamilton, J.D.: *Time Series Analysis*. Princeton University Press, Princeton, NJ (1994)
16. Harvey, A.C.: *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press (1989)
17. Illian, J., Penttinen, A., Stoyan, H., Stoyan, D.: *Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice*, Wiley (2008)
18. Jones, R.H.: Likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**(3), 389–395 (1980)

19. Jones, R.H.: Time series analysis with unequally spaced data. In: Hannan, E.J., Krishnaiah, P.R., Rao, M.M. (eds.) *Time Series in the Time Domain, Handbook of Statistics*, vol. 5, chap. 5, pp. 157–177. Elsevier Science Publishers B.V., Amsterdam, North-Holland (1985)
20. Kiliç, E.: Explicit formula for the inverse of a tridiagonal matrix by backward continued fractions. *Appl. Math. Comput.* **197**, 345–357 (2008)
21. Kim, J., Stoffer, D.S.: Fitting stochastic volatility models in the presence of irregular sampling via particle methods and the em algorithm. *J. Time Ser. Anal.* **29**(5), 811–833 (2008)
22. Koehler, E., Brown, E., Haneuse, S.J.: On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am. Stat.* **63**(2), 155–162 (2009)
23. Miller, J.I.: Testing cointegrating relationships using irregular and non-contemporaneous series with an application to paleoclimate data. *J. Time Ser. Anal.* **40**(6), 936–950 (2019)
24. Moore, M.I., Visser, A.W., Shirtcliffe, T.: Experiences with the brillinger spectral estimator applied to simulated irregularly observed processes. *J. Time Ser. Anal.* **8**(4), 433–442 (1987)
25. Muñoz, A., Carey, V., Schouten, J.P., Segal, M., Rosner, B.: A parameteric family of correlation structures for the analysis of longitudinal data. *Biometrics* **48**(3), 733–742 (1992)
26. Mudelsee, M.: Climate time series analysis: classical statistical and bootstrap methods, Atmospheric and Oceanographic Sciences Library, vol. 51, 2nd edn. Springer International Publishing (2014)
27. Nair, V.N.: Q-Q plots with confidence bands for comparing several populations. *Scand. J. Stat.* **9**(4), 193–200 (1982)
28. Ojeda, C.: Analysis of irregularly spaced time series. Ph.D., Pontificia Universidad Católica de Chile, Santiago de Chile (2019). <https://repositorio.uc.cl/handle/11534/48405>
29. Ojeda, C., Palma, W., Eyheramendy, S., Elorrieta, F.: An irregularly spaced first-order moving average model. *math.ST* (2021). <https://arxiv.org/abs/2105.06395>
30. Parzen, E.: On spectral analysis with missing observations and amplitude modulation. *Sankhyā Indian J. Stat. Ser. A* (1961–2002) **25**(4), 383–392 (1963)
31. Parzen, E. (ed.): *Time Series Analysis of Irregularly Observed Data*. Lecture Notes in Statistics, vol. 25. Springer (1984)
32. Reinsel, G.C., Wincek, M.A.: Asymptotic distribution of parameter estimators for nonconsecutively observed time series. *Biometrika* **74**(1), 115–124 (Mar 1987)
33. Robinson, P.M.: Estimation of a time series model from unequally spaced data. *Stochast. Process. Appl.* **6**, 9–24 (1977)
34. Stout, W.F.: *Almost Sure Convergence*. Probability and Mathematical Statistics, No. 24. Academic Press (1974)
35. Thornton, M.A., Chambers, M.J.: Continuous-time autoregressive moving average processes in discrete time: representation and embeddability. *J. Time Ser. Anal.* **34**(5), 552–561 (2013)
36. Wang, Z.: cts: An R package for continuous time autoregressive models via Kalman filter. *J. Stat. Softw.* **53**(5), 1–19 (2013)
37. Zhang, S.: Nonparametric bayesian inference for the spectral density based on irregularly spaced data. *Comput. Stat. Data Anal.* **151**, 107019 (2020). <https://doi.org/10.1016/j.csda.2020.107019>

Part II

Econometric and Forecasting

Using Natural Language Processing to Measure COVID-19-Induced Economic Policy Uncertainty for Canada and the USA



ShafiuIlah Qureshi, Ba Chu, Fanny S. Demers, and Michel Demers

Abstract In this paper, we develop an economic policy uncertainty (EPU) index for the USA and Canada using natural language processing (NLP) methods. Our EPU-NLP index is based on an application of several algorithms, including the rapid automatic keyword extraction (RAKE) algorithm, a combination of the RoBERTa and the Sentence-BERT algorithms, a PyLucene search engine, and the GrapeNLP local grammar engine. For comparison purposes, we also develop an index based on a strictly Boolean method. We find that the EPU-NLP index captures COVID-19-related uncertainty better than the Boolean index. Using a structural VAR approach, we find that a one-standard deviation (SD) economic policy uncertainty shock with EPU-NLP leads, both for Canada and the USA, to larger declines in key macroeconomic variables than a one SD EPU-Boolean shock. In line with the COVID-19 impact, the SVAR model shows an abrupt contraction in economic variables both in Canada and the USA. Moreover, an uncertainty shock with the EPU-NLP caused a much larger contraction for the period including the COVID-19 pandemic than for the pre-COVID-19 period.

Keywords EPU · Impulse response · NLP · BERT · Uncertainty index · COVID-19 · RoBERTa · SBERT

1 Introduction

The sudden incursion of the COVID-19 pandemic and the worldwide recession that followed have generated great interest in measuring the resulting uncertainty and its impact on macroeconomic variables. An increase in uncertainty has been

S. Qureshi (✉)

Carleton University, Department of Economics, Ottawa, ON, Canada

Department of Economics, NUML, Islamabad, Pakistan

e-mail: shafiuIlahqureshi@cmail.carleton.ca; suqureshi@numl.edu.pk

B. Chu · F. S. Demers · M. Demers

Carleton University, Department of Economics, Ottawa, ON, Canada

shown to have a very important impact on economic decisions, particularly on investment decisions, if firms face irreversibility, [6, 8, 10], fixed costs [7], or financial constraints, and also on consumption decisions when consumers are risk averse, prudent, or face binding budget constraints. Obtaining a measure of the degree of uncertainty is important for assessing its macroeconomic impact and for guiding policymakers in making appropriate monetary and fiscal policy decisions. Furthermore, policy itself may lead to uncertainty. Thus, for example, [2] investigate the impact of tax-policy uncertainty on the dynamic investment decisions of the firm. Several authors have given priority to developing an index to measure uncertainty. One prominent example is the forward-looking Baker-Bloom-Davis newspaper-based economic policy uncertainty index [3]. Other notable examples are the model-based uncertainty measures of [14] for the USA and [16] for Canada. With the COVID-19 shock as a backdrop, [1] note that while model-based measures have the benefit of being well grounded in a model in which the role and the nature of uncertainty are well-defined, such measures are essentially backward-looking and are based on the premise that the underlying model has not changed and that the statistical relationship among variables is still the same even after large and unprecedented shocks. Furthermore, the macroeconomic variables (leading indicators) in the underlying model are only available with a lag, and hence, not available in real time. In the wake of the COVID-19 shock, [1] thus point to the importance of having alternative measures of uncertainty that are *forward looking* and available in *real time*. As mentioned above, an important and very widely used measure of uncertainty is the economic policy uncertainty (EPU) index developed by [3, henceforth,BBD]. Being forward-looking in nature, the BBD-EPU newspaper-based index has been found by various authors to successfully capture uncertainty, especially policy uncertainty. Currently, an index is available for 26 countries (including the USA and Canada). The use of this index is so widespread that data providers such as Bloomberg, FRED, Haver, and Reuters also make the EPU available for users on their website. Their index has also been used in numerous economics articles since its development. We describe in detail the development of the BBD-EPU index in Sect. 2. Let us simply note here, however, that the BBD-EPU (at least the one for the USA) was very human-input intensive and expensive to develop. In this paper, we suggest an alternative newspaper-based, and (almost entirely) computer-based, approach to developing an EPU index directly related to COVID-related uncertainty for Canada and the USA, by appealing to natural language processing (NLP) techniques [11]. These techniques are widely used by Data scientists, but have not yet received much attention in economics. Our index circumvents the necessity to rely very heavily on human resources, which is less expensive and faster to obtain. These attributes make it useful for developing EPUs for country-specific policy categories and subcategories, for developing monthly or daily EPUs, and also for EPUs for countries not yet having their own BBD-EPU. We use a “text mining” approach that uses NLP to transform unstructured data (such as ordinary texts) into structured data (i.e., data or texts that are organized into categories, such as username, user ID, address, etc.) that in turn permit computers to understand, interpret, and classify human language. Our method differs markedly from a Boolean method. In contrast to the latter, our method is capable of capturing

contextual and implied meanings of EPU-related terms, thanks to our use of the RoBERTa [15] algorithm¹ which we combine with its specialization for semantic searches, SBERT [18]. To ensure greater accuracy and robustness with respect to capturing the contextual meaning of words, we also use an additional independent NLP algorithm, namely, GrapeNLP, developed by [20], which is based on Unitex-GramLab [12, 17].

In order to highlight the important difference between a Boolean method and the EPU-NLP, we also develop an alternative, strictly Boolean, index (EPU-Boolean) which we compare with our NLP-based one. We show that the EPU-NLP is better able to track COVID-19-related uncertainty than the EPU-Boolean. We also compare the EPU-NLP with other leading uncertainty indices, such as the BBD-EPU, BBD's equity market volatility (EMV) index, and the Chicago Board Options Exchange's (CBOE) volatility (VIX) index, and find that it is closely correlated with them.

We then conduct a structural vector autoregression (SVAR) analysis to observe the impact of a one-standard deviation (SD) EPU-NLP shock on some macroeconomic variables for Canada and the USA. We also compare the impact of a one-SD EPU-NLP shock on pre-COVID-19 data (January 2015–December 2019) with its impact on the data range including COVID-19 data (January 2015–October 2020). The VAR results show that a one-SD shock in the EPU-NLP index provokes a larger contraction in real GDP and other macroeconomic variables for Canada and the USA than a one-SD shock in the EPU-Boolean index. Moreover, a EPU-NLP shock results in a stronger decline in these variables for the span of time that includes the COVID-19 pandemic than the one that excludes it. The remainder of the paper proceeds as follows. Section 2 explains the development of the BBD-EPU. Section 3 presents the stages of the construction of the EPU-NLP and describes the algorithms that were used in its development. Section 4 presents the SVAR results, and Sect. 5 concludes. Figures are given in the Appendix.

2 The Development of the Baker-Bloom-Davis EPU (BBD-EPU)

As [3] explain extensively in their paper, the construction of their index was done in two stages over 2 years and involved a great deal of human resources. The authors first developed a 65-page guideline over a 6-month period. Then, under close supervision by the authors, different teams of students were trained as readers (auditors) on the basis of the guideline and were given the task of sifting through 12,000 newspaper articles to identify those that contained three terms, one from each of the following three sets: (economic *or* economy), (uncertain *or* uncertainty),

¹ RoBERTa is the “robustly optimized” version of [9]’s seminal neural network-based BERT (*Bidirectional Encoder Representations from Transformers*). We describe these algorithms below.

and *at least one* policy-related term from the third set (Congress, deficit, Federal Reserve, legislation, regulation, or White House.) [3, p. 1594]. The (human) auditors gave a coding of $EPU^H=1$ or $EPU^H=0$ depending on whether they contained the three categories of terms or not. The authors also generated a computer-based coding of articles which they compared to the human-based records to eliminate the false positives and false negatives generated by the computer-based records. They found a correlation of 0.89 between their human- and computer-generated indices for the 1989–2012 period. They also developed for the US *specialized EPU* indices for 11 different policy categories and subcategories (such as fiscal, tax, monetary, healthcare, national security, etc.) As the authors themselves indicate, the development of this index was very intensive in terms of human input and required substantial resources (i.e., it was “expensive” [3, p. 1608]). In addition, one of the reasons for the false positives and false negatives generated during their computer-based records is due to the Boolean nature of their procedure, and to words being evaluated out of context by their computer-based method. They were able to sift these out by comparing the computer-generated records with their very meticulously developed human-based records.²

3 Constructing the EPU-NLP Index: Data, Methodology, and Algorithms

We describe here the procedure used to develop our EPU-NLP index. To ensure greater accuracy and robustness, we use two NLP techniques consecutively to refine our search for relevant articles: the RoBERTa/SBERT algorithm and the GrapeNLP approach. Our motivation for using these techniques is that each of them was very successful in a Kaggle competition whose aim was to extract summary tables from the COVID-19 Open Research Dataset comprising 500,000 COVID-19-related articles. In particular, [22] used GrapeNLP to find the impact of temperature and humidity on the spread of the virus. We base ourselves on articles gathered from eight newspapers from Canada and seven newspapers from the USA from January 2015 to October 2020.³ We first enumerate the six steps that we followed and then explain the procedure in greater detail below. (1) Use the *RAKE* algorithm to search for frequently used economy, uncertainty, and policy-related words in the newspapers. (2) Select articles that contain the words obtained from step 1 using a Python filter (1,182,945 articles for Canada and 720,266 for the USA). (3) Use a

² It would seem that in the case of the BBD-EPU for countries with a native language other than English, there was a more cursory verification process and that the selection of articles was entirely Boolean in nature (see the online Appendix to [3]).

³ We used the *Calgary Herald*, the *Financial Post*, the *Montreal Gazette*, the *National Post*, the *Ottawa Citizen*, the *Toronto Star*, and the *Vancouver Sun* for Canada and *USA Today*, the *Los Angeles Times*, *The Wall Street Journal*, *The Dallas Morning News*, the *Miami Herald*, and *The New York Times* for the USA.

combination of RoBERTa and SBERT to filter out those articles having a cosine-similarity score of 0.75 or more. (We remain with 622,948 articles for Canada and 379,166 for the USA.) (4) Use a rapid Python-based Apache Lucene search engine (PyLucene) to break the short-listed articles into words and form an index containing the word-id, the number of documents in which it is present, and the exact position of the word in that document. (5) Develop a “local grammar” with Unitex/GramLab based on the keywords obtained from the previous steps and use it with the GrapeNLP Python package developed by [20] (We finally remain with 18,526 articles for Canada and 18,032 articles for the USA.). (6) Calculate the EPU-NLP index following the method indicated in BBD [3].

To highlight the benefits of using the NLP approach, we also constructed another index on the basis of the same data set, using a strictly *Boolean* approach, assigning a $EPU^{Boolean} = 1$ or 0 depending on the presence or absence of uncertainty related keywords (but without the ability to discern the context in which these keywords appear). We now explain in greater detail the RAKE, RoBERTa/SBERT, and GrapeNLP algorithms that were used in the construction of the EPU-NLP and the calculations used in the last step.

3.1 *The RAKE (Rapid Automatic Keyword Extraction) Algorithm*

RAKE is a language-independent, unsupervised ML algorithm developed by [19]. A “keyword” (also called a “token”) is defined as a sequence of one or more words. This algorithm splits text into a list of keywords, by using “stop words” (like “the,” “a,” “for,” etc.), and punctuation as a means of separating one string of contiguous words from another. These strings are candidates for keywords. The algorithm then creates a table of “co-occurrences” (i.e., words that occur together within the string) and assigns a score to each word based on its frequency of occurrence within the entire text ($\text{freq}(w)$) and also on the number of times it appears in conjunction with another word ($\text{deg}(w)$). The score assigned to a word is the ratio of $\text{deg}(w)/(\text{freq}(w))$, and the score assigned to a keyword string is the sum of the score assigned to the words composing it. We built an initial short list of simple keywords, such as uncertain, economic, recession, COVID-19, and coronavirus, as a means of initializing the RAKE algorithm and fed these words into RAKE in order to observe their frequency of occurrence in the articles. We thus produced a list of bigrams (two-word groups), and trigrams which had a high frequency of occurrence and a high score, such as *coronavirus crisis* (3218 times), *virus crisis* (3290), *economic crisis* (2029), *job losses* (2009), *virus lockdown* (2309), *global recession* (1081), and *make ends meet* (583). These terms also helped us in developing a “local grammar” as we will see below.

3.2 The BERT, RoBERTa, and SBERT Algorithms

In step 3 of our procedure, we use a combination of RoBERTa and SBERT in order to develop *sentence embeddings* and further refine our selection of articles. Word or sentence embedding is a technique in machine learning that is used to map words or phrases into vectors of real numbers. We develop a list of *queries* (full sentences) on the basis of the RAKE results, which are then processed by the SBERT and RoBERTa algorithms to extract sentence embeddings. These algorithms are extensions of the neural network-based BERT (*Bidirectional Encoder Representations from Transformers*) algorithm developed by [9] at Google. One of the important particularities of the BERT algorithm is that it is *bidirectional*. That is, it can better detect the context within which a word occurs by taking into account words that appear both *before* and *after* the keyword (i.e., both to its left and to its right) and will perform a different embedding depending on the context. Thus, for example, the following sentences given below will be embedded (encoded) differently in view of the different contextual meaning of the word *taxing*: *Bicycling up this steep hillside is very taxing on one's legs. The new policy involves taxing the rich at a higher rate.* In this respect, it can resolve ambiguities related to words having different meanings in different contexts and therefore can better avoid false positives or false negatives and better select the relevant articles. The BERT algorithm is pre-trained by using a technique called *masked learning modeling* which essentially “hides” (or “masks”) 15% of the keywords in each query by replacing them with another token or mask and requiring the algorithm to predict the true keywords.⁴ The RoBERTa algorithm developed by [15] is a “robustly optimized” version of the BERT algorithm that uses a much larger training data set (160G instead of 16G). The RoBERTa algorithm presents several other advantages over the BERT algorithm. Thus, for example, it uses dynamic “masking” as opposed to the static masking used in BERT. The dynamic masking pattern of RoBERTa implies that the pattern of masking changes with each sequence that is fed into the algorithm, whereas in BERT, the pattern is set at the initial sequence and remains fixed throughout. The RoBERTa algorithm requires fine-tuning depending on the application, and we fine-tuned it by using unlabeled newspaper articles and using the “hugging face” open-source NLP platform.

The SBERT algorithm [18] is a further refinement of BERT that is better suited for semantic searches and that uses two identical (“siamese”) sub-networks (both of them RoBERTa algorithms) so as to better compare two sentence (or document) embeddings. It then applies the cosine-similarity measure to assess the degree of similarity between sentence or document pairs.⁵ In the third step of our procedure,

⁴ To be more precise, 10% of the masked terms are replaced with randomly selected keywords, 10% are replaced with the true word, and 80% are replaced with the token [MASK].

⁵ The cosine-similarity measure is given by $\sigma(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2}$ where A and B are n dimensional vectors and $\|\cdot\|_2$ is L^2 norm. The cosine-similarity measure $\sigma(A, B)$ takes the value 1 when the two vectors are exactly the same and the value -1 when they are completely dissimilar. Comparing this

we pair the articles obtained in step 2 with sentences from the queries that we developed on the basis of the RAKE results and feed them into SBERT. After a pooling process⁶ needed to transform the contextual embeddings obtained from each RoBERTa sub-network into vectors of fixed length, SBERT calculates the cosine-similarity measure between the articles and the queries. We choose articles that have a cosine-similarity measure of 0.75 or greater.

3.3 *GrapeNLP Grammar*

In step 5, we use GrapeNLP grammar to further refine our choice of articles. As mentioned above, we follow [22] who used this approach (developed in [20]) to extract research papers from the *COVID-19 Open Research Dataset*. We build a local grammar using the Unitex grammar editor⁷ [17]. We then use GrapeNLP to convert the grammar to a form that may be processed by Python. This approach involves a “human-assisted” training of grammar. We review the results of an initial trial and then change the “grammar” accordingly. While space constraints preclude us from providing the extensive grammars that we have developed, we illustrate the concept by using an example given in [21]. This example (shown in Fig. 1 in the Appendix) is a grammar that is built to recognize sentences that may be used by someone requesting to make a phone call. The sentence may take different forms: I (want/would like) (to) (call) (<E>a/my/the) (TOKEN) or (phone number). The box containing an <E> is optional, in the sense that none of the terms in that box need to be present (e.g., I want to call 911). Boxes without an <E> are compulsory and at least one of the words in that box must be present. The box labeled TOKEN may contain any name (e.g., Mary, mother, emergency, etc.), while the box (phone number) is a subgrammar (i.e., another grammar that is evoked by this one) which recognizes phone numbers and which may contain symbols such as + in the case of country codes, or parentheses, etc. Alternatively, the sentence might take the form “Could you call my sister, please?” We again note that the comma and the word “please” are optional.

For our purpose of finding phrases expressing policy or COVID-19-related uncertainty in the newspaper articles, we adapt this methodology by building four grammars (which are linked to one another) to find phrases such as “economic

similarity measure to cross-entropy or to mean-squared error (which uses Euclidean distance as its measure of closeness), this measure has the advantage of being dependent only on the direction of the vectors and not on their magnitudes and, hence, is independent of the scaling of the two vectors.

⁶ We use the RoBERTa model to map tokens in a sentence to the contextual word embeddings from RoBERTa. The next layer in our model consists of averaging (“mean-pooling”) all contextualized word embeddings obtained from RoBERTa. In other words, each sentence is passed first through the word_embedding_model (in Roberta) and then through the pooling_model to give fixed-sized vectors. Vectors of fixed length are required by SBERT.

⁷ Unitex/GramLab is an open-source, cross-platform, multilingual, lexicon- and grammar-based corpus processing tool. It can be downloaded from <https://unitexgramlab.org/>.

uncertainty caused by the coronavirus lockdown” or more complex ones such as “During the prolonged period of the coronavirus crisis, targeted transfers are urgently needed to stay above the poverty line.” We use the frequently encountered terms that were selected by RAKE in developing these grammars.⁸

3.4 Calculating the EPU-NLP

In step 6, we follow [3, p.1599]’s method to calculate the EPU. Let $c_{i,t}$ $i = 1, \dots, N$, $t = 1, \dots, T$ denote the raw count of articles found to be uncertainty-related in newspaper i in month t and let $Total_{it}$ be the *total* number of articles in newspaper i in month t . The scaled count is then given by C_{it}^* where $C_{it}^* = \frac{c_{it}}{Total_{it}}$, $i = 1, \dots, N$; $t = 1, \dots, T$. Compute the standard deviation of scaled counts as $\sigma_i^2 = (1/T) \sum_{t=1}^T (C_{it}^* - \bar{C}_i)^2$ where $\bar{C}_i = (1/T) \sum_{t=1}^T C_{it}^*$ is the average over the entire period of the scaled counts of articles in newspaper i . Divide the scaled counts by the standard deviation : $Y_{it} = C_{it}^*/\sigma_i$, $i = 1, \dots, N$; $t = 1, \dots, T$ (Thus, e.g., $Y_{it} = 2$ would indicate that this scaled count obtained in month t is two standard deviations above the mean for newspaper i for the entire time period and would point to a period of higher uncertainty.). Compute the $Z_t = (1/N) \sum_{i=1}^N Y_{it}$, where Z_t is the average of the scaled standardized counts over all newspapers for month t . Calculate $M = (1/T) \sum_{t=1}^T Z_t$, where M is the average of the scaled standardized counts over all newspapers and for all months in the data set. Calculate the normalized EPU time-series index as $EPU_t^{NLP} = \left(\frac{100}{M}\right) Z_t$. With this normalization, the EPU-NLP has a mean of 100.

4 Testing the Model

We adopt a structural vector autoregression approach (SVAR) to test our EPU-NLP index and our EPU-Boolean index for both Canada and the USA using data from January 2015 to October 2020. As in [4] and [1], we detrend all the variables using the Hamilton filter [13] and then take the first difference of the log of these variables.

Even though VAR models may not be used to establish causality, as [1] note, they can indicate whether uncertainty shocks are precursors to a slowdown in economic activity, such as a fall in GDP and employment. Using a vector autoregression analysis, they find significant contractions in economic variables during COVID-19 for both the USA and the UK. Baker et al. [4] use stock market volatility, newspaper-

⁸ We used the GPU accelerator offered by Google Colab Pro for data cleaning, semantic search, and RAKE (Python), the Nvidia Tesla P100 GPU offered by the Kaggle platform for the grammar, and an Intel Xeon i9-7980Xe 36-core server with Nvidia Titan V GPU and 126GB DDR4 RAM for fine-tuning the RoBERTa model.

based economic uncertainty, and subjective uncertainty in business expectation surveys to measure the COVID-19-induced uncertainty for the USA. As mentioned earlier, [16] constructed an uncertainty measure for Canada by applying the method of [14] and assessed the impact of COVID-induced uncertainty on economic variables using a SVAR analysis for Canada. To indicate the advantages of a SVAR, let us start with a standard VAR model which in our case may be described as follows:

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (1)$$

where \mathbf{y}_t is a 5×1 vector of four macroeconomic variables and one uncertainty index and p denotes the number of lags. The \mathbf{A}_i , $i = 1, \dots, p$ are 5×5 matrices of parameters, while $\boldsymbol{\varepsilon}_t$ is a 5×1 vector of innovations with $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_s) = \mathbf{0}$ for all $s \neq t$. Equation (1) is a standard VAR which describes a reduced form model and which does not allow contemporaneous effects of the endogenous variables on each other. It also has the underlying counterfactual assumption that the innovations of the different equations are mutually uncorrelated. Since the innovations are in fact correlated in our model (as an analysis of the covariance matrix $\boldsymbol{\Sigma}$ reveals), a shock to one variable will have an impact on the innovation of another variable, precluding a clear interpretation of the impulse responses. We therefore adopt a SVAR approach and use a Cholesky decomposition in order to identify the shocks. Our equation may be written as:

$$\mathbf{A}(\mathbf{I}_5 - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p) \mathbf{y}_t = \mathbf{B} \boldsymbol{\varepsilon}_t \quad (2)$$

where \mathbf{A} is a lower triangular matrix with ones in the diagonal, \mathbf{B} is a diagonal matrix, and $\boldsymbol{\varepsilon}_t$ is a 5×1 vector of orthogonalized innovations with $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I}_5)$ and $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_s) = 0$ for all $s \neq t$ such that $\mathbf{B} \boldsymbol{\varepsilon}_t \equiv \mathbf{A} \boldsymbol{\varepsilon}_t$. The structure of the matrices \mathbf{A} and \mathbf{B} is set in accordance with the order of the variables in the VAR. The order of the variables matters for the results. The variable that is listed first is assumed to have a contemporaneous impact on the rest of the variables, while none of the other variables has a contemporaneous impact on the first. Similarly, each of the variables will have a contemporaneous impact on the rest of the variables that are listed after it, but will not be affected by them contemporaneously. In our case, since we wish to analyze the impact of a EPU shock on the macroeconomic variables and since our uncertainty measures are news-based, it seems reasonable to consider them to be exogenous and to order them first. Consequently, we adopt the following order of the variables for Canada : EPU-NLP (or EPU-Boolean), TSX, employment, industrial production, and GDP. For the USA, the order of the variables is EPU-NLP (or EPU Boolean), S&P 500, employment, industrial production, and consumption.^{9,10}

⁹ We use industrial production at the monthly frequency as a proxy for real GDP in the case of the USA since real GDP is not available on a monthly basis for the USA.

¹⁰ Since our primary purpose is not to make a comparison of the EPU-NLP^{Canada} and EPU-NLP^{USA} indices, using different variables is not consequential.

In accordance with the optimal lag criteria SBIC and HQIC, we choose three lags for Canada and one lag for the USA.

Before analyzing the impulse response functions, we first present a comparison of the EPU-NLP and EPU-Boolean. As is shown in Fig. 2, EPU-NLP better captures the large increase in uncertainty in March and April 2020 due to COVID-19 compared to EPU-Boolean for both Canada and the USA.

We then estimate the response of the economic variables to an uncertainty shock as captured by a one-standard deviation (SD) innovation in EPU-NLP and EPU-Boolean, respectively, for Canada (Fig. 4) and for the USA (available upon request) for the period including COVID-19 (January 2015–Oct 2020). The Canada SVAR results indicate that a one-SD EPU-NLP shock leads to declines of 1.04% in real GDP, 0.95% in employment, 1% in industrial production, and 1.08% in TSX, respectively. By contrast, a one-SD shock with EPU-Boolean results in declines of only 0.42% in real GDP, 0.41% in employment, 0.33% in industrial production, and 1.02% in TSX. Similarly for the USA, we find that a one-SD EPU-NLP shock results in a 0.90% drop in industrial production, 0.70% in real consumption, 0.83% in employment, and 2.1% in S&P 500. On the other hand, one-SD shock to uncertainty with EPU-Boolean provokes only a 0.19% fall in industrial production, 0.11% in real consumption, 0.16% in employment, and 0.60% in S&P 500. Hence, for both the USA and Canada, we observe a much less pronounced response to a one-SD shock when we use the EPU-Boolean instead of the EPU-NLP.

We also experimented with different orders of the macroeconomic variables (while keeping the EPU variable as the first) but did not observe any difference in the results. In addition, in view of the small sample size, we also conducted a Bayesian VAR analysis using a conjugate Minnesota (multivariate normal) prior distribution for the regression coefficients and an inverse-Wishart prior distribution for the error covariance as recommended by [5]. While space constraints prevent us from reporting these results, we observed that they are qualitatively analogous to the ones for the SVAR, and in particular, a comparison of the EPU-NLP and EPU-Boolean shocks again indicates that a EPU-NLP shock has a greater impact on the macro variables.

Next, with the SVAR approach, we follow [1] and compare the impact of a one-SD shock with our EPU-NLP index for the period *including* COVID-19 (January 2015–October 2020) with the *pre-COVID-19* period (January 2015–December 2019) for Canada (Fig. 5) and for the USA (available upon request). These results are striking. For the period including COVID-19, a one-SD innovation with EPU-NLP leads to declines of 1.04% in real GDP, 1% in industrial production, 0.95% in employment, and 1.08% in TSX, respectively. By contrast, for the pre-COVID-19 period, a one-SD innovation with EPU-NLP results in no change in real GDP, 0.025% in employment, 0.6% in TSX, and 0.17% in industrial production. We obtain similar results for the USA: For the period including COVID-19, one-SD shock to uncertainty (with EPU-NLP) results in a drop of 0.90% in industrial production, 0.70% in real consumption, 0.83% in employment, and 2.1% in S&P 500. By contrast, for the pre-COVID-19 period, one-SD shock to uncertainty (with EPU-NLP) provokes a fall of 0.06% in industrial production, 0.05% in real

consumption, 0.015% in employment, and 0.34% in S&P 500. In other words, for both the USA and Canada, we observe a less pronounced response to a one-SD shock to uncertainty (with EPU-NLP) for the *pre-COVID-19 period*. Hence, the EPU-NLP index is able to capture the COVID-19-induced uncertainty and its severe negative impact on economic variables.

We also compared EPU-NLP to other uncertainty measures such as VIX and BBD-EPU. As is shown in Fig. 3, EPU-NLP and VIX closely follow each other with matching peaks and troughs during the COVID-19 period for Canada and the USA. We find a correlation of 0.85 between EPU-NLP^{Canada} and VIX, and of 0.80 between EPU-NLP^{USA} and VIX, whereas [3] found a correlation of 0.58 between the BBD-EPU and VIX. As they explain, their EPU is more specialized in policy uncertainty as opposed to financial uncertainty captured by the VIX. They therefore developed an EMV (equity market volatility) index that better captures financial uncertainty. In our case, in view of our search words, the EPU-NLP may be more attune to generalized uncertainty, and COVID-19 uncertainty in particular. This may be an explanation of its closer correlation with the VIX. We find a correlation of 0.79 between BBD's EMV index and our EPU-NLP^{Canada} and of 0.70 between the EMV index and EPU-NLP^{USA}. The correlation between the EPU-NLP^{USA} index and BBD-EPU^{USA} is 0.85 and is 0.72 between EPU-NLP^{Canada} and BBD-EPU^{Canada}.

5 Conclusion

This paper described a new approach based on NLP techniques for constructing an EPU index based on newspaper articles. For this purpose, we use the RAKE, RoBERTa/SBERT, and GrapeNLP algorithms. RAKE is used to determine high-frequency words or phrases related to policy uncertainty and COVID-19, which are then used to filter articles and develop search queries and local grammars. We use the RoBERTa algorithm which is pre-trained on a large new dataset (CC-News). We fine-tune it on our own newspaper data and combine it with SBERT which is better suited for semantic searches. Finally, we use the GrapeNLP grammar engine to select the final EPU-related articles on the basis of which we calculate our EPU-NLP index.

We compare the EPU-NLP index with a EPU-Boolean index which we construct on the basis of the same dataset using a strictly Boolean approach. We observe that EPU-NLP better captures the COVID-19-induced uncertainty than the EPU-Boolean. We also compare the EPU-NLP with other leading uncertainty indices and find that it is closely correlated with several of them (BBD-EPU, EMV, and VIX). We further assessed the impact of EPU-NLP and EPU-Boolean using a VAR model with Canadian and US economic variables. We found that EPU-NLP created a greater dip in economic variables compared to EPU-Boolean. Lastly, we compare the impact of a one-SD EPU-NLP shock on pre-COVID-19 data (January 2015–Dec. 2019) with its impact on the entire data range including COVID-19 data (January 2015–October 2020). The VAR results showed once more that EPU-NLP

generated a greater dip in economic variables for Canada and the USA for the period including COVID-19 than for the pre-COVID-19 period. The proposed method can be employed to construct a high-frequency EPU index which can then be used to predict financial variables. This point is left for future research.

Acknowledgments This research is partially funded by a Carleton University FPA Research Engagement Grant. The authors sincerely thank three anonymous referees for comments, and Javier Sastre, Asma Djaidri, and Raheel Ahmed for their support.

Appendix

See Figs. 1, 2, 3, 4, 5.

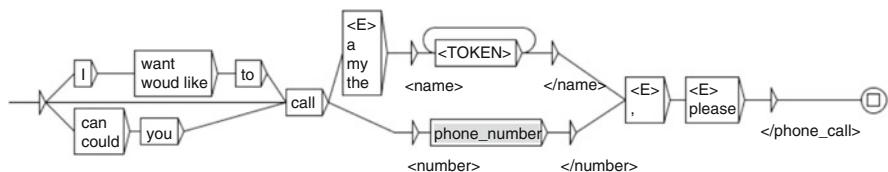


Fig. 1 A GrapeNLP grammar diagram

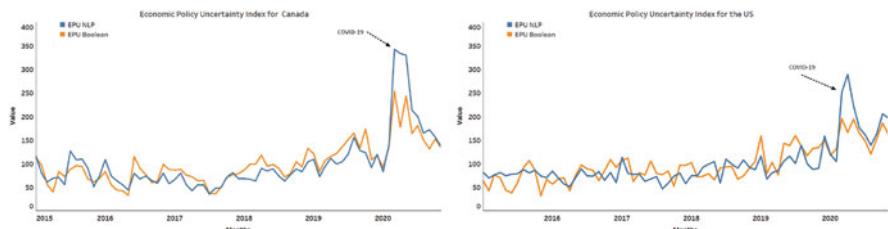


Fig. 2 EPU-NLP vs EPU-Boolean for Canada (left subfigure) and the USA (right subfigure)

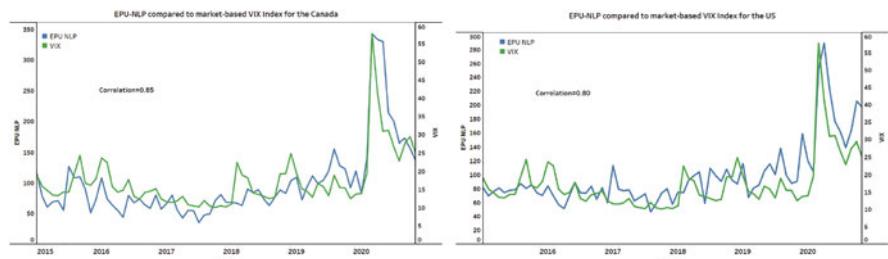


Fig. 3 EPU-NLP vs VIX for Canada and the USA

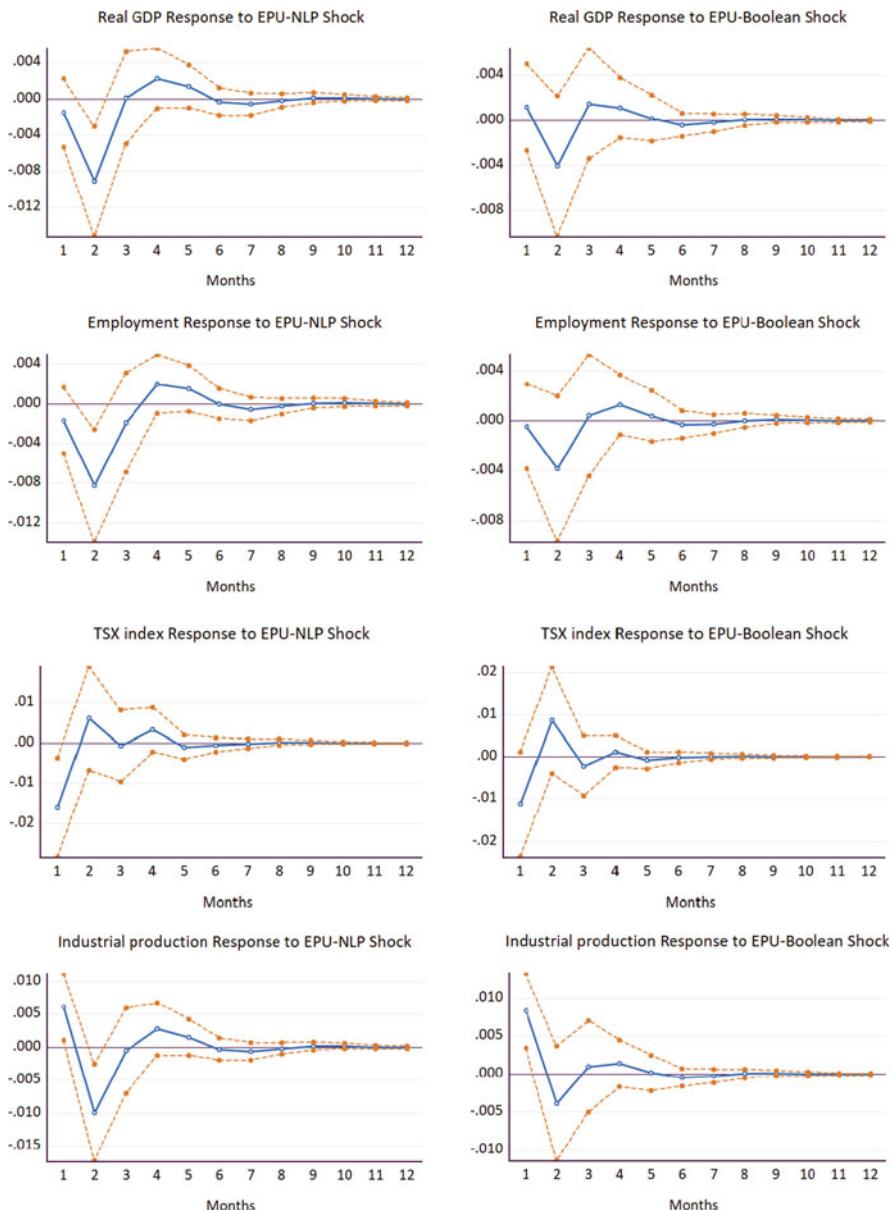


Fig. 4 SVAR results, Canada: Comparing EPU-NLP and EPU-Boolean

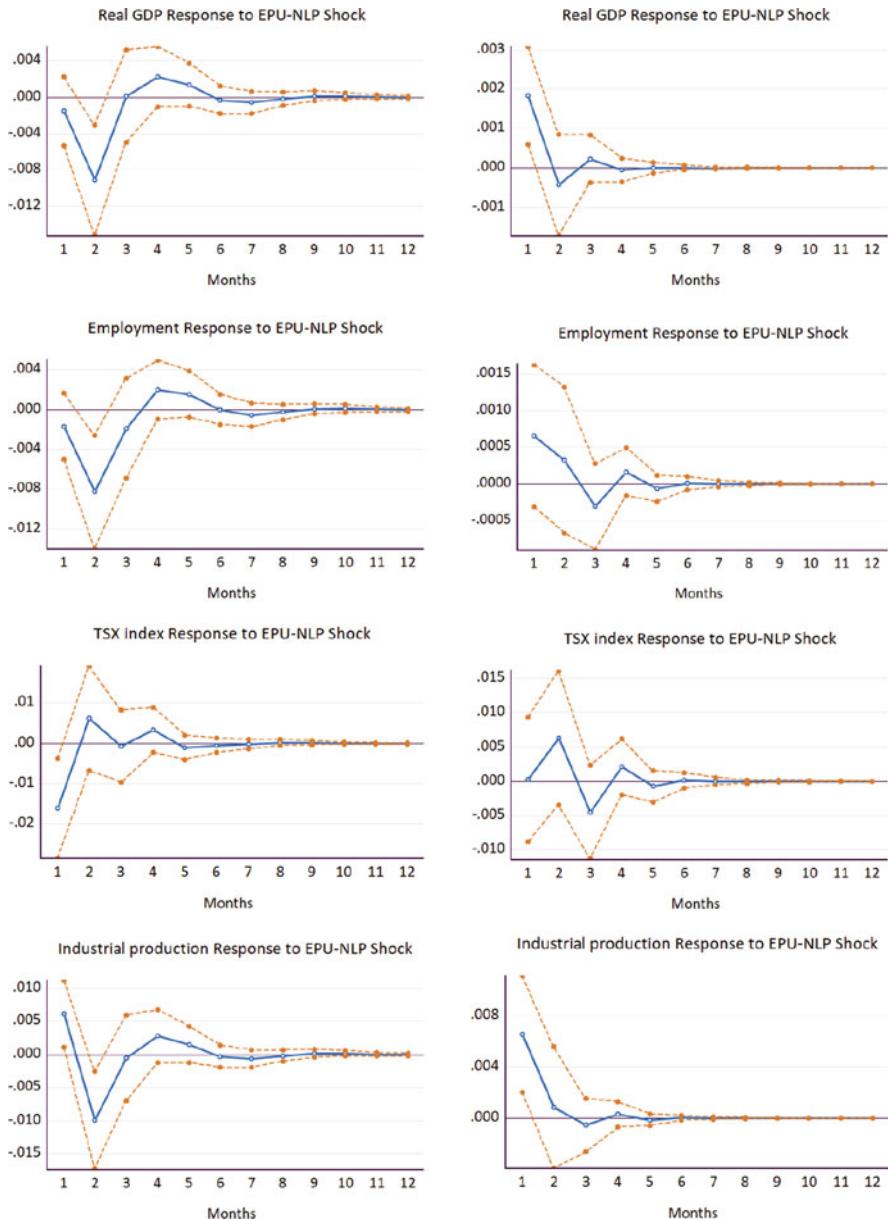


Fig. 5 EPU-NLP shock for Canada: period including COVID-19 (left panel) vs pre-COVID-19 period (right panel)

References

1. Altig, D., Baker, S., Barrero, J.M., Bloom, N., Bunn, P., Chen, S., Davis, S.J., Leather, J., Meyer, B., Mihaylov, E., et al.: Economic uncertainty before and during the covid-19 pandemic. *J. Public Econ.* **191**, 104274 (2020)
2. Altug, S., Demers, F.S., Demers, M.: The investment tax credit and irreversible investment. *J. Macroecon.* **31**(4), 509–522 (2009)
3. Baker, S.R., Bloom, N., Davis, S.J.: Measuring economic policy uncertainty. *Q. J. Econ.* **131**(4), 1593–1636 (2016)
4. Baker, S.R., Bloom, N., Davis, S.J., Terry, S.J.: Covid-induced economic uncertainty. Technical report, National Bureau of Economic Research (2020)
5. Banbura, M., Giannone, D., Reichlin, L.: Large bayesian vars. Technical report, European Central Bank (2008)
6. Bernanke, B.S.: Irreversibility, uncertainty, and cyclical investment. *Q. J. Econ.* **98**(1), 85–106 (1983)
7. Caballero, R.J., Engel, E.M.: Explaining investment dynamics in us manufacturing: a generalized (s, s) approach. *Econometrica* **67**(4), 783–826 (1999)
8. Demers, M.: Investment under uncertainty, irreversibility and the arrival of information over time. *Rev. Econ. Stud.* **58**(2), 333–350 (1991)
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint (2018). arXiv:1810.04805
10. Dixit, A., Pindyck, R.: *Investment Under Uncertainty*. Princeton University Press, Princeton, NJ (1994)
11. Gentzkow, M., Kelly, B., Taddy, M.: Text as data. *J. Econ. Lit.* **57**(3), 535–574 (2019)
12. Gross, M.: The construction of local grammars. In: Roche, E., Shabes, Y. (eds.) *Finite-State Language Processing*. MIT Press, Cambridge, MA (1997)
13. Hamilton, J.D.: Why you should never use the hodrick-prescott filter. *Rev. Econ. Stat.* **100**(5), 831–843 (2018)
14. Jurado, K., Ludvigson, S.C., Ng, S.: Measuring uncertainty. *Am. Econ. Rev.* **105**(3), 1177–1216 (2015)
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. Preprint (2019). arXiv:1907.11692
16. Moran, K., Stevanović, D., Touré, A.K.: Macroeconomic Uncertainty and the Covid-19 Pandemic: Measure and Impacts on the Canadian Economy. CIRANO (2020)
17. Paumier, S., Nakamura, T., Voyatzis, S.: Unitex, A Corpus Processing System with Multi-Lingual Linguistic Resources. In: eLEX2009, vol. 173 (2009)
18. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. Preprint (2019). arXiv:1908.10084
19. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. *Text Mining Appl. Theory* **1**, 1–20 (2010)
20. Sastre, J.: Efficient Finite-State Algorithms of Application of Local Grammars. Ph. D. thesis (2011)
21. Sastre, J.: Grapenlp grammar engine in a kaggle notebook (2020). Available at <https://www.kaggle.com/javiersastre/grapenlp-grammar-engine-in-a-kaggle-notebook>
22. Sastre, J., Vahid, A.H., McDonagh, C., Walsh, P.: A text mining approach to discovering covid-19 relevant factors. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 486–490. IEEE (2020)

Asymptotic Expansions for Market Risk Assessment: Evidence in Energy and Commodity Indices



Daniel Velásquez-Gaviria Andrés Mora-Valencia and Javier Perote

Abstract The increasing volatility experienced in financial and commodity markets has motivated the search of frequency functions with more complex attributes to characterize their asset returns distribution. In this research, two semi-nonparametric distributions are proposed and compared, the Gram-Charlier expansion and a novel Edgeworth expansion for the Student's t, to estimate the value-at-risk and the expected shortfall in four indices related to energy, metals, mining, and physical commodities. Backtesting performance is assessed in terms of Kupiec and Independence tests for value at risk and the recent proposal by Acerbi and Székely for the expected shortfall. Our results indicate that the Student's t expansion density adequately fits the returns of different indices and exhibits the best performance for value at risk and expected shortfall backtesting. Consequently, the Student's t expansion density, which encompasses the Gram-Charlier distribution as the degrees of freedom parameter tends to infinity, reveals as a flexible and accurate methodology for risk management purposes in energy and commodity markets.

Keywords Commodity and energy markets · Edgeworth expansion · Student's t distribution · Value-at-Risk · Expected shortfall · Backtesting

D. Velásquez-Gaviria

School of Business and Economics, Maastricht University, Maastricht, Netherlands
e-mail: d.velasquezgaviria@maastrichtuniversity.nl

A. Mora-Valencia

School of Management, Universidad de los Andes, Bogotá, Colombia
e-mail: a.mora262@uniandes.edu.co

J. Perote

School of Economics and Business, University of Salamanca, Salamanca, Spain
e-mail: perote@usal.es

1 Introduction

From a parametric perspective, the non-normality of financial returns has been traditionally tackled by proposing distributions capable of featuring thick tails and asymmetries by adding additional parameters. For instance, Student's t, inverse Gaussian, hyperbolic, exponential, gamma, Weibull distributions, among many others, have been generalized with many different parametrizations. The extensions of these distributions include an entire family of Student's t (see, e.g., [1–3]; Jones and Faddy [4], [5, 6]), but also generalized inverse Gaussian [7], generalized hyperbolic [8, 9], generalized exponential [10], generalized Weibull [11], or generalized gamma [12] have been widely studied.

A direct and rigorous framework to generalize (continuous and differentiable) probability density functions (pdfs) is the semi-nonparametric approach (SNP) method that expands parametric pdfs in terms of their derivatives, which allow asymptotically approximating any frequency function. Such expansions arise from the early work of Edgeworth [13], who provided a former GC Type A series of orthogonal polynomials, named Hermite polynomials, derived from the expansion of a Gaussian probability density function. The SNP expansions incorporate an arbitrary degree of flexibility that allows to model the moments of the returns and provide outstanding performance for risk assessment, although at the cost of some instability in the density that might result in negative values when finite expansions are considered. This problem, stated by Barton and Dennis [14], has been tackled by studying the positivity regions [15] or imposing positivity transformations [16]. However, from an empirical perspective, it only requires that maximum likelihood (ML) algorithms converge to global optima.

In this research, we compare two SNP distributions, the traditional Gram-Charlier expansion (GCE) and the novel expansion on Student's t basis, STE, whose Hermitian-type of polynomials are arbitrarily derived up to the eighth term. The polynomials of the STE are more complex than those of the GCE since they depend on the degrees of freedom parameter. Nevertheless, the series converges to the GCE as this parameter goes to infinity. This property makes the STE a valid asymptotic expansion more flexible than the GCE, meaning that similar data fits might be obtained through shorter expansions. The estimation of the STE expansion, however, seems more challenging and computationally demanding. The empirical results for daily data of four indices on energy and commodity indices indicate that both expansions seem to be accurate representations of these indices' daily returns since the 90s. However, the STE seems to provide better risk assessment measures in terms of both VaR and ES, thus being an accurate tool for risk management.

The rest of this work is divided as follows. Section 2 presents the methodology and introduces our proposed STE density. Section 3 analyzes the VaR and ES backtesting results for four energy and commodity indices, and Sect. 4 summarizes the main conclusions.

2 Methodology

2.1 Gram-Charlier Expansion

The traditional theory for SNP density modeling lies in the construction of orthogonal polynomials. According to Abramowitz and Stegun [17], a system of s polynomials, $P_s(x)$, is orthogonal to a weight function $\omega(x)$, if

$$\int_{-\infty}^{\infty} P_s(x) P_j(x) \omega(x) dx = 0, \quad \forall \neq j, s, j = 0, 1, 2, \dots, \quad (1)$$

The Hermite polynomials consider the standard normal pdf, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, as the weight function and form a natural basis for defining a family of pdfs. These polynomials emerge from the s -th order derivative of the weight function and thus $H_s(x)$ are obtained by solving Eq. (2)

$$H_s(x) = (-1)^s \phi(x)^{-1} \frac{d^s \phi(x)}{dx^s}, \quad (2)$$

$$\begin{aligned} \text{e.g. } H_1(x) &= x, & H_2(x) &= x^2 - 1, & H_3(x) &= x^3 - 3x, \\ H_4(x) &= x^4 - 6x^2 + 3, & H_5(x) &= x^5 - 10x^3 + 15x, \\ H_6(x) &= x^6 - 15x^4 + 45x^2 - 15, & H_7(x) &= x^7 - 21x^5 + 105x^3 - 105x \text{ and} \\ H_8(x) &= x^8 - 28x^6 + 210x^4 - 420x^2 + 105. \end{aligned} \quad (3)$$

Based on these polynomials, the GCE can be expressed as

$$f_n(x, \mathbf{d}_s) = \left[1 + \sum_{s=1}^n d_s H_s(x) \right] \phi(x). \quad (4)$$

where d_s is directly related to the s -th order cumulant (or moment) of the pdf $f(x)$. This GCE “truncated” at the order n is a well-defined pdf in virtue of the orthogonality condition in Eq. (1) and provided that $f_n(x, \mathbf{d}_s) \geq 0$. The (central) moment of order s of the GCE, $\mu_s = E[x^s]$, can be expressed as a linear function of the first s even (odd) parameters, in case s is even (odd)—see Kendall and Stuart [18] for this and other properties of the GCE and Trespalacios, Cortés, and Perote [19] for a recent application to energy markets. Further characterizations of the GCE in terms of the moment generating function are possible (see [20]). For quantile computation (q), exists a closed form for the cumulative distribution function (cdf)—see, e.g., Cortés, Mora-Valencia, and Perote [21].

$$F_n(x, \mathbf{d}_s) = \int_{-\infty}^q \phi(x) dx - \phi(q) \sum_{i=1}^s d_i P_{i-1}(q). \quad (5)$$

2.2 Student's t Expansion

The pdf of a random variable x that is distributed as Student's t with $v > 2$ degrees of freedom— $\Gamma(\cdot)$ stands for the gamma function—is given by

$$t_v(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{v+1}{2}}. \quad (6)$$

According to the theory on GCE series, a direct expansion can be defined by considering $t_v(x)$ as the weight function and computing orthogonal polynomials on the basis of its derivatives [22], i.e.

$$P_s(x) = \frac{(-1)^s}{t_v(x)} \cdot \frac{d^s t_v(x)}{dx^s} \quad (7)$$

Therefore for a vector $\delta'_s = (\delta_1, \delta_2, \dots, \delta_s)$ such that $t_n(x, v, \delta_s) \geq 0$, the STE with v degrees of freedom and $P_s(x)$ can be characterized in terms of the pdf

$$t_n(x, v, \delta_s) = \left[1 + \sum_{s=1}^n \delta_s P_s(x) \right] t_v(x), \quad (8)$$

where $t_v(x)$ is the Student's t pdf in Eq. (6) and $P_s(x)$ the s -th order orthogonal polynomial in Eq. (7), particularly the first eight polynomials are:

$$P_1(x) = \frac{x(v+1)}{x^2+v}, \quad P_2(x) = \frac{(v+1)(x^2(v+2)-v)}{(x^2+v)^2},$$

$$P_3(x) = \frac{(v+1)(v+3)(x^3(v+2)-3xv)}{(x^2+v)^3},$$

$$P_4(x) = \frac{(v+1)(v+3)(x^4(v+4)(v+2)+3v^2-6x^2v(v+4))}{(x^2+v)^4},$$

$$P_5(x) = \frac{(v+1)(v+3)(v+5)(x^5(v+4)(v+2)+15xv^2-10x^3v(v+4))}{(x^2+v)^5},$$

$$P_6(x) = \frac{(v+1)(v+3)(v+5)\left(x^6(v+6)(v+4)(v+2)+45x^2v^2(v+6)-15v^3-15x^4v(v+6)(v+4)\right)}{(x^2+v)^6},$$

$$\begin{aligned}
P_7(x) &= \frac{(v+1)(v+3)(v+5)(v+7)}{\left(x^7(v+6)(v+4)(v+2)+105x^3v^2(v+6)\right.} \\
&\quad \left.-105xv^3-21x^5v(v+6)(v+4)\right)} \quad \text{and} \\
P_8(x) &= \frac{(v+1)(v+3)(v+5)(v+7)}{\left(x^8(v+8)(v+6)(v+4)(v+2)+210x^4v^2(v+8)(v+6)+105v^4\right.} \\
&\quad \left.-420x^2v^3(v+8)-28x^6v(v+8)(v+6)(v+4)\right)} \\
&\quad \frac{(x^2+v)^8}{(9)}
\end{aligned}
\tag{9}$$

As in the case of the GCE, the even (odd) moment of order s depends linearly on the first s even (odd) parameters. For instance, the first eight moments are given by:

$$\mu_1 = \delta_1, v > 1, \quad \mu_2 = \frac{v}{v-2} + 2\delta_2, v > 2, \quad \mu_3 = 6\delta_3 + 3\delta_1 \frac{v}{v-2}, v > 3,$$

$$\begin{aligned}
\mu_4 &= 3 \left[\frac{v^2}{v^2 - 6v + 8} + 4\delta_2 \frac{v}{v-2} + 8\delta_4 \right], v > 4, \\
\mu_5 &= 15 \left[\frac{v [4\delta_3(v-4) + v\delta_1]}{(v-4)(v-2)} + 8\delta_5 \right], v > 5,
\end{aligned}$$

$$\begin{aligned}
\mu_6 &= \frac{15\Gamma\left(\frac{v}{2}-3\right)[48\delta_6(v-6)(v-4)(v-2) \\
&\quad + 24\delta_4v(v-6)(v-4) + 6v^2(v-6) + v^3]}{8\Gamma\left(\frac{v}{2}\right)}, v > 6,
\end{aligned}$$

$$\begin{aligned}
\mu_7 &= \frac{105[\delta_1v^3 + 6(v-6)[\delta_3v^2 + 4(v-4)[\delta_5v + \delta_72(v-2)]]]}{(v-6)(v-4)(v-2)}, v > 7 \text{ and} \\
\mu_8 &= \frac{\Gamma\left(\frac{v}{2}-4\right)\left[v^4 + 147456\delta_8 + 8v[(v-8)\right.} \\
&\quad \left.\left[v^2\delta_2 + 6(v-6)[v\delta_4 + 4\delta_6(v-4)]\right]\right. \\
&\quad \left.+ 48\delta_8(v-10)[40 + (v-10)v]\right]}{2}, v > 8.
\end{aligned}
\tag{10}$$

Note that unlike the GCE, where moments of all order exist, in the STE the existence of moments depends directly on the value of the degrees of freedom. This shortcoming is partially solved by the fact that in practical applications, degrees of

freedom of the STE seems to increase from those of the Student's t. Furthermore, the GCE is nested on the STE since, by construction, it follows that as $v \rightarrow \infty$.

$$t_n(x, v, \delta_s) = \left[1 + \sum_{s=1}^n \delta_s P_s(x) \right] t_v(x) \rightarrow \left[1 + \sum_{s=1}^n \delta_s H_s(x) \right] \phi(x) = f_n(x, \mathbf{d}_s) \quad (11)$$

since $P_s(x) \rightarrow H_s(x)$ and $t_v(x) \rightarrow \phi(x)$.

In addition, the cdf of the STE can be obtained similarly as the cdf of the GCE in Eq. (5)

$$T_n(x, v, \delta_s) = \int_{-\infty}^q t_v(x) dx - t_v(q) \sum_{s=1}^n \delta_s P_{s-1}(q). \quad (12)$$

2.3 Model and Maximum Likelihood Estimation

Asset returns use to exhibit a small predictable component in conditional mean, but clusters and long memory in conditional volatility. Consequently, an AR(1)-GARCH(1,1) has traditionally been used for capturing such behavior. Therefore, we consider that asset returns, r_t , can be modeled as:

$$r_t = \varphi + \phi r_{t-1} + \varepsilon_t, \quad (13)$$

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad (14)$$

$$z_t = \frac{\varepsilon_t}{\sigma_t}, \quad z_t \sim \text{iid } G(\boldsymbol{\theta}), \quad (15)$$

where $-1 < \phi < 1$, $\varphi > 0$, $\alpha > 0$, $\beta > 0$, $\alpha + \beta < 1$ and $\omega > 0$. z_t represents independent and identically distributed innovations characterized by the pdf $G(\boldsymbol{\theta})$, which may be either a GCE or STE—or their corresponding nested distributions, normal and Student's t, respectively. Models are estimated by ML in one step, which, for the case of STE, imply maximizing the following log-likelihood function given a sample of size T :

$$\begin{aligned} L_{STE} = T \log & \left[\frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right) \sqrt{\pi(v-2)}} \right] - \frac{1}{2} \sum_{t=1}^T \log(\sigma_t^2) \\ & - \frac{v+1}{2} \sum_{t=1}^T \log \left[1 + \frac{\varepsilon_t^2}{(v-2)\sigma^2} \right] + \sum_{t=1}^T \log \left[1 + \sum_{s=1}^n \delta_s P_s \left(\frac{\varepsilon_t^2}{\sigma_t^2} \right) \right]. \end{aligned} \quad (16)$$

The ML estimation overcomes the problem of negativity values, provided that the algorithms converge to a global optimum. In order to avoid local optima in joint estimation, different procedures can be used to refine the initial values and simplify the convergence. The initial values for the second step can be obtained through the method of moments in Del Brio and Perote [23]. However, to guarantee the convergence to global optima, it is convenient to perform a last step with a joint estimation of the total density parameters. For this purpose, Newton-Raphson and Broyden-Fletcher-Goldfarb-Shanno methods implemented in R packages.

2.4 Risk Measures

The performance of the SNP expansions is assessed in terms of both VaR and ES. The former is the most standard risk measure since it is the maximum expected loss for a given confidence and time horizon and corresponds to a quantile of the distribution. Given a stochastic variable X with cdf F_X , the VaR can be defined as:

$$VaR_\alpha(X) = \inf \{x \in \mathbb{R} : F_X(x) \leq \alpha\}, \quad (17)$$

where α represents the significance level, which conforming to the regulation is set at 1% and 2.5%. According to this notation, VaR is quantified in the left tail of the density and it is denoted VaR at 99% and ES at 97.5%. Therefore, VaR can be computed from the inverse cdf or quantile function, $VaR_\alpha(X) = F_x^{-1}(\alpha)$, displayed in Eqs. (5) and (12).

The ES is the expected loss conditioned on the fact that the VaR has been exceeded. This measure has been recently proposed to replace VaR by international organizations on banking supervision. In the left tail of the density, it can be computed as

$$ES_\alpha(x) = \mathbb{E}[-x | x \leq -VaR_\alpha] = -\frac{1}{\alpha} \int_0^\alpha VaR_\xi d\xi = -\frac{1}{\alpha} \int_0^{-VaR_\alpha} xf(x) dx. \quad (18)$$

Therefore, and according to Del Brio, Mora-Valencia, Perote [24] the ES for the GCE can be computed as

$$\begin{aligned} ES_\alpha(X) &= -\frac{\phi(\Phi^{-1}(\alpha))}{\alpha} \\ &\times \left[1 + \sum_{s=3}^n d_s \left[H_s(\Phi^{-1}(\alpha)) + sH_{s-2}(\Phi^{-1}(\alpha)) \right] \right]. \end{aligned} \quad (19)$$

then, the ES for the STE is

$$\begin{aligned} \text{ES}_\alpha(X) = & -\frac{T_v(t_v^{-1}(\alpha))}{\alpha} \left(\frac{\nu + (t_v^{-1}(\alpha))^2}{\nu - 1} \right) \\ & \times \left[1 + \sum_{s=3}^n \delta_s \left[P_s(t_v^{-1}(\alpha)) + s P_{s-2}(t_v^{-1}(\alpha)) \right] \right]. \end{aligned} \quad (20)$$

2.5 Backtesting

The performance of the GCE and STE is tested through backtesting techniques along with a rolling window size of 500 observations. For the backtesting of VaR, the unconditional coverage test proposed by Kupiec [25] and the independence test of Christoffersen [26]. Kupiec's test identifies the VaR exceptions, which happens when realized losses exceed the estimated VaR with a significance level α during the backtesting period. The "unconditional coverage" examines if the sequence of exceptions follows an iid Bernoulli process with probability $\alpha(1 - \alpha)$, the null hypothesis assuming that the number of exceptions is correct. The Christoffersen test investigates the independence of the VaR exceptions, which holds under the null hypothesis.

For the backtesting of ES, we follow Acerbi and Székely [27], who prove that their Z_{ES} test is not to be sensitive to possible values of VaR and ES is equivalent to

$$ES_{1-\alpha,t} = VaR_{1-\alpha,t} - \frac{1}{\alpha} E[(X_t + VaR_{1-\alpha,t}) I_t], \quad (21)$$

where I_t is an indicator function that takes the value 1 if $X_t + VaR_{1-\alpha,t} < 0$, and 0 otherwise. Then, the Z_{ES} statistic becomes

$$Z_{ES} \left(\bar{X} \right) = \sum_{t=1}^T \frac{\alpha (ES_{1-\alpha} - VaR_{1-\alpha}) + (X_t + VaR_{1-\alpha}) I_t}{T \alpha ES_{1-\alpha,t}}, \quad (22)$$

Under the null, the model represents adequate performance for ES and under the alternative ES is rejected regardless of the VaR. The test's critical values are obtained by Monte Carlo simulation, for details see Velásquez-Gaviria, Mora-Valencia, and Perote [28].

3 Empirical Results

3.1 Data

The database consists of four indices observed at daily frequency: MSCI World Energy Sector Index (World Energy), MSCI World Metals & Mining Index (World Metals & Mining), S&P GSCI Industrial Metals Spot Index (Metals), and the Bloomberg Commodity Spot Index (Commodity). For the first two indices, we have 6300 observations (January 1995/August 2019), for the third index 7400 observations (January 1990/August 2019), and for the fourth index 7600 observations (January 1990/August 2019). Logarithmic returns are calculated as $r_t = \log(V_t/V_{t-1})$, where V_t is the value of each index at time t . Table 1 reports the main descriptive statistics of the series.

3.2 In-Sample Analysis

GCE Innovations The AR(1)-GARCH(1,1) model is firstly fitted with GCE innovations. For each index, five different models are estimated according to the involved parameters in the GCE density to compare the risk measure performance with larger expansions. The first model considers the d_3 and d_4 parameters, as in most financial applications. Then, a new parameter is added in each model until the fifth model, which considers parameters from d_3 to d_8 . Considering the four series, a total of 20 models, denoted by ML(1)–ML(20), were estimated by ML. Results are not provided for the sake of saving space but are available upon request.

STE Innovations Table 2 contains the estimations of the AR(1)-GARCH(1,1) model with STE innovations. The first model includes δ_1 , δ_2 , and δ_3 parameters. Subsequently, a new parameter is added in each model, and the sixth model incorporates from δ_1 parameter to δ_8 parameter. In this case, a total of 24 models—ML(21)–ML(44)—were estimated by ML. The δ_1 parameter, related to the mean, is significant and positive in all cases, except for the Commodity index. The δ_2 parameter is negative and significant in all cases. This may be explained because the degrees of freedom estimation vary between 7 and 10 for the analyzed returns.

Thus, the ratio $v/(v - 2)$ results to be greater than the second moment of the empirical distribution. The δ_3 parameter is negative and significant in all cases. This parameter is related to the asymmetry and represents the leverage effect of financial returns. The δ_4 parameter results to be significant in all the expansions for the Metals and Commodity indices, while for the World Energy index, it is significant only in the fourth and fifth-order expansions, this parameter is related to the shape of the tails in returns. On the other hand, for the Commodity index, this parameter is significant when the expansions of orders six and seven are involved. In summary, the estimation of the δ_4 parameter is positive in expansions of orders four and five

Table 1 Descriptive statistics of the return series

Index	Mean	Standard deviation	Skewness	Kurtosis	Min	Max	Jarque-Bera	Ljung-Box	Augmented Dickey-Fuller
World Energy	-0.0006 (-14.91%)	0.0213 (33.88%)	-0.5858	14.9857	-0.16	0.1600	16,521.3	0.06026	0.01
World Metals & Mining	-0.0003 (-7.04%)	0.0203 (32.25%)	-0.3926	8.8561	-0.145	0.1582	5249.7	0.0021	0.01
Metals	0.0002 (5.12%)	0.0100 (15.84%)	-0.4849	6.2073	-0.059	0.0524	1021.8	0.0000	0.01
Commodity	0.0001 (0.70%)	0.0203 (32.18%)	-0.4490	8.5372	-0.143	0.1508	4097.3	0.1509	0.01

The first column refers to the index name, the second the length of the sample, the third, the mean, and in the parenthesis the annualized mean ($\mu \times 252$). The fourth refers to the standard deviation. In parenthesis, the annualized volatility ($\sigma \times \sqrt{252}$), the following four columns refer to the skewness, the kurtosis, the minimum, and the maximum of the data sample. Column number nine contains the estimated Jarque-Bera statistic. Null hypothesis implies normality. Column number ten contains the p-value of the first lag Ljung-Box test. The null hypothesis implies the absence of autocorrelation of order one. The last column refers to the p-value of the augmented Dickey-Fuller with only constant, selecting with BIC criteria the deterministic part. In this test null hypothesis implies unit root

Table 2 In-sample estimation for AR(1)-GARCH(1,1) with STE innovations

MSCI World Energy sector Index							MSCI World Metals & Mining Index					
	ML(21)	ML(22)	ML(23)	ML(24)	ML(25)	ML(26)	ML(27)	ML(28)	ML(29)	ML(30)	ML(31)	ML(32)
φ	0.0005 (0.0001)	0.0005*** (0.0001)	0.0004*** (0.0001)	0.0006*** (0.0001)	0.0007*** (0.0001)	0.0002 (0.0001)	0.0005 (0.0001)	0.0004 (0.0001)	0.0002 (0.0001)	0.0001 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)
ϕ	0.1048*** (0.0124)	0.0968*** (0.0131)	0.1033*** (0.0133)	0.1113*** (0.0159)	0.1042*** (0.0137)	0.1082*** (0.0133)	0.2234*** (0.0156)	0.2176*** (0.0130)	0.2265*** (0.0137)	0.2243*** (0.0129)	0.2299*** (0.0141)	0.2281*** (0.0132)
ω	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)						
α	0.0987*** (0.0081)	0.0649*** (0.0076)	0.0655*** (0.0075)	0.0739*** (0.0091)	0.0738*** (0.0079)	0.0713*** (0.0077)	0.0636*** (0.0059)	0.0705*** (0.0070)	0.0765*** (0.0074)	0.0554*** (0.0060)	0.0542*** (0.0060)	0.0546*** (0.0055)
β	0.9228*** (0.0074)	0.9223*** (0.0079)	0.9225*** (0.0078)	0.9229*** (0.0081)	0.9228*** (0.0081)	0.9228*** (0.0081)	0.9429*** (0.0071)	0.9222*** (0.0071)	0.9225*** (0.0069)	0.9429*** (0.0069)	0.9429*** (0.0069)	0.9429*** (0.0085)
ν	10.381*** (0.1764)	10.392*** (0.2147)	10.307*** (0.3381)	10.338*** (0.3273)	10.342*** (0.5503)	10.344*** (0.2098)	9.9235*** (0.3125)	10.175*** (0.2088)	10.206*** (0.2528)	9.7118*** (0.1816)	9.6951*** (0.2215)	10.161*** (0.3372)
δ_1	0.0548*** (0.0188)	0.0541*** (0.0177)	0.0703*** (0.0189)	0.0528*** (0.0203)	0.0648*** (0.0213)	0.0541*** (0.0218)	0.0418*** (0.0190)	0.0409*** (0.0186)	0.0115** (0.0201)	0.0354* (0.0209)	0.0410* (0.0219)	0.0371* (0.0217)
δ_2	0.1645*** (0.0097)	0.0686*** (0.0100)	0.0530*** (0.0101)	0.0272*** (0.0098)	0.0282*** (0.0121)	0.0126*** (0.0101)	0.0898*** (0.0104)	0.0255*** (0.0105)	-0.0317*** (0.0105)	-0.0228*** (0.0106)	-0.0155*** (0.0094)	-0.0105*** (0.0112)

(continued)

Table 2 (continued)

	MSCI World Energy sector Index								MSCI World Metals & Mining Index							
	ML(21)	ML(22)	ML(23)	ML(24)	ML(25)	ML(26)	ML(27)	ML(28)	ML(29)	ML(30)	ML(31)	ML(32)				
δ_3	0.0456*** (0.0085)	0.0425*** (0.0076)	0.0508*** (0.0094)	0.0564*** (0.0102)	0.0539*** (0.0109)	-0.0536*** (0.0116)	-0.0322*** (0.0089)	0.0312*** (0.0087)	-0.0288*** (0.0102)	-0.0296*** (0.0107)	-0.0294*** (0.0114)	-0.0290*** (0.0112)				
δ_4	0.0155*** (0.0035)	0.0124*** (0.0035)	0.0043 (0.0044)	0.0050 (0.0051)	0.0108* (0.0064)	0.0065 (0.0036)	0.0047 (0.0037)	0.0047 (0.0037)	-0.0101** (0.0047)	-0.0101** (0.0044)	-0.0101** (0.0044)	-0.0062 (0.0057)				
δ_5		0.0030** (0.0016)	0.0034*** (0.0018)	0.0022 (0.0018)	0.0028 (0.0027)	0.0022 (0.0031)	0.0028 (0.0018)	0.0009 (0.0018)	0.0009 (0.0017)	0.0010 (0.0027)	0.0009 (0.0017)	0.0009 (0.0027)				
δ_6			-0.0015*** (0.0007)	-0.0018*** (0.0007)	-0.0018 (0.0015)	-0.0018 (0.0015)	-0.0018 (0.0015)		-0.0018*** (0.0007)	-0.0018*** (0.0007)	-0.0018*** (0.0007)	-0.0018*** (0.0007)	-0.0012 (0.0013)			
δ_7				0.0002 (0.0002)	0.0003 (0.0003)					0.0000 (0.0002)	0.0000 (0.0002)	0.0000 (0.0002)	0.0001 (0.0001)			
δ_8					0.0003*** (0.0001)								0.0000 (0.0001)			
LL	19,803.3	19,799.1	19,803.3	19,809.1	19,808.7	19,811.6	19,102.7	19,096.8	19,101.0	19,103.5	19,103.0	19,102.6				
AIC	39,622.6	39,614.1	39,622.7	39,634.1	39,639.1	39,633.4	38,221.4	38,209.6	38,218.1	38,223.0	38,222.0	38,221.3				

(continued)

Table 2 (continued)

S&P GSCI Industrial Metals Spot Index							Bloomberg Commodity Spot Index						
	ML(33)	ML(34)	ML(35)	ML(36)	ML(37)	ML(38)		ML(39)	ML(40)	ML(41)	ML(42)	ML(43)	ML(44)
φ	0.0002 (0.0001)	0.0004 (0.0001)	0.0004 (0.0001)	-0.0002 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)	0.0000 (0.0001)	0.0004 (0.0001)	0.0004 (0.0001)	0.0000 (0.0001)	0.0002 (0.0001)	0.0003 (0.0001)	
ϕ	-0.0153 (0.0104)	-0.0131 (0.0121)	-0.0145 (0.0122)	-0.0089 (0.0095)	-0.0092 (0.0103)	-0.0125 (0.0124)	-0.0202*** (0.0120)	-0.0202*** (0.0120)	-0.0202*** (0.0121)	-0.0163*** (0.0137)	-0.0203*** (0.0105)	-0.0151*** (0.0120)	-0.0201*** (0.0120)
ω	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)
α	0.0597*** (0.0046)	0.0720*** (0.0079)	0.0740*** (0.0081)	0.0691*** (0.0069)	0.0612*** (0.0097)	0.0562*** (0.0063)	0.0454 (0.0047)	0.0650*** (0.0070)	0.0664*** (0.0068)	0.0423*** (0.0052)	0.0439 (0.0056)	0.0552*** (0.0068)	
β	0.9434*** (0.0088)	0.9225*** (0.0074)	0.9225*** (0.0073)	0.9434*** (0.0087)	0.9434*** (0.0082)	0.9434*** (0.0086)	0.9555*** (0.0067)	0.9227*** (0.0071)	0.9225*** (0.0071)	0.9225*** (0.0081)	0.9537*** (0.0090)	0.9538 (0.0094)	0.9351*** (0.0073)
v	7.9985*** (0.1957)	8.2958*** (0.2762)	8.3388*** (0.4206)	8.3559*** (0.2810)	8.1449*** (0.2393)	8.3209*** (0.3115)	7.5177 (0.3643)	7.6361*** (0.4235)	7.8041*** (0.4220)	7.4964 (0.411)	7.5017*** (0.3634)	7.8079*** (0.3294)	
δ_1	0.0508*** (0.0180)	0.0503*** (0.0176)	0.0285*** (0.0194)	0.0880*** (0.0201)	0.0758*** (0.0219)	0.0639*** (0.0217)	-0.0019 (0.0179)	-0.0008 (0.0174)	-0.0269 (0.0193)	0.0154 (0.0202)	0.0082 (0.0222)	0.0084 (0.0221)	
δ_2	-0.0483*** (0.0093)	0.0133*** (0.0102)	-0.0069*** (0.0109)	-0.1405*** (0.0111)	-0.0793*** (0.0109)	-0.0216*** (0.0081)	0.0348*** (0.0100)	-0.0258*** (0.0110)	-0.0182*** (0.0110)	-0.0072*** (0.0533)	-0.0328*** (0.0127)	-0.1582*** (0.0122)	

(continued)

Table 2 (continued)

	S&P GSCI Industrial Metals Spot Index								Bloomberg Commodity Spot Index							
	ML(33)	ML(34)	ML(35)	ML(36)	ML(37)	ML(38)	ML(39)	ML(40)	ML(41)	ML(42)	ML(43)	ML(44)				
δ_3	-0.0296*** (0.0087)	-0.0291*** (0.0086)	-0.0356*** (0.0105)	-0.0343*** (0.0107)	-0.0399*** (0.0121)	-0.0395*** (0.0121)	0.0038 (0.0089)	-0.0038 (0.0086)	-0.0088 (0.0107)	-0.0081 (0.0112)	-0.0034 (0.0126)	-0.0033 (0.0126)				
δ_4	0.0054*** (0.0035)	0.0052*** (0.0036)	0.0163*** (0.0049)	0.0170*** (0.0048)	-0.0135*** (0.0049)	-0.0170*** (0.0049)	0.0081** (0.0036)	0.0082*** (0.0036)	-0.0183*** (0.0196)	-0.0183*** (0.0054)	-0.0183*** (0.0054)	-0.0094*** (0.0063)				
δ_5		0.0018	0.0015	0.0037	0.0036			0.0032** (0.0017)	0.0031* (0.0017)	0.0031* (0.0016)	0.0013 (0.0026)	0.0013 (0.0027)				
δ_6				-0.0026*** (0.0006)	-0.0029*** (0.0006)	-0.0022* (0.0012)			-0.0033*** (0.0011)	-0.0034*** (0.0006)	-0.0011*** (0.0012)					
δ_7					-0.0002 (0.0002)	-0.0002 (0.0002)				0.0002 (0.0002)	0.0002 (0.0002)	0.0002 (0.0002)				
δ_8						0.0000 (0.0001)							0.0001* (0.0000)			
LL	-22,759.8	-22,752.4	-22,756.8	-22,768.9	-22,769.4	-22,766.6	-23,856.8	-23,848.6	-23,856.6	-23,864.9	-23,868.6	-23,873.4				
AIC	45,535.6	45,520.9	45,529.6	45,553.9	45,554.8	45,549.3	47,729.7	47,713.3	47,729.2	47,745.8	47,753.2	47,756.8				

φ and ϕ are the parameters of the AR(1); ω , α , and β are the parameters of the GARCH(1,1) and δ_i $\forall i = 1, 2, \dots, 8$ the parameters of the STE. LL and AIC are the log-likelihood and Akaike's information values. The asterisks signal the significance of the parameter whose p-value is in parenthesis

for all cases. For higher orders, the estimation turns out to be negative. This implies that the pointing and the shape of the tail shifts as higher order polynomials are involved in the estimation. Regarding the other parameters, δ_5 estimation presents a positive sign in every case, whereas the δ_7 parameter estimation is not significant but remains positive. The δ_6 parameter is significant in most cases, and it is negative. The δ_8 parameter is significant for the World Energy and Commodity indices, with a positive sign when expansions of orders two, four, and six are jointly estimated.

According to the log-likelihood and AIC criteria, the STE distribution with the first four-order expansion is the model that best fits the whole sample for all the analyzed indices. This is in line with results obtained for GCE density. The significance of the estimated parameters for STE innovations highlights that the SNP distributions better capture asymmetry, heavy-tails, and multimodality in the tails. This is confirmed in Fig. 1 that depicts the fit for the left tails of the standardized

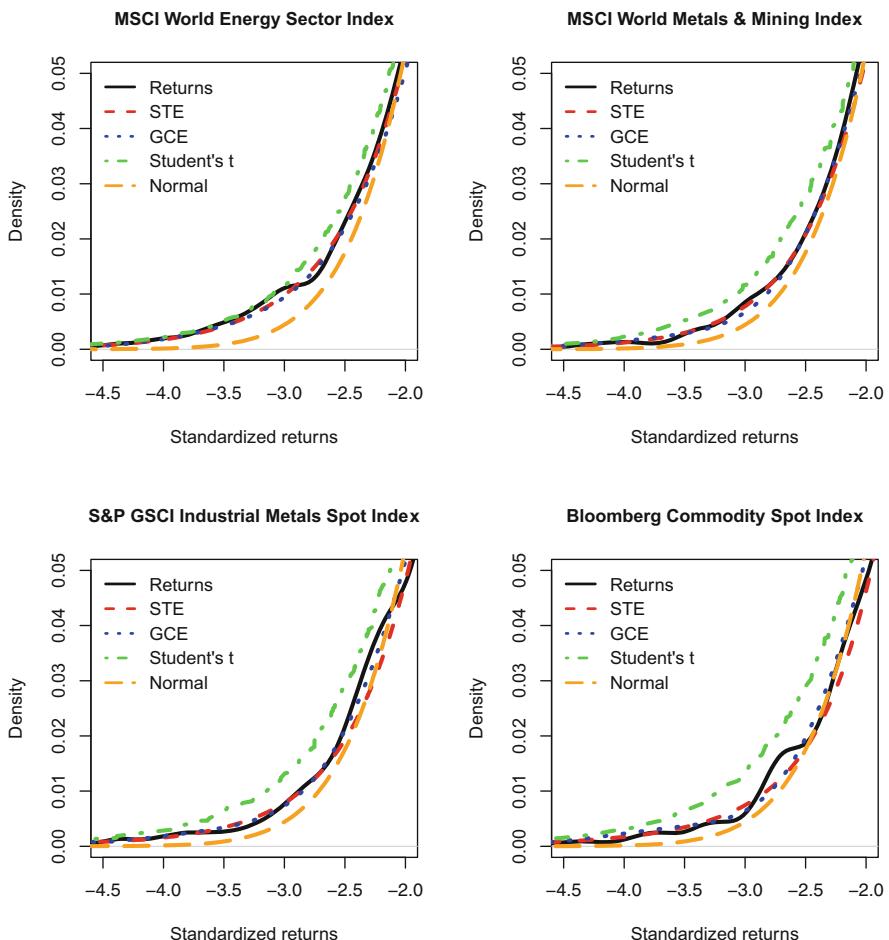


Fig. 1 Left tail of empirical and fitted distributions for standardized residuals

returns. The plots provided the comparison of the larger GCE and STE with their respective basis distributions, the normal and standard Student's t.

The plot highlights the outstanding performance of the expansions, especially the STE, for fitting both the peak at the mean and the behavior of the tails. This evidence has a remarkable implication for risk management purposes since underestimating the impact of the extreme values will not enable the model to anticipate extreme losses and could lead to bankruptcy.

3.3 Backtesting

This section presents the backtesting validation techniques applied through a rolling window of 500 days on a backtesting period of 5800 observations for World Energy and World Metals & Mining indices, 6800 observations for Metals index, and 7100 observations for Commodity index. The analysis covers the comparison between different versions of the GCE, the STE, and their respective basis distributions as the benchmark. Notably, we consider the most widely used GCE with two versions with 3 and 4 parameters. For the case of GCE we choose parameters d_3 , d_4 , d_6 , and d_8 (since a single parameter seems to be enough to capture skewness) and for the case of STE we consider δ_1 , δ_2 , δ_3 , and δ_4 (since the properties for higher moments should be studied, emphasizing on the importance of the correlations of all parameters with v). We employ the traditional Kupiec and Independence tests for testing purposes, which have been consistent with the results of new and more sophisticated techniques for validating VaR amounts. Moreover, we apply the recent methodology proposed by Acerbi and Székely [27] to test ES, and it seems to be an adequate technique considering the lack of elicitability property for ES itself.

VaR Backtesting Results Following the recommendation of the Basel Committee, the backtesting displayed in Table 3 is validating at VaR at 99% (backtesting at 97.5% provides similar, and these results are available upon request). For the models in red (green) the null is rejected (not rejected). As expected, the normal distribution performs the worst. For Student's t, the results are better. There is no statistical evidence to reject the null hypothesis of the Kupiec and the independence tests for Metals and Commodity indices. For the case of GCE distribution with three- and four-order expansion, the null hypothesis for the Kupiec and independence test cannot be rejected for the Metals and World Metals & Mining indices. For larger expansions, GCE truncated up to the sixth order does not perform well for all cases. In contrast, the largest expansion considered in our applications only works well for the World Metals & Mining index. This is consistent with the in-sample analysis, where the larger expansions did not provide significant improvements. The STE density that considers the first four parameters performs the best, while STE with the first three parameters works well for the World Energy and World Metals & Mining indices.

Table 3 99%-VaR backtesting. Kupiec and independence tests

Distribution	MSCI World energy sector Index ev=58				MSCI World Metals & Mining Index ev=58				S&P GSCI Industrial Metals Spot Index ev=68				Bloomberg Commodity Spot Index ev=71			
	V	VR	KT	IT	V	VR	KT	IT	V	VR	KT	IT	V	VR	KT	IT
Normal	124	2.13	29.35	35.26	118	2.03	27.12	31.05	115	1.69	24.23	28.62	149	2.09	18.65	21.46
Student t	93	1.60	18.03	26.18	87	1.50	12.69	13.02	80	1.16	1.76	1.77	70	1.05	0.27	3.95
GCE 3,4	79	1.36	6.89	14.81	65	1.12	1.85	4.63	82	1.20	2.45	3.92	84	1.18	3.65	5.87
GCE 3,4,6	87	1.50	12.69	16.44	72	1.24	6.45	7.03	87	1.27	4.24	7.54	89	1.25	4.65	6.83
GCE 3,4,6,8	85	1.46	11.10	17.83	70	1.22	3.69	5.63	90	1.32	4.97	8.12	88	1.19	4.04	6.37
STE 1,2,3	66	1.14	1.06	5.04	45	0.77	3.18	3.89	53	0.78	7.07	7.97	49	0.69	12.64	13.17
STE 1,2,3,4	70	1.20	2.35	5.78	54	0.93	0.28	1.30	61	0.89	1.34	3.76	58	0.82	3.62	5.42

ev is the expected number of exceptions. V and VR are the number and ratio of violations, respectively. The critical value of the Kupiec test (KT) is 3.84. The critical value of the Independence test (IT) is 5.99. For Kupiec's test, the null hypothesis implies correct VaR exceedances. In the independence test, the null hypothesis implies correct and independent exceedances

Table 4 Backtesting of ES at 97.5%. Test Z_{es}

Distribution	Test Z_{es}			
	MSCI World energy sector Index	MSCI World Metals & Mining Index	S&P GSCI Industrial Metals Spot Index	Bloomberg Commodity Spot Index
	T=5800 cv=-0.2455	T=5800 cv=-0.2455	T=6800 cv=-0.2434	T=7100 cv=-0.2432
Normal	-0.4003	-0.4025	-0.4267	-0.4195
Student t	-0.3378	-0.3402	-0.3318	-0.3167
GCE 3,4	-0.2611	-0.2209	-0.2417	-0.2410
GCE 3,4,6	-0.2668	-0.2426	-0.2465	-0.2502
GCE 3,4,6,8	-0.2653	-0.2376	-0.2426	-0.2516
STE 1,2,3	-0.2179	-0.1904	-0.2215	-0.2101
STE 1,2,3,4	-0.2035	-0.1716	-0.2173	-0.2067

The backtesting period (T) and the critical value (cv) for the test are displayed in the third row of the table. The Null hypothesis is rejected when the estimated test is greater than the cv

ES Backtesting Results For ES testing, we employ Acerbi and Székely [27] proposal, where the model performs well if the Z_{ES} statistic—Eq. (22)—is close to zero, otherwise, the model underestimates risk if it is significantly negative. Therefore, the critical value is approximately -0.24 calculated by employing Monte Carlo simulation. For more details about the calculation of the critical value, see the algorithm proposed in Acerbi and Székely [27] and the references therein. The results show that normal and Student's t underpredict risk in all cases. Again, the GCE with the shortest expansion performs better than other expansions, while all the analyzed STE expansions present adequate performance for all the indices. These results are summarized in Table 4. For the models in red (green) the null is rejected (not rejected).

4 Conclusions

In general, commodities as oil and related assets present higher volatility than equity assets. Among the reasons is greater exposure to geopolitical, climatical, and the speculation of investors, highlighting the need to resort to models capable of adapting to unexpected extreme losses. This study examines two important risk measures, namely VaR and ES applied to four indices: MSCI World Energy Sector Index (World Energy), MSCI World Metals & Mining Index (World Metals & Mining), S&P GSCI Industrial Metals Spot Index (Metals), and the Bloomberg Commodity Spot Index (Commodity). To this end, we filter the returns of the indices through the AR-GARCH process with four different distributions for the innovations: Normal and Student's t and their respective semi-nonparametric versions, Gram-Charlier expansion (GCE), and the Student's t expansion (STE) densities.

We introduce the STE in terms of the derivatives of the Student's t and provide some of its main properties that result in valid asymptotic approximations to the true density and its risk measures obtained as a by-product. Mainly, we provide its pdf and log-likelihood function for future applications and replication purposes. Furthermore, we derive a closed-form expression to calculate ES for STE density. Interestingly, GCE and STE distributions share multiple similarities, among them that they are Edgeworth expansions of a weight function, the resulting polynomials of the expansion are orthogonal to each other, the even (odd) moment of order s depend linearly on the first s even (odd) parameters. The even polynomials are related to kurtosis and the odd ones to asymmetry. In fact, the STE converges to the GCE as the degrees of freedom parameter tends to infinity. This parameter is notably higher than that of the basis Student's t distribution and its co-movements with the rest of the density parameters seem to be the cornerstone for the accuracy of this expansion.

Model performance is tested through backtesting techniques. The VaR backtesting is performed through the well-known Kupiec and Independence tests, while we apply the proposal of Acerbi and Székely [27] for ES testing. The results show that GCE and STE densities with the shortest expansions fit adequately to the empirical distributions of the indices. In addition, STE density is the model that performs the best in our VaR and ES backtesting results.

Consequently, we recommend the use of the STE distribution for risk measure purposes since its capacity to anticipate extreme values originated in the commodities, energy, and metals & mining markets. Future research will be focused on applying our proposed distribution to different assets and different ES backtesting techniques. Moreover, the performance of STE density may be compared with other parametric and flexible distributions.

References

1. Hansen, B.E.: Autoregressive conditional density estimation. *Int. Econ. Rev.* **35**, 705–730 (1994)
2. Fernández, C., Steel, M.F.: On Bayesian modeling of fat tails and skewness. *J. Am. Stat. Assoc.* **93**(441), 359–371 (1998)
3. Theodossiou, P.: Financial data and the skewed generalized t distribution. *Manag. Sci.* **44**(12-part-1), 1650–1661 (1998)
4. Jones, M.C., Faddy, M.J.: A skew extension of the t-distribution, with applications. *J. R. Stat. Soc., B: Stat. Methodol.* **65**(1), 159–174 (2003)
5. Bauwens, L., Laurent, S.: A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *J. Bus. Econ. Stat.* **23**(3), 346–354 (2005)
6. Cardona, E., Mora-Valencia, A., Velásquez-Gaviria, D.: Testing expected shortfall: an application to emerging market stock indices. *Risk Manag.* **21**(3), 153–182 (2019)
7. Jørgensen, B.: Statistical Properties of the Generalized Inverse Gaussian Distribution, vol. 9. Springer (1982)
8. Barndorff-Nielsen, O.E.: Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Stat.* **24**(1), 1–13 (1997)
9. Mencía, J., Sentana, E.: Estimation and testing of dynamic models with generalized hyperbolic innovations. Available at SSRN 790704 (2005)
10. Gupta, R.D., Kundu, D.: Generalized exponential distribution: different method of estimations. *J. Stat. Comput. Simul.* **69**(4), 315–337 (2001)
11. Carrasco, J.M., Ortega, E.M., Cordeiro, G.M.: A generalized modified Weibull distribution for lifetime modeling. *Comput. Stat. Data Anal.* **53**(2), 450–462 (2008)
12. Níguez, T.-M., Paya, I., Peel, D., Perote, J.: Flexible distribution functions, higher-order preferences and optimal portfolio allocation. *Quant. Finance* **19**, 669–703 (2019)
13. Edgeworth, F.: The asymmetrical probability-curve. *The London, Edinburgh, and Dublin Philosophical Magazine and J. Sci.* **41**(249), 90–99 (1896)
14. Barton, D.E., Dennis, K.E.: The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. *Biometrika* **39**(3/4), 425–427 (1952)
15. Jondeau, E., Rockinger, M.: Gram-Charlier densities. *J. Econ. Dyn. Control.* **25**(10), 1457–1483 (2001)
16. León, Á., Mencía, J., Sentana, E.: Parametric properties of semi-nonparametric distributions, with applications to option valuation. *J. Bus. Econ. Stat.* **27**(2), 176–192 (2009)
17. Abramowitz, M., Stegun, I. A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Series 55, Tenth Printing. National Bureau of Standards Appl. Math. (1972)
18. Kendall, M., Stuart, A.: Distribution Theory The Advanced Theory of Statistics, vol. 1. Griffin, London (1977)
19. Trespalacios, A., Cortés, L.M., Perote, J.: Uncertainty in electricity markets from a semi-nonparametric approach. *Energy Policy* **137**, 111091 (2020)
20. Níguez, T.-M., Perote, J.: Moments expansion densities for quantifying financial risk. *N. Am. J. Econ. Finance* **42**, 56–69 (2017)
21. Cortés, L.M., Mora-Valencia, A., Perote, J.: The productivity of top researchers: a semi-nonparametric approach. *Scientometrics* **109**(2), 891–915 (2016)
22. Mauleon, I., Perote, J.: Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to the Students t. *Eur. J. Financ.* **6**, 225–239 (2000)
23. Del Brio, E.B., Perote, J.: Gram-Charlier densities: maximum likelihood versus the method of moments. *Insur.: Math. Econ.* **51**(3), 531–537 (2012)
24. Del Brio, E.B., Mora-Valencia, A., Perote, J.: Risk quantification for commodity ETFs: backtesting value-at-risk and expected shortfall. *Int. Rev. Financ. Anal.* **70**, 101163 (2020)

25. Kupiec, P.: Techniques for verifying the accuracy of risk measurement models. *J. Deriv.* **3**(2) (1995)
26. Christoffersen, P.F.: Evaluating interval forecasts. *Int. Econ. Rev.* **39**, 841–862 (1998)
27. Acerbi, C., Szekely, B.: General properties of backtestable statistics. Available at SSRN 2905109 (2017)
28. Velásquez-Gaviria, D., Mora-Valencia, A., Perote, J.: A comparison of the risk quantification in traditional and renewable markets. *Energies.* **13**(11), 2805 (2020)

Predicting Housing Prices for Spanish Regions



Paloma Taltavull de La Paz 

Abstract This paper aims to forecast the long-term trend of housing prices in the Spanish cities with more than 25,000 inhabitants, a total of 275 individual municipalities. Based on a causal model explaining housing prices based on six fundamental variables (changes in population, income, number of mortgages, interest rates, vacant and housing prices), a pool VECM technique is used to estimate a housing price model and calculate the ‘stable long-term price’, a central concept defined in the formal valuation process. The model is estimated for the period 1995–2020, and the long term is approached from 2000 to 2026, so the prediction exercise includes backcast and forecast period allowing to extract the long-term cycle housing price have followed during last 20 years and project it further 6 years. The analytical process follows three steps. Firstly, it identifies the cities following a common pattern in their housing market by clustering twice the cities: (1) using house price time series and (2) using a machine learning approach with the six fundamental variables. Results give a comprehensible evolution of the long-term component of housing prices, and the model also permits the understanding of the main drivers of housing prices in each Spanish region. Clustering cities with two statistical tools gives pretty similar results in some cities but is different in others. The challenge of finding the correct grouping is critical to understanding the housing market and forecasting their prices.

Keywords Housing prices · Panel VECM forecast · Time series · Housing valuation

P. Taltavull de La Paz 
University of Alicante, Alicante, Spain
e-mail: paloma@ua.es

1 Introduction and Motivation

The interest in knowing the future evolution of residential prices has different motivations. On the one hand, anticipating future prices allows assessing the residential household wealth in an economy and its potential as a generator of consumption growth through the wealth effect [1]. On the other hand, predicting future prices reduces uncertainty in investment markets and facilitates the movement of capital and the decision to build.

Other relevant reasons support the interest in advancing housing prices, but one stands out for its great relevance to the economic system. Residential prices and their evolution are part of macroprudential policy. To the extent that housing (and real estate in general) serves as collateral for the financing granted for its provision or purchase, a stable property value is part of the risk assumed by the financial system. In an economy with a developed mortgage market, correctly pricing real estate can be vital to keeping risk levels under control and avoiding situations that can lead to a loss of confidence with negative results for the institution or the financial system as a whole. The reason lies in the fact that the credits backing real estate are long term, and even if the financial institution acts correctly in granting them taking care of their risk levels, the economic situation can change completely during their lifetime so that operations that are robust at one point in time would fail when the cycle changes. A massive fall in residential prices resulting from an economic or another shock would dramatically increase the risk level of the loans granted (and the financial assets issued on them). In contrast, a generalized increase in prices would generate the opposite effect, encouraging the financial system to grant more credit with very low-risk levels (at the moment), leading to increased exposure to real estate risk. Therefore, it is understandable that there is an interest in detecting real estate price bubbles or price corrections.

Predicting residential prices is not a simple matter. On the one hand, housing is a highly heterogeneous good, and its value depends on different groups of factors. The literature identifies the most relevant as location (AMM model, [48]); housing characteristics [49]; neighbourhood and demander characteristics [2]; but also the evolution of a set of the so-called fundamental factors ([1, 2, 3]) that determine the existence of demand pressure (generally due to migratory movements, [50]), and the payment capacities of potential demanders or their financial activity. It is considered the fundamentals that delimit the evolution of prices in the long run, although their local particularities determine the specific value levels of residential goods. Property heterogeneity and the bundle of variables affecting housing prices convert price prediction into a complex and challenging task.

Housing prices are critical for the financial system due to property acts as collateral of the loan. It is why appraisal techniques have been developed and embraced in most of the developed countries. In the Spanish system, financial institutions take, as a reference value when granting mortgages, the value resulting from the appraisal of the property (the so-called mortgage price, [42]). Such calculation is under complex rules that include the precise measurement of the

property, the location and certain adjustments that attempt to identify its market price independently of the price that has been offered in the transaction or declared. Valuations in Spain are calculated in real-time by institutions endorsed by the Bank of Spain and specialized training.

The financial crisis has globally evidenced how property values can be influenced by market shocks from stable levels (with precise valuations), towards scenarios of extreme risk, so that interest in anticipating these shocks is growing among institutions with macroprudential responsibility (see the alert mechanism of the EU's Macroeconomic Imbalance Procedure). Institutions responsible for property valuation methodology (IVS—International Valuation Standards, white book; TEGoVA, European Valuations Standards, blue book; RICS, red book) agree on the relevance of determining a stable value, although only some regulations in Europe, such as the Spanish ECO/805, include the obligation to estimate a long-term value to serve as a guide for the granular valuations carried out. This price concept is known as the 'Equilibrium Value' and would be the long-term value around which the observed price of the property evolves.

This paper aims to estimate long-term value for housing prices in Spanish municipalities and predict the following 5 years.

2 Previous Experiences and Evidence: Literature Review of Housing Price Prediction

There is ample evidence of price predictability in the literature. There is also a debate between those who advocate whether residential prices are predictable and those who do not [4], which stems from the difficulty of approximating complex goods' prices and the multiple influences they receive from the environment and their dynamics.

The literature is extensive, covering the issue of predicting prices, most of them paying attention to price formation from different perspectives as a means of evaluating its model and prediction. The main contributions follow the theory of residential price behaviour and assess the effects of aggregate economic variables (the fundamentals) on their dynamics.

Existing works explain that housing prices evolved based on their demand fundamentals as drivers. Economic fundamentals are macroeconomic and general variables affecting the housing market, such as demographic [51], income levels and ability to pay [2], investment flows, inflation [52]; or market expectations [5, 6]; with previous studies demonstrating that housing prices receive differential effects according to territorial behaviour [2]. The indirect effect of housing prices is spread out to the economy and society. Evidence shows that residential prices play a relevant role in driving economic growth by developing income and wealth effects [1, 7], and their effects are maximized at the urban level [8]. Wealth effects have been estimated in numerous countries, and the debate remains on the table,

with different studies finding strong wealth effects on those economies with housing markets led by homeownership while others reject the hypothesis of wealth effect influence [53]. Housing prices also receive influences from and would affect the Monetary Policy. Several studies have found causality between housing prices and lending activity [9]; and a rule relating monetary policy and property prices [10] in the way that changes in housing prices would affect significantly macroeconomic and financial variables through the housing price channel ([11, 12, 47]).

The impact of residential prices on the economy is amplified through the wealth effect and liquidity channels in periods of significant overgrowth. These phases have been referred to as bubbles ([3, 44]), and there is also a rich literature that attempts to estimate the evolution and reasons for them ([3, 13] would be two initial papers) and to establish mechanisms for their detection [14]. Analyses that delve deeper into the bubble identify price reactions in the short run.

Demand fundamentals are not sufficient to justify the different dynamics by geographical areas. The literature support that the structure and composition of the development sector, as well as its reaction to the market price signal (the supply response to changes in prices), are relevant factors to explain the actual evolution of housing construction ([15, 43]). The determinants of the supply function are construction costs and technology (coming from the production function theorem), but also geographical aspects [16] and regulation [17], which are decisive in identifying in which markets developers react to market signal. Regulatory structure or/and geographical components would have more significant weight driving the development reactions in particular markets [18, 19]. This reaction is relevant for its subsequent effects on prices [20].

Price estimation correcting for existing spatial effects is also a very fruitful line of work with Basu and Thibodeau [21], Anselin [22], Montero et al. [23], among others. See [46] for a summary of this literature.

Techniques used for estimating housing prices predictions fall under an extensive range of methods, from pure econometrics ([24–27], among others) to those statistical ([28] using Kriging tool; Pagourtzi et al. [29] using the PYTHIA model,; using network lasso to cluster variables). An increasing number of papers use Machine Learning algorithms, like in Gu et al. [30] or Rico-Juan and Taltavull [31], among others. Kauko and D'Amato [32] summarize the methods used for Mass Appraisal.

3 Theoretical Basis

Housing markets experience a well-known pseudo-equilibrium situation derived from a mismatch between prices in the short run and an adjustment over time. The long-run price behaviour adopted in this paper is considered to depend on their fundamental factors, such as changes in population, income, financing flows, and interest rates (Mayo, 1981, [3, 4, 33], as relevant references), and adjusted for the supply response in each municipality. Supply responses to prices are captured here

through the vacancies, according to the housing supply literature. The model could be represented as in Eq. (1).

$$ph_{it} | Hs_{it} = \Psi (\Delta pop_{it}, Inc_{it}, h_n_{it}, ir_t, \mu_t) \quad (1)$$

where Δpop is the change in the resident population in each municipality, Inc is the average income of the city, h_n captures capital flows for house purchases, Hs is an indicator of housing supply and corrects for the price reaction in each market adjusted for its idiosyncratic particularities, and ir represents the interest rate in nominal terms. The subscript ‘i’ refers to the municipality and t to the observed period. The function is a dynamic operator in a panel framework, which allows the prediction of the dependent variable. This approach defines a dynamic model representing causal relationships, used to forecast the housing price trend-cycle to approach the ‘*long-run price*’. It would approximate the stable long-run value associated with the residential market determinants in each market location.

The prediction minimizes the difference between the observed and estimated price value with the long-term components at each point in time and location (Eq. 2).

$$ph_i^{obs} - \widehat{ph}_i = \mu_i \quad (2)$$

where the first term, ph_i^{obs} refers to the observed price, $(ph)^{\wedge}_i$ hat is the estimated price, and μ is the error component.

4 Data

The data used in this article comes from secondary sources of Spanish statistics. The six variables included in the long-term housing market model are: population changes (which are the proxy for potential new demand), municipality income (which is a proxy for the population’s ability to pay and level of income in the city), investment flows for housing purchases (which are proxied by financial flows or mortgages reflecting new funds coming to the market to facilitate purchase), interest rates (which proxy for the user cost of capital), and housing supply which captures the idiosyncratic supply elasticity response (constraining the price reaction) in each market. House prices are measured using MITMA statistics on appraised prices [45]. The average price per square metre is used. This information has been obtained for Spanish cities with more than 25,000 inhabitants, 275 cities. The cities included in the analysis are listed in the appendix. The model uses annual data, and the time series covers 25 years, 1995–2020.

The whole period for some variables is not available, and the missing observations are extrapolated. The explanation of the data reconstruction process can be found in Table 1 and their sources and basic statistics for each variable.

The variables allow constructing a panel with 275 cross-sections and 25 years (1995–2020) and six variables to analyse the causal relationships between house

Table 1 Variables and statistical description

Variable	Population	Income	Mortgages	Interest rates	Stock	Housing prices
Mean	100,369.5	23,971.37	30,355.89	4.083669	48,498.89	1409.39
Median	47,416.25	23,807.92	15,808	3.427167	24,175.27	1271.98
Maximum	3,273,049	90,334.42	164,464	9.4535	1,575.484	4078.15
Minimum	13,605	7164.415	321	1.910417	1,592.975	283.552
Std. dev.	226,478.6	8134.153	35,497.3	1.805062	107,517.8	650,411
Skewness	9.86069	1.302487	1.980493	1.039491	9.793845	0.91549
Kurtosis	121.7349	8.791829	6.688305	4.048857	119,6178	3,5187
# Obs	6600	6600	6600	6600	6600	6600
Cross-sections	275	275	275	275	275	275
Sources	INE	City Audit, Eurostat and INE ^a	INE	Bank of Spain	INE ^b	MTMA

^aData from 1995 to 2000 has been estimated based on the GDP at province level

^bData has been calculated to obtain the yearly series by adding lagged starts to the housing stock

prices, and four demand-oriented variables considered capturing the long-term fundamentals and controlling for housing supply.

5 Methodology

5.1 Empirical Strategy to Estimate the Model

The econometric strategy is developed through the following steps. In the first step, cities are clustered to find those followed the closer housing market reaction pattern according to the literature. Three cities grouping are calculated: firstly, a clustering based on proximity (arbitrary decision, obtaining 17 clusters); secondly, clustering time series of prices following [34–36, 41] (obtaining eight clusters). Thirdly, using Machine Learning range of methods; the best one based on Catboost (16 clusters).¹

The second step analyses the stationarity properties of the variables through the panel unit root tests giving, as a result, that all of them show a unit root. The third step investigates the presence of a cointegration relationship between the variables. Pool unit root tests developed by Levin, Lin and Chu [37] and Im, Pesaran and Shin [38] are calculated to check the stationarity properties of the variables. LLC tests the existence of a common unit root process across the cross-sections against the null of no unit root, while IPS tests whether individual unit root processes exist across the cross-sections. The first test suggests a homogeneous autoregressive root, while IPS tests for the existence of heterogeneous autoregressive coefficient, both under the alternative hypothesis.

After cointegration relationships are identified, Pedroni (Engle-Granger Based) Cointegration Tests allowing for individual-specific fixed effects are applied to test for a long-run relationship among the variables contained in the model. The cointegration tests imply long-run relationships among variables, identifying the causal patterns of residential market price reaction. Thus, model (3) is estimated through a panel VECM approach to examine the causal relationship between the variables in which the error correction term (ECT) is included in the VAR system as an additional variable. In this step, long- and short-run causality are investigated and serve to define the final model.

The best model is chosen in the next step regarding its predictive capacity among those potential models, which are demonstrated to be stable (all roots inside the circle) and LM test rejecting the null of residual serial correlation.

Model (1) is fitted through the Vector Autoregressive framework, adjusting an Error Correction Model. Technically, the whole model captures the effects of

¹ Details of the 275 cities clustered are available upon request.

residuals in the long-run relationship (ECT) and changes in its components in the short run. Formally, the models could be represented by (3).

$$\Delta X_t = A + X_{t-1}\Gamma_m + \Delta X_{t-j}B_j + N_t \quad (3)$$

or in a matrix formalization:

$$\begin{bmatrix} \Delta Ph_{i,t} \\ \Delta(\Delta Pop_{i,t}) \\ \Delta In_{i,t} \\ \Delta hn_{i,t} \\ \Delta ir_{i,t} \\ \Delta(\Delta Stock_{i,t}) \end{bmatrix} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \end{bmatrix} ECT_{i,t-1} \\ + \sum_{j=1}^T \begin{bmatrix} \beta_{11,j} & \dots & \beta_{16,j} \\ \vdots & \ddots & \vdots \\ \beta_{61,j} & \dots & \beta_{66,j} \end{bmatrix} \begin{bmatrix} \Delta Ph_{i,t-j} \\ \Delta(\Delta Pop_{i,t-j}) \\ \Delta In_{i,t-j} \\ \Delta hn_{i,t-j} \\ \Delta ir_{i,t-j} \\ \Delta(\Delta Stock_{i,t-j}) \end{bmatrix} + \begin{bmatrix} \mu_{1,t} \\ \mu_{2,t} \\ \mu_{3,t} \\ \mu_{4,t} \\ \mu_{5,t} \\ \mu_{6,t} \end{bmatrix}$$

If extracting the price function from the endogenous system, the functional form is (4).

$$\begin{aligned} \Delta Ph_{it} = \alpha_1 + \Omega_{1..6} [& Ph_{it-1} + \delta_{1,1}d(Pop)_{it-1} + \delta_{1,2}In_{it-1} \\ & + \delta_{1,3}hn_{it-1} + \delta_{1,4}ir_{it-1} + \delta_{1,5}d(Stock)_{it-1} + c_1] \\ & + \sum_{i=1}^j \beta_j \Delta X_{it-j} + \mu_{1,t} \end{aligned} \quad (4)$$

where X is a matrix including the variables in the model so that $X=\{ph, \Delta(Pop), Inc, h_n, int, \Delta(stock)\}$, the subscript ‘i’ refers to the city, m is the number of variables in the model ($m = 6$). In specification (3), the matrix expression reflects the structure of the computed system of endogenous equations.

The first component on the right-hand side of Eq. (4) is the long-run relationship. If it exists and is statistically significant, it represents the long-term causal pattern that quantifies how long-term prices contribute to equilibrium convergence in the short run. There can be more than one long-term relationship, so the omega parameter can take different values depending on calculated ‘n’ relationships. Each long-run relationship would be capturing an economic mechanism that acts autonomously on the evolution of prices in each cluster of cities. These mechanisms show permanent effects on the dependent variable. The second component is known as the Error Correction and captures the short-run reactions. This block identifies and quantifies the factors that produce deviations from equilibrium in the short run and have temporary effects. The number of lags is computed as ‘j’. It is possible

to determine the short-term causality through the significance of the lagged error correction term based on the t-test.

Each model is estimated separately for each cluster, in which the test mentioned above is calculated to define the final cluster-model functional form. Models should fulfil the following conditions: (1) Be stable (all roots into the circle) and VEC residual normality test failing to reject the null of multivariate normal. (2) The lowest number of lags, in case of inconclusive VAR lag Order Criteria (including LR, FPE, AIC, SC AND HQ tests) and (3) having the lowest AIC tests of the full model.

When multiple estimated models are fulfilling the conditions, the final model chosen for use in the forecasting step is the better predictions and lower error predicting the out-of-sample data one (trial data are in the three last years). As the objective is to extend the long-term cycle by forecasting the period further to the observed data (predicting the future), when multiple potential models are acceptable, the choice is made by the expert judgement approach [39]. The model is estimated starting in the earlier period (backcast prediction), seeking to identify whether the future model prediction is consistent with the previous long-term cycle observed in the data. This decision is taken to reduce the subjectivity implicit in the judgement approach.

5.2 Forecast Methodology

Model (3) is fitted separately for each group of municipalities required, and the forecast is calculated as in (5).

$$\begin{aligned} \Delta \widehat{Ph}_{t+k} = & \widehat{\alpha}_1 + \widehat{\Omega_{1..n}} [\widehat{Ph}_{t+k-1} + \widehat{\delta_{1,1}} d(\text{Pop})_{t+k-1} + \widehat{\delta_{1,2}} In_{t+k-1} \\ & + \widehat{\delta_{1,3}} h_{nt+k-1} + \widehat{\delta_{1,4}} ir_{t+k-1} + \widehat{\delta_{1,5}} d(\text{Stock})_{t+k-1} + \widehat{c_1}] \\ & + \sum_{i=1}^j \widehat{\beta}_i \Delta X_{t+k-j} \end{aligned} \quad (5)$$

where 'k' is the number of future periods calculated, and the sign $\widehat{}$ (hat) refers to the estimated value. Note that this methodology predicts all the variables and uses each of the predictions in one period to calculate the next period according to the estimated model, where the first prediction is the one made with the parameters set in the base period. These multiple predictions allow assessing whether the future quantification is according to the economic logic. The forecast method is multidirectional based on performing a dynamic-stochastic simulation using the estimated model, following Broyden solver with iterative calculations until reaching the convergence (allowed a maximum of 5000 iterations) at 95% of the confidence interval. The covariance matrix is scaled to equation specified variances, and the system allows for a maximum of 1000 repetitions converging to $(1/10^8)$.

The model is accepted depending on the forecast precision in two steps. The first step evaluates the out-of-sample prediction during the period 2017–2020. The model with better precision (lower error) is the chosen and which fulfils two conditions of having large explanatory capacity (R^2) and lower AIC test. The second condition is when the backcast prediction from 2000 to 2020 gives a long-term cycle with lower errors. The RMA measures errors.

6 Results Empirical Evidence

The entire exercise has produced three groups of predictions estimated over the three cluster methods of the cities explained above. The first corresponds with the naïve forecast based on the arbitrary aggregation of cities. The second is made using clustering based on housing prices time series, and the third corresponds with clustering based on machine learning methods.

Figures 1 and 2 represent the estimation results in a selected number of cities estimated in the second grouping (clustering based on time series) with the dynamic prediction made for the model's backcast and forecast for the period 2004–2026. A

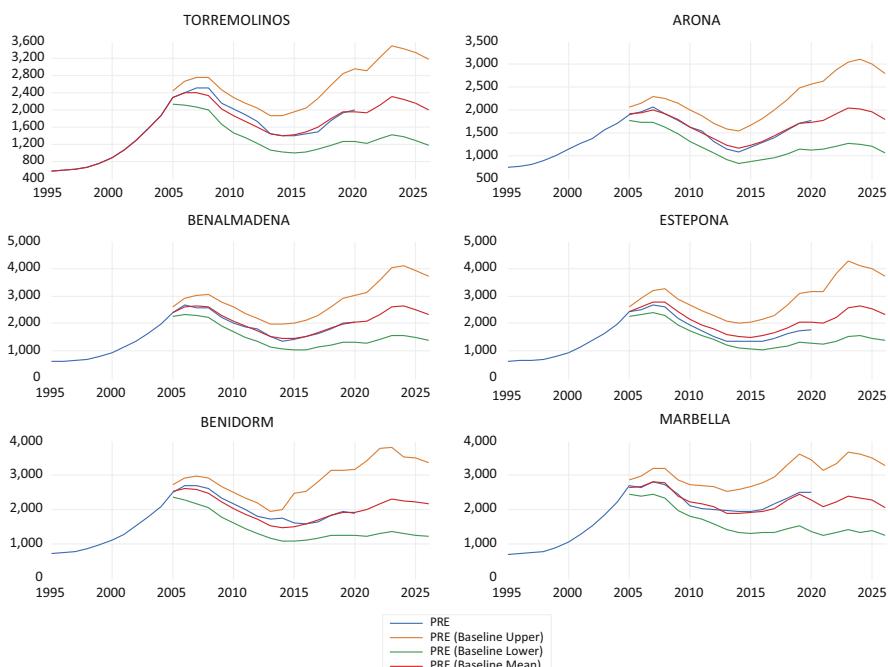


Fig. 1 Price backcast and forecast, cluster 11th of grouping B. Main tourist cities

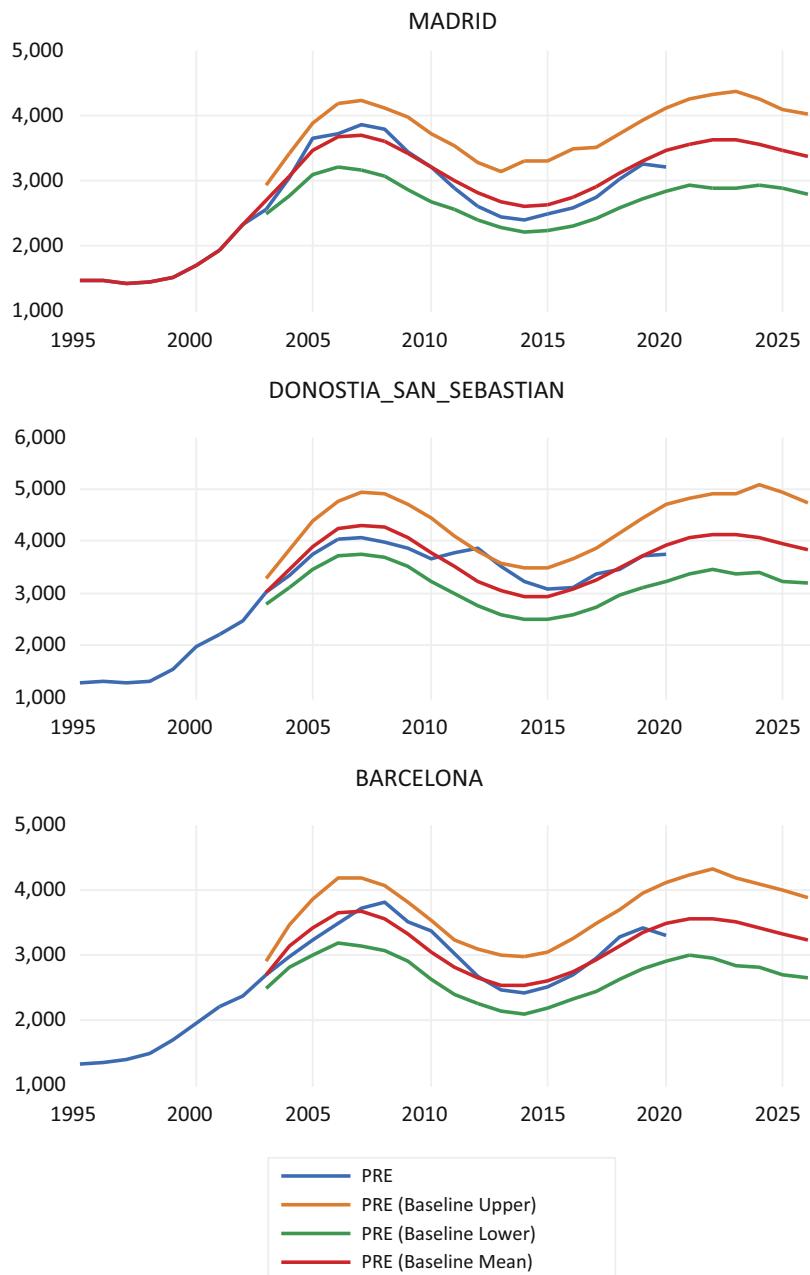


Fig. 2 Price backcast and forecasts, cluster 12th of grouping B. Main capitals

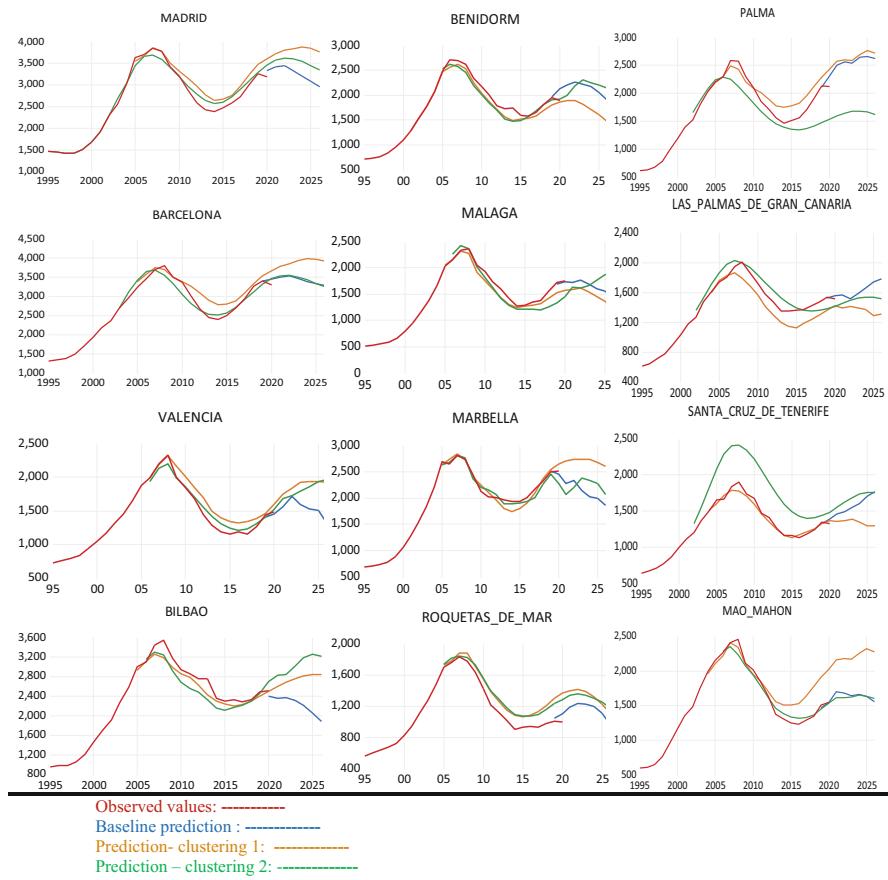


Fig. 3 Forecasting comparison of housing prices in selected cities. Main capitals, touristic and selected cities' results by housing price predictions

comparison of the forecast estimated in the three grouping for selected cities is in Fig. 3.²

6.1 Discussion

Forecast housing prices requires a complex estimation. Firstly, the causal model explaining the long-term trend of housing prices identifies common housing price patterns in a particular group of cities, suggesting that each market receives shocks

² Full results are not presented here due to the limited space. They are available upon request.

from fundamental variables and generates the reaction of prices similarly but pretty closer in each group of cities. Those could be interpreted as that housing prices react to distinct stimuli from fundamentals following a long-term pattern and that patterns represent both economic and idiosyncratic features combination. The exercise demonstrates similar patterns in the housing price reaction across Spain and not necessarily associated with physical proximity and that the prediction is feasible and very accurate in some areas. The prediction errors are found to be the lowest in the two Spanish archipelagos, Bask Country and central cities, which confirms the general belief that proximity (and accessibility as the ripple effect principles support) acts as a convergence channel in housing prices and fundamentals in particular areas with some degree of isolation, although not in others.

The clusters of municipalities calculated reveal group of cities with common reactions of their housing prices to changes in key variables so-called fundamentals. Cluster shows different patterns of price responses between them, but capture those municipalities with close housing price responses whithin them, that is showing those with a common pattern. The aggregation of cities made through statistical tools does capture the common housing market patterns, allowing us to estimate the long-term trend of housing prices and the potential deviations from the equilibrium in each city due to idiosyncratic features. However, it fails in classifying some of the cities accordingly.

In addition, the method is critical to determine the trend in the future. Clustering using the dependent variable time series fits better the data but tends to estimate an upward trend for prices. On the contrary, clustering using machine learning methods combining the six fundamental variables time series gives more smooth long-term cycles. Both clustering methods produce more accurate and closed results than a baseline prediction in general.

This analysis represents the first attempt to estimate the long-term cycle for housing prices in Spain at the municipal level and their forecast, as well as it sets a methodology to advance the trend of prices allowing to prevent future shocks affecting housing prices, with its strong effects on the financial system. It serves to make decisions at a macroprudential policy level. The critical issue to obtain accurate predictions is determining the standard pattern to which every city pertains, which offers better precision.

7 Conclusions

This paper contains an empirical application of a method to forecast the future evolution of residential prices for 270 municipalities in Spain, grouped according to relevant parameters. This paper is the first one providing a precise forecast of long-term housing prices at the municipality level.

The estimation uses annual information from 1995 to 2020 in residential prices, interest rates, and mortgage concessions. It applies non-stationary time-series forecasting methods based on a conventional long-run behavioural model based on

fundamentals, according to the main variables identified by the literature as long-run drivers of housing prices.

The analysis strategy consists of applying a Vector Error Correction model (VECM) to estimate the housing price long-term trend through forecasting and backcasting the observed data. The model allows forecasting 6 years onwards based on the behavioural mechanisms that have been identified by the model chosen with the best functional form (with lower error) fitting the actual data used.

For the modelling, municipalities are grouped. The groups are built, firstly arbitrarily (grouping by proximity), and secondly by calculating clusters based on (1) residential prices, time-series clusters and (2) machine learning approach, using the six time-series variables. Seventeen clusters were chosen in the first grouping method, 8 in the second case and 16 in the latter. In each estimated cluster model, the method identifies two types of influences, on housing prices. The first is the long-run relationships calculated as cointegrating relationships (linear combinations representing a stable and permanent long-run relationship between variables) that determine the fundamental evolution of prices in the long run. The second group identifies the short-term effects of changes in the variables on price developments and is responsible for the equilibrium's price deviation. These short-run components are considered to have transitory effects.

The best model for 1995–2017 is selected for the forecast estimation phase and done in two steps. Firstly, forecasting out-of-sample 2017–2020 period and secondly, with a backcast estimation from 2004 to 2026 applying the estimated parameters. As VECMs are a dynamic system of equations with endogenous variables, their structure allows predicting the actual data. Broyden algorithm has been used, with a maximum of 5000 iterations until convergence is reached in the parameters to obtain a dynamic solution. Results suggest:

- Less accuracy in the models resulting from the first grouping (grouping by proximity). Inaccurate prediction appears in a more significant number of municipalities than in the other two grouping methods.
- The second prediction is more accurate than the third. In most municipalities, it can adjust the evolution of house prices with minor deviations and shallow errors.
- There are a low number of cities where the third estimate is the best accurate. Nevertheless, the third prediction shows better accuracy in the long-term cycle than the second forecast in 40% of cases.

The forecasting exercise reveals the economic mechanisms leading housing markets in the groups of municipalities. In each of them, the behaviour of housing markets in the long and short term has been disentangled, identifying those hidden patterns that act permanently and the sources of short-term price deviations.

The detailed analysis allows identifying two types of reactions across Spanish municipalities that lead the responses of residential prices. These reactions are found in different markets and allow us to understand the dynamics in cities. Interestingly, cities are grouped with others far away, contradicting the principle of proximity to determine housing prices and supporting the evidence of ripple effect [40].

The inference derived from the different mechanisms is consistent and reflects a heterogeneous group of responses that affect the diversity of the mechanisms at work in Spanish housing markets.

References

1. Case, K.E., Quigley, J.M., Shiller, R.J.: Comparing wealth effects: the stock market versus the housing market. *Adv. Macroecon.* **5**(1) (2005)
2. Hwang, M., Quigley, J.M.: Economic fundamentals in local housing markets: evidence from US metropolitan regions. *J. Reg. Sci.* **46**(3), 425–453 (2006)
3. Case, K.E., Shiller, R.J.: Is there a bubble in the housing market? *Brook. Pap. Econ. Act.* **2003**(2), 299–362 (2003)
4. Hwang, M., Quigley, J.M.: Housing price dynamics in time and space: predictability, liquidity and investor returns. *J. Real Estate Financ. Econ.* **41**(1), 3–23 (2010)
5. Clayton, J.: Rational expectations, market fundamentals and housing price volatility. *Real Estate Econ.* **24**(4), 441–470 (1996)
6. Taltavull, P., McGreal, S.: Measuring price expectations: evidence from the Spanish housing market. *J. Eur. Real Estate Res.* **2**(2), 186–209 (2009)
7. Case, K.E., Quigley, J.M.: How housing booms unwind: income effects, wealth effects, and feedbacks through financial markets. *Eur. J. Hous. Policy.* **8**(2), 161–180 (2008)
8. Quigley, J.M.: Urban diversity and economic growth. *J. Econ. Perspect.* **12**(2), 127–138 (1998)
9. Iacoviello, M.: House prices, borrowing constraints, and monetary policy in the business cycle. *Am. Econ. Rev.* **95**(3), 739–764 (2005)
10. Taylor, J.B.: Housing and Monetary Policy (No. w13682). National Bureau of Economic Research (2007)
11. Caldera, A., Johansson, Å.: The price responsiveness of housing supply in OECD countries. *J. Hous. Econ.* **22**(3), 231–249 (2013)
12. Mishkin, F.S.: Housing and the monetary transmission mechanism. NBR working paper, Housing and the Monetary Transmission Mechanism | NBER (2007)
13. Mayer, C., Quigley, J.M.: Is there a bubble in the housing market?. Comments and discussion. *Brook. Pap. Econ. Act.* **2003**(2), 343–362 (2003)
14. Hagemann, D., Wohlmann, M.: An early warning system to identify house price bubbles. *J. Eur. Real Estate Res.* **12**(3), 291–310 (2019)
15. Gyourko, J.: Housing supply. *Annu. Rev. Econ.* **1**(1), 295–318 (2009)
16. Saiz, A.: The geographic determinants of housing supply. *Q. J. Econ.* **125**(3), 1253–1296 (2010)
17. Gyourko, J., Molloy, R.: Regulation and housing supply. In: *Handbook of Regional and Urban Economics*, vol. 5, pp. 1289–1337. Elsevier (2015)
18. Taltavull de la Paz, P.: New housing supply and price reactions: evidence from Spanish markets. *J. Eur. Real Estate Res.* **7**(1), 4–28 (2014)
19. Taltavull de La Paz, P., Gabrielli, L.: Housing supply and price reactions: a comparison approach to Spanish and Italian markets. *Hous. Stud.* **30**(7), 1036–1063 (2015)
20. Glaeser, E.L., Gyourko, J., Saiz, A.: Housing supply and housing bubbles. *J. Urban Econ.* **64**(2), 198–217 (2008)
21. Basu, S., Thibodeau, T.G.: Analysis of spatial autocorrelation in house prices. *J. Real Estate Financ. Econ.* **17**(1), 61–85 (1998)
22. Anselin, L.: GIS research infrastructure for spatial analysis of real estate markets. *J. Hous. Res.* **9**(1), 113–133 (1998)
23. Montero, J.M., Mínguez, R., Fernández-Avilés, G.: Housing price prediction: parametric versus semi-parametric spatial hedonic models. *J. Geogr. Syst.* **20**(1), 27–55 (2018)

24. Anglin, P.M., Gencay, R.: Semiparametric estimation of a hedonic price function. *J. Appl. Econ.* **11**(6), 633–648 (1996)
25. Cajias, M.: Is there room for another hedonic model? The advantages of the GAMLSS approach in real estate research. *J. Eur. Real Estate Res.* **11**(2), 204–245 (2018)
26. Crone, T.M., Voith, R.P.: Estimating house price appreciation: a comparison of methods. *J. Hous. Econ.* **2**(4), 324–338 (1992)
27. Englund, P., Quigley, J.M., Redfearn, C.L.: The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *J. Real Estate Financ. Econ.* **19**(2), 91–112 (1999)
28. Dubin, R.A.: Spatial autocorrelation: a primer. *J. Hous. Econ.* **7**(4), 304–327 (1998)
29. Pagourtzi, E., Makridakis, S., Assimakopoulos, V., Litsa, A.: The advanced forecasting information system PYTHIA: an application in real estate time series. *J. Eur. Real Estate Res.* **1**(2), 114–138 (2008)
30. Gu, J., Zhu, M., Jiang, L.: Housing price forecasting based on genetic algorithm and support vector machine. *Expert Syst. Appl.* **38**(4), 3383–3386 (2011)
31. Rico-Juan, J.R., Taltavull de La Paz, P.: Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Syst. Appl.* **171**, 114590 (2021)
32. Kauko, T., d'Amato, M. (eds.): *Mass appraisal methods: an international perspective for property valuers*. Wiley, Oxford (2008)
33. DiPasquale, D., Wheaton, W.C.: *Urban economics and real estate markets*. Prentice Hall, Englewood Cliffs, NJ (1996)
34. Corduas, M., Piccolo, D.: Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal.* **52**(4), 1860–1872 (2008)
35. Piccolo, D.: A distance measure for classifying ARIMA models. *J. Time Ser. Anal.* **11**(2), 153–164 (1990)
36. Xiong, Y., Yeung, D.Y.: Time series clustering with ARMA mixtures. *Pattern Recogn.* **37**(8), 1675–1689 (2004)
37. Levin, A., Lin, C.F., Chu, C.S.J.: Unit root tests in panel data: asymptotic and finite-sample properties. *J. Econ.* **108**(1), 1–24 (2002)
38. Im, K.S., Pesaran, M.H., Shin, Y.: Testing for unit roots in heterogeneous panels. *J. Econ.* **115**(1), 53–74 (2003)
39. Hyndman, R.J., Athanasopoulos, G.: *Forecasting: Principles and Practice*. OTexts, online <https://otexts.com> (2018)
40. Meen, G.: Regional house prices and the ripple effect: a new interpretation. *Hous. Stud.* **14**(6), 733–753 (1999)
41. Alonso, A.M., Peña, D.: Clustering time series by linear dependency. *Stat. Comput.* **29**(4), 655–676 (2019)
42. ECO/805, ECO/805/2003, Orden de 27 de marzo, sobre normas de valoración de bienes inmuebles y de determinados derechos para ciertas finalidades financieras. Disponible en <https://www.boe.es/eli/es/o/2003/03/27/eco805/con>
43. Gabrielli, L., Taltavull de La Paz, P., Ortúñoz Padilla, A.: Long-term regional house prices cycles. A city-based index for Italy. *J. Eur. Real Estate Res.* **10**(3), 303–330 (2017)
44. Ljungqvist, A., Nanda, V., Singh, R.: Hot markets, investor sentiment, and IPO pricing. *J. Bus.* **79**(4), 1667–1702 (2006)
45. Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA). *Housing Statistics*. <https://www.mitma.gob.es/informacion-para-el-ciudadano/informacion-estadistica/vivienda-y-actuaciones-urbanas/estadisticas/vivienda-y-suelo>. Accessed 01 July 2021
46. Taltavull de La Paz, P., López, E., Juárez, F.: Ripple effect on housing prices. Evidence from tourist markets in Alicante, Spain. *Int. J. Strateg. Prop. Manag.* **21**(1), 1–14 (2017)
47. Taltavull de La Paz, P., White, M.: Fundamental drivers of house price change: the role of money, mortgages, and migration in Spain and the United Kingdom. *J. Prop. Res.* **29**(4), 341–367 (2012)

48. Brueckner Jan K.: Analysing third-world urbanization: a theoretical model with empirical evidence. BEBR faculty working paper; no. 1389 (1987)
49. Goodman Allen C.: Hedonic prices, price indices and housing markets. *J. Urban Econ.* **5.4**, 471–484 (1978)
50. Saiz Albert.: Immigration and housing rents in American cities. *J. Urban Econ.* **61.2**, 345–371 (2007)
51. Poterba, James M.: Tax policy to combat global warming: on designing a carbon tax. (1991)
52. Summers, Lawrence H.: Inflation, the stock market, and owner-occupied housing. No. w0606. National Bureau of Economic Research, (1980)
53. Guren, Adam M., et al.: Housing wealth effects: The long view. *The Review of Economic Studies* **88.2**, 669–707 (2021)

Optimal Combination Forecast for Bitcoin Dollars Time Series



Marwan Abdul Hameed Ashour and Iman A. H. Aldahhan

Abstract Bitcoin has been the most used blockchain platform in business and finance in recent years. This paper aims to find a reliable prediction model that improves a combination of prediction models. Exponential smoothing, ARIMA, artificial neural networks (ANNs) models, and forecasts combination models are among the techniques used in this Paper. The effect of artificial intelligence models in enhancing the results of compound prediction models is the study's most obvious finding. The second major finding was that a model of a robust combination forecasting model that responds to the many variations that occur in the bitcoin time series and Error improvement should be adopted. The results of the prediction accuracy criterion and matching curve fitting in this paper showed that if the residuals of the changed model are white noise, the forecasts are unbiased. A future study investigating robust combination forecasting would be very interesting.

Keywords Exponential smoothing · ARIMA model · ANNs · Combination forecast · Optimization · Robust predictions

1 Introduction

In recent years, there has been a growing interest in forecasting economic and financial time series data which is a challenging task due to uncertain events or incomplete information in current economies. The volatility in time series is high in this situation. Authors have over time applied increasingly sophisticated predicting techniques to predict it more accurately. The most common cryptocurrency in the world is Bitcoin.

M. A. H. Ashour (✉)

Administration and Economics College, University of Baghdad, Baghdad, Iraq

I. A. H. Aldahhan

Continuing Education Center, University of Baghdad, Baghdad, Iraq

Since Bitcoin values are highly volatile, we need to use a robust model. Using many different approaches on the same time series and averaging the resulting predictions is a simple way to increase forecast accuracy.

John Bates and Clive Granger wrote a famous paper, showing that combining forecasts often leads to better forecast accuracy. In recent years, Clemen wrote the results almost universal agreement exists that integrating numerous forecasts improves forecast accuracy. By merely averaging the projections, one may often increase performance dramatically. Time series forecasting may also be done using ARIMA models. The two most generally used techniques for time series forecasting are exponential smoothing, ARIMA, and artificial neural network models, which provide complementary approaches to the problem. ARIMA models try to characterize the autocorrelations in the data, whereas exponential smoothing methods are based on a description of the data's trend and seasonality.

Artificial neural networks (ANNs) are one of the nonlinear models and are the foundation of artificial intelligence (AI). It has the property of self-learning and adaptation. Research efforts on neural networks as forecasting models are commendable, and many studies have reported on the use of ANNs for forecasting. Although some theoretical and empirical issues remain unresolved, the field of neural network forecasting has unquestionably advanced over the last decade. It is not surprising that the next decade will see even more progress and success.

Previous studies have primarily concentrated on Comparison between traditional (exponential smoothing, ARIMA) and modern models (artificial neural network ANN) of forecasting methods to determine the best model, without interest in the fluctuations in time series behavior that may arise in the future [1–7].

The contribution of this study is obvious as the resulting outcomes can be capitalized as guidelines for Comparison between traditional and modern models of prediction methods.

In this paper, the well-known implementation of actual data collected from the period January 2018 to February 2021, which is the closing price data for Bitcoin (digital currency). This paper is divided into five sections: introduction (the objective of the study, and literature review), theoretical, implementation, analysis, and conclusion.

2 Method

2.1 *Exponential Smoothing Model*

Exponential smoothing is a method for forecasting time series based on univariate observations that can be applied to data with a systemic trend or seasonal compound. It is a strong method of forecasting that can be used as an alternative to the common Box-Jenkins ARIMA family of methods.

To forecast data with trends, Holt expanded single exponential smoothing to linear exponential smoothing with trends. Holt's linear exponential smoothing

consists of two constants, α and β (with values between 0,1), and three equations [8–10]:

$$L_t = \alpha y_t + (1 - \alpha) (L_{t-1} + b_{t-1}) \quad (1)$$

$$b_t = \beta (L_t - L_{t-1}) + (1 - \beta) b_{t-1} \quad (2)$$

$$F_{t+m} = I_t + b_t m \quad (3)$$

where:

y_t = time series

L_t :estimation of the time series level at time t.

b_t : an estimate of the slop time series at time t.

F_{t+m} : forecast at time $t + m$.

2.2 Optimization

The majority of exponential smoothing techniques need the definition of several smoothing parameters (constant). These determine how quickly the forecast reacts to changes in data. Because the computer time required to optimize these parameters was so costly, methods involving more than one or two parameters were seldom employed, and parameter values were limited to a narrow number of options. With the introduction of considerably quicker commutes, choosing a nonlinear optimization technique to optimize parameters is rather simple. All competent forecasting software will automatically optimize parameter values [7, 11–13].

2.3 ARIMA Model

A variable's future value is supposed to be a linear function of many past observations and random errors in an autoregressive integrated moving average model. As a result, a nonseasonal time series can be modeled as a mixture of past values and errors, which can be expressed as ARIMA (p,d,q) or as follows [1, 8, 11, 14, 15]:

$$\phi(B)(1 - B)^d y_t = \theta(B)\epsilon_t \quad (4)$$

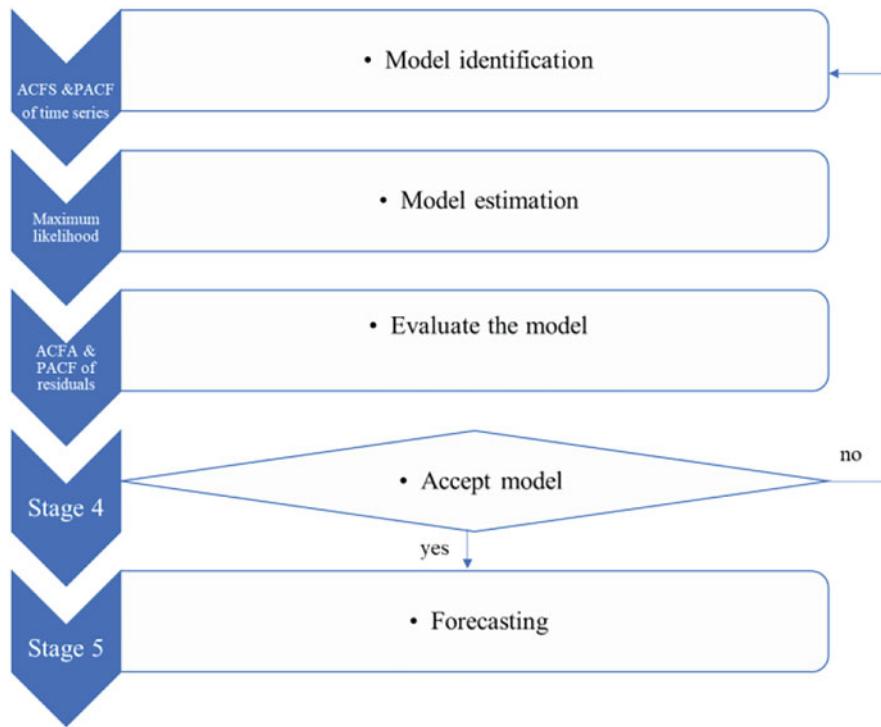


Fig. 1 ARIMA model methodology

where:

y_t : is the time series

ϵ_t : error

B: backward shift operator.

p: order of the autoregressive part.

d: a degree of first differencing involved.

q: order of the moving average part.

Figure 1 illustrates the methodology for the ARIMA model.

2.4 Artificial Neural Network

An artificial neural network (ANN) is a computing framework inspired by a biological neural system and is made up of small, interacting processors known as neurons. Weighted connections bind the neurons, allowing signals to flow through them. Each neuron receives multiple inputs proportional to its contact weights from other neurons and produces a single output that can be propagated to multiple

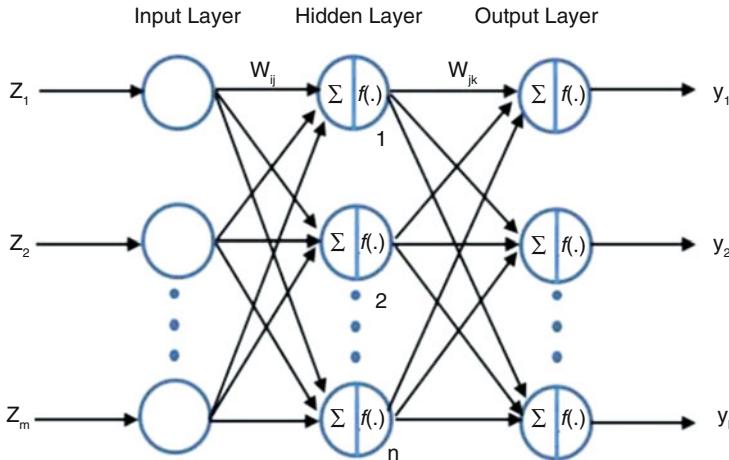


Fig. 2 BP structure

other neurons. An artificial neural network is capable of learning and generalizing relationships in a data set, as well as providing fast and accurate estimates.

In computing applications, the Back Propagation Neural Network (BP) is the most widely used neural network technique. It's a multilayer artificial neural network (ANN) with a feed-forward connection from the input layer to the hidden layers and finally to the output layer. The BP algorithm aims to reduce the mean square error between the forecast and desired outputs. Figure 2 shows the structure of the BP.

2.5 Forecast's Combination

Combining data enhances predicting accuracy without a doubt. This empirical observation holds when it comes to statistical forecasting, judging estimations, and averaging statistical and subjective forecasts. Also, combining results in a significant reduction in the variance of post-sample forecasting inaccuracy. The empirical findings contradict the statistical theory, requiring a reassessment of what constitutes effective forecasting methods and how they should be used [3, 8, 16].

Forecasting is combined as well as the best mix of forecasting. Furthermore, the root means square error (RMSE), which measures the variance or level of uncertainty in our forecast, is lower with a simple combination than with either the individual approach or the best combination. The rationale for this is that the average reduces the RMSE by canceling big prediction errors. The fact that a simple combination decreases the RMSE of the post-sample forecast is another incentive to use it in reality; reduced error equals less uncertainty, which translates to smaller inventories and, thus, minimizes costs [1, 3, 6, 8].

2.6 Measuring Forecast Accuracy

In most forecasting situations, accuracy is treated as an overarching criterion for the selection of the forecasting method. In many cases, the word “accuracy” refers to “goodness of fit,” which in turn refers to the extent to which the forecasting model can reproduce the data already known to the forecast consumer. The accuracy of the future forecast is the most important thing. The most common measure of error accuracy is [1, 8, 10, 16, 17]:

Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_t - f_t)^2}{n}} \quad (5)$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{|y_t - f_t|}{y_t}} \quad (6)$$

$$R^2 = 1 - \frac{RSS}{TSS} \quad (7)$$

where:

n : number of observations (number of non-missing data points)

y_t : actual value

f_t : forecast value

R^2 : coefficient of determination

RSS: sum of squares of residuals.

TSS: the total sum of squares.

3 Results and Discussion

Figure 3 presents the time series from 1/1/2018 to 18/2/2021 for bitcoin’s daily closed price (Source data: the wall street journal website).

As shown in Fig. 3 there is a clear trend in time series, the closed price of bitcoin is rising significantly between the end of 2020 and the end of 2021, with a slight decline at the end of February. Because of the high level of volatility in the market, such as during the COVID-19 pandemic, there is inconsistency in the actions of the bitcoin closing price sequence. The following are the forecasting methods’ findings for the time series under study:

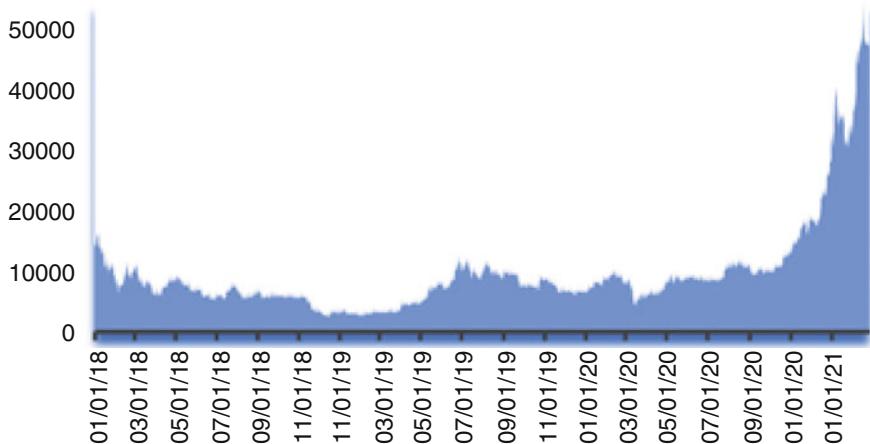


Fig. 3 Closed price bitcoin series

Table 1 Optimal parameter of Holt model

Optimal parameter	Estimate
α	0.98
β	0.18

Table 2 Model fit statistics of Holt model

Model fit statistics		
R ²	RMSE	MAPE
0.994	600.864	2.63

3.1 Exponential Smoothing Model Result

Figure 3 illustrates the series has a linear trend, so the best model is Holt's linear exponential smoothing. Use a statistical program (V.12) to find optimal parameters, and the results are as follows (Table 1).

The results of the evaluation of this model were as follows (Table 2):

The finding of the present study suggests that it is clear from the result that significant parameter and the ACF and PACF coefficients of the residues are random behavior and white noise. This model is appropriate and the best (Figs. 4 and 5).

3.2 ARIMA Model Result

The best ARIMA model for the time series under research is ARIMA (1,0,0), based on the analysis of ACF and PACF. The ARIMA model was computed using SPSS (V.23), and the model parameter estimates are shown in Table 3.

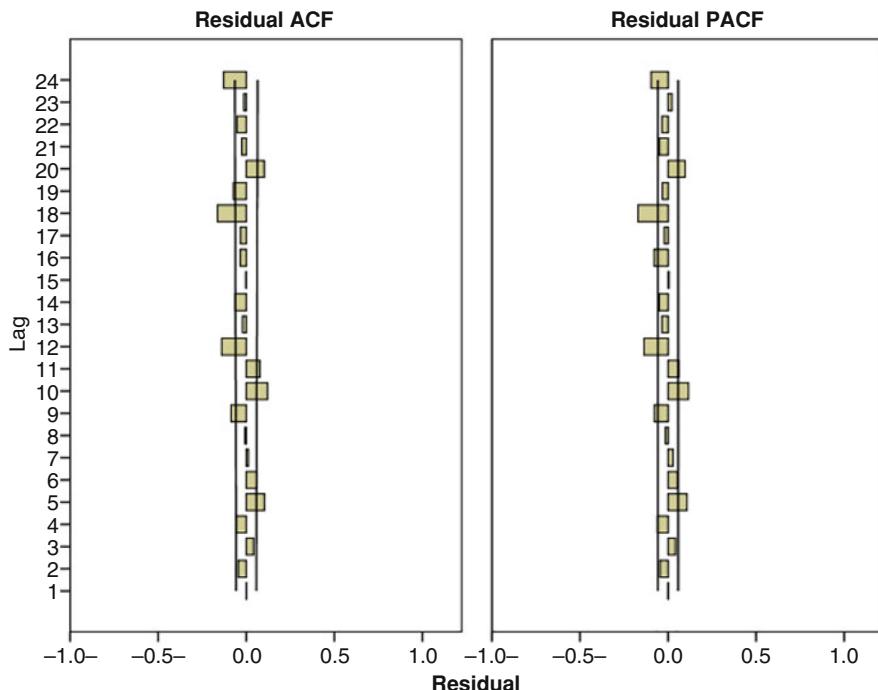


Fig. 4 Autocorrection and partial autocorrection function for residual

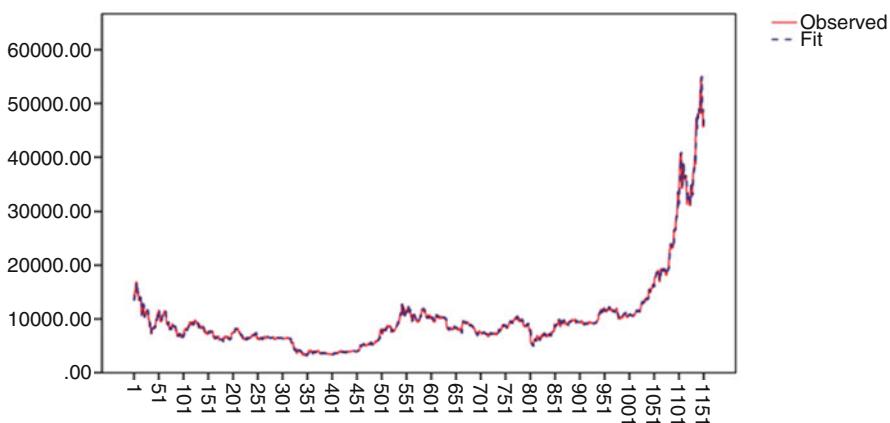


Fig. 5 Curve fitting of exponential smoothing model

Table 3 ARIMA model parameters

	Estimate	SE	t	Sig.
	0.96	.002	634.324	.00

Table 4 Model fit statistics of ARIMA model

R ²	RMSE	MAPE
0.994	605.626	2.64

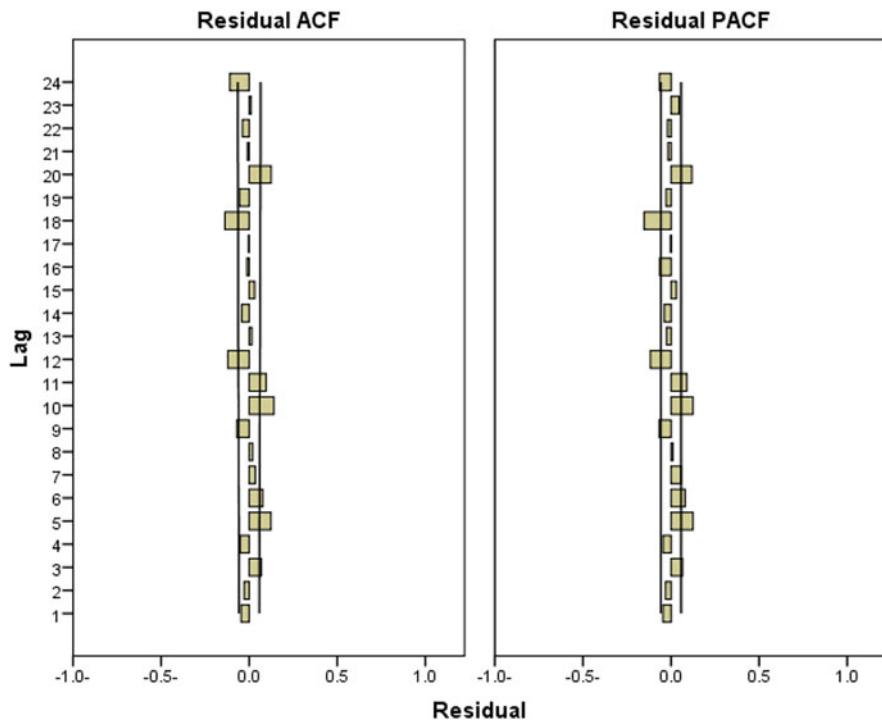


Fig. 6 Autocorrection and partial autocorrection function for residual

The results of the evaluation of this model were as follows (Table 4):

Thus, the mathematical model of the ARIMA model is according to the following formula:

$$y_t = 0.99 y_{t-1} + \epsilon_t \quad (8)$$

The findings suggest that it was found that the ACF and PACF of the residuals are random behavior and white noise, and the significance of the Ljung-Box test, and curve fitting (Figs. 6 and 7), the estimated model is the best.

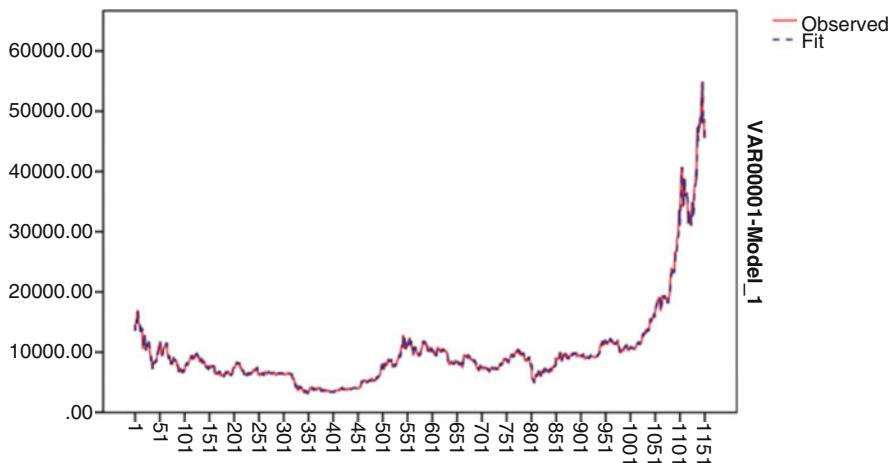


Fig. 7 Curve fitting of ARIMA model

Table 5 Accuracy error of ANN model

R ²	RMSE	MAPE
0.997	502.95	2.53

3.3 Artificial Neural Networks

MATLAB was used to provide the following results for the time series under search from the BP:

Input layer: This layer has one node, which is represented by the variables y_{t-1} , with a one-degree time series lag.

Hidden layer: The maximum number of nodes in this layer is 15, and there is only one layer (after several trials).

The output layer consists of only one node, which is represented by the y_t vector. Figures 3 and 4 show the performer's performance for the BP network for the time series under consideration. Table 5 shows the results of the error evaluation by comparing the two methods used.

Time series response and the response of output element for time series are shown in Fig. 8.

From the data in Fig. 8, it is apparent that network evaluation results are significant. It appears from Table 5 that the error accuracy results are good (Fig. 9).

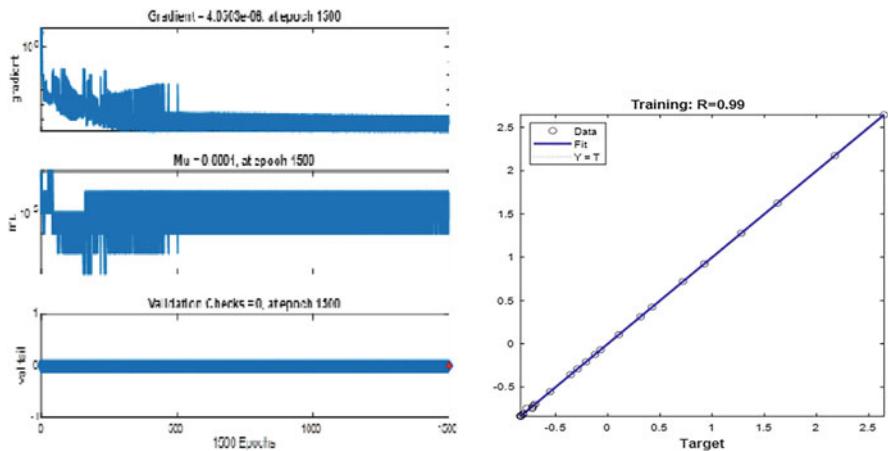


Fig. 8 Performance of ANN

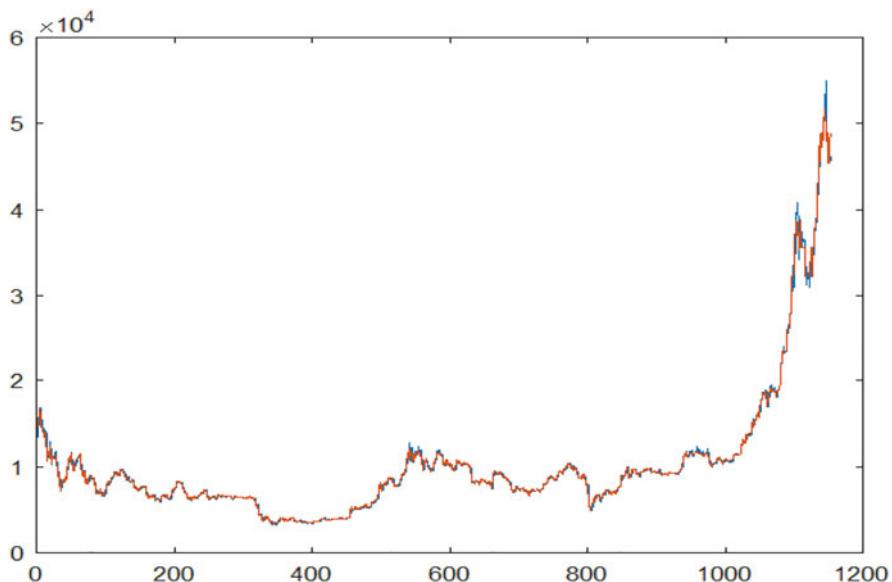


Fig. 9 Curve fitting of ANN

3.4 Combination Model Result

In this section two combination models will be used, the first includes the traditional models only and the second includes the traditional and modern models.

Model I

The first model includes ARIMA and exponential smoothing, based on the error weight of each model. Therefore, the mathematical model is as follows:

$$c_{t+1} = w_1 \text{ARIMA} (1, 0, 0) + w_2 \text{EXP} \quad (9)$$

$$\begin{aligned} c_{t+1} &= 0.49 (0.99 y_t) \\ &+ 0.51 (0.98 y_t + 0.02 (I_{t-1} + b_{t-1}) + 0.18 (I_t - I_{t-1}) + 0.82 b_{t-1}) \end{aligned} \quad (10)$$

Table 6 demonstrates that the evaluation of model I.

Model II

The combination model combines ANN, ARIMA, and exponential smoothing, based on the error weight of each model. Therefore, the mathematical model is as follows:

$$c_{t+1} = w_1 \text{ANN} + w_2 \text{ARIMA} (1, 0, 0) + w_3 \text{EXP} \quad (11)$$

$$\begin{aligned} c_{t+1} &= 0.34 [t] + 0.32 (0.99 y_t) \\ &+ 0.34 (0.98 y_t + 0.02 (I_{t-1} + b_{t-1}) + 0.18 (I_t - I_{t-1}) + 0.82 b_{t-1}) \end{aligned} \quad (12)$$

where:

t: output of ANN

Table 6 provides the results of the evaluation of this model.

As Table 6 shows, there is a significant difference between the two models. The combination model that includes artificial neural network models enables error minimization.

The author found that modern models (ANN) have improved error results, which is in good agreement with the results of the present study. The finding provides evidence of the efficiency of the results of traditional and modern models.

Table 7 demonstrates the prediction values for the next 25 days.

Table 6 Fit statistics of two combination model

Models	Criteria		
	R ²	RMSE	MAPE
Model I	0.994	603.25	2.63
Model II	0.995	467.09	2.59

Table 7 Forecasting values

Period	Forecast
02/27/21	48,476.98
02/28/21	48,789.19
03/01/21	49,101.41
03/02/21	49,413.62
03/03/21	49,725.83
03/04/21	50,038.04
03/05/21	50,350.25
03/06/21	50,662.47
03/07/21	50,974.68
03/08/21	51,286.89
03/09/21	51,599.1
03/10/21	51,911.32
03/11/21	52,223.53
03/12/21	52,535.74
03/13/21	52,847.95
03/14/21	53,160.17
03/15/21	53,472.38
03/16/21	53,784.59
03/17/21	54,096.8
03/18/21	54,409.02
03/19/21	54,721.23
03/20/21	55,033.44
03/21/21	55,345.65
03/22/21	55,657.86
03/23/21	55,970.08

4 Conclusion

Important conclusions drawn from this work include:

1. These findings suggest that in general all traditional and modern methods are competitive and have proven to be efficient.
2. The relevance of the combination model is supported by the current findings.
3. The results of this study indicate that artificial neural network models minimize error and improve the results of the model compound.
4. The results of this investigation show that residual behavior is white noise.
5. The results of this study also suggest that the combination model with ANN is best the model.
6. The results presented here may facilitate improvements in the forecasting and adopt a model of a robust forecasting model that replies to the many fluctuations that occur in the bitcoin time series.

References

1. Ashour, M.A.H., Al-Dahhan, I.A.H.: Turkish lira Exchange rate forecasting using time series models. (2020).
2. Azari, A.: Bitcoin price prediction: an ARIMA approach. arXiv Prepr. arXiv1904.05315 (2019).
3. Bates, J.M., Granger, C.W.J.: The combination of forecasts. *J. Oper. Res. Soc.* **20**(4), 451–468 (1969)
4. Derbentsev, V., Datsenko, N., Stepanenko, O., Bezkorovainyi, V.: Forecasting cryptocurrency prices time series using machine learning. In: CEUR Workshop Proceedings, pp. 320–334 (2019)
5. Munim, Z.H., Shakil, M.H., Alon, I.: Next-day bitcoin price forecast. *J. Risk Financ. Manag.* **12**(2), 103 (2019)
6. Paterson, S.J.C.: A comparison between 8 common cost forecasting methods. (2018).
7. Wirawan, I.M., Widyaningtyas, T., Hasan, M.M.: Short term prediction on bitcoin price using ARIMA method. In: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), pp. 260–265 (2019)
8. Abdul, M., Ashour, H., Al-dahhan, I.A.H., Hassan, A.K.: Forecasting by Using the Optimal Time Series Method. Springer International Publishing (2020). <https://doi.org/10.1007/978-3-030-44267-5>
9. Chen, Y., Yang, B., Dong, J., Abraham, A.: Time-series forecasting using flexible neural tree model. *Inf. Sci. (NY)*, **174**(3–4), 219–235 (2005)
10. Oancea, B., Ciucu, S.C.: Time series forecasting using neural networks. arXiv Prepr. arXiv1401.1333 (2014).
11. Abdul, M., Ashour, H., Jamal, A., Alayham, R., Helmi, A.: Effectiveness of Artificial Neural Networks in Solving Financial Time Series. *October* (2018). <https://doi.org/10.14419/ijet.v7i4.11.20783>
12. Ashour, M.A.H., Abbas, R.A.: Improving time series' forecast errors by using recurrent neural networks. In: Proceedings of the 2018 7th International Conference on Software and Computer Applications, pp. 229–232 (2018)
13. Guiné, R.P.F., Matos, S., Gonçalves, F.J., Costa, D., Mendes, M.: Evaluation of phenolic compounds and antioxidant activity of blueberries and modelization by artificial neural networks. *Int. J. Fruit Sci.* **18**, 1–16 (2018)
14. Tang, Z., De Almeida, C., Fishwick, P.A.: Time series forecasting using neural networks vs. Box-Jenkins methodology. *Simulation*, **57**(5), 303–310 (1991)
15. Wang, K.W., Deng, C., Li, J.P., Zhang, Y.Y., Li, X.Y., Wu, M.C.: Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. *Epidemiol. Infect.* **145**(6), 1118–1129 (2017). <https://doi.org/10.1017/S0950268816003216>
16. Abbas, R.A., Jamal, A., Ashour, M.A.H., Fong, S.L.: Curve fitting prediction with artificial neural networks: a comparative analysis. *Period. Eng. Nat. Sci.* **8**(1), 125–132 (2020)
17. Castel, S., Burr, W.: Assessing statistical performance of time series interpolators. *Eng. Proc.* **5**, 57 (2021). <https://doi.org/10.3390/engproc2021005057>

The Impact of the Hungarian Retail Debt Program



An Estimation of the Past and Future Effects of the Retail Sector on Hungarian Public Debt

Bianka Biró, Dávid Tran, András Stark, and András Bebes

Abstract This paper presents an analysis of both the past and future of the Hungarian retail debt program from a cost-risk standpoint. A quarter of the Hungarian central government debt is held through retail securities. From purely a nominal coupon point of view and analyzed in isolation, retail debt seems to be a comparatively more expensive form of funding. The paper has two goals. First, to estimate the historical cost of the retail debt program compared to alternative domestic sources of funding, taking portfolio effects and risks into account. Second, to simulate the future effects of retail debt based on security-level transaction data and a Vector Error Correction macroeconomic model in order to utilize quantitative tools for the perspective rethink of the retail debt strategy once the current strategic objectives are achieved in the near future.

Keywords Public debt · Retail debt program · Household assets · Macroeconomics forecasting · Vector error correction model

1 Introduction

The mission of the Hungarian Government Debt Management Agency (“ÁKK”) is to finance Hungary’s central government (“CG”) debt at the lowest possible cost with acceptable risks. ÁKK is acting on behalf of Hungary when managing Hungary’s CG debt, that amounted to HUF 36,684 billion or 77% of the GDP at the end of 2020. Of this debt, approximately 26% was in retail securities, a sizeable increase compared to the 2.3% figure of end-2011. Simultaneously, during this period the share of FX debt decreased from nearly 50% to under 20%. This shift, however, required Hungary to pay a higher interest rate for retail debt compared

B. Biró · D. Tran · A. Stark · A. Bebes (✉)
Government Debt Management Agency Pte. Ltd., Budapest, Hungary
e-mail: bebes.andras@akk.hu; strategia@akk.hu

to domestic wholesale government securities with similar maturities. The average interest of HUF retail debt was approximately 4.2% compared to the 2.5% of domestic wholesale domestic bonds, while the average term-to-maturity of the former was only 3.1 years as opposed to the 5.6 years of wholesale domestic bonds as of December 31, 2020.

However, due to limited demand for domestic wholesale bonds taking into account the Government's strategic objective to gradually reduce foreign participation, it has to be investigated whether the realistic alternative to retail financing is wholesale domestic bonds or FX debt. The latter is a higher risk alternative compared to the perceived stability and diversification advantage of having the retail sector as a significant source of financing.

The goal of this paper is twofold. First, to estimate the historical cost of the retail debt program compared to alternative sources of funding, taking portfolio effects, and risks into account. Second, to calculate the future effects of the Hungarian retail debt program from the end of 2020 to the end of 2025.

2 The Hungarian Retail Debt Program

2.1 Main Objectives

The Government considers the expansion of outstanding debt owned by households a key objective since 2012. The share of retail securities dropped from the 7.3% figure of 1999 to 2.3% by end-2011 due to competing investment products and the relatively low interest rates of retail securities compared to the domestic wholesale market.

The primary goals of the retail debt program are to improve the attitude of the household sector regarding savings and investment, diversify the investment base of government debt in order to make funding more diverse and secure, reduce FX exposure, and comparatively reduce the reliance on the wholesale HUF government bond market.

In order to reverse the 1999–2012 downwards trend, retail debt had to be made into an attractive investment opportunity through product development, better and cheaper accessibility through financial institutions and also directly via the Hungarian State Treasury's branch network and electronic platforms as well as higher yields compared to competing investment products. Consistently providing positive real interest rates for the household sector has been the primary consideration behind the pricing of retail securities to ensure financial inclusion of the widest possible range of retail investors. This economic policy goal, against the backdrop of the prevailing negative real interest rate environment globally are the two main factors behind the higher cost of retail debt compared to the domestic wholesale market rates.

The favorable conditions and security of retail debt also offer households a viable alternative to holding cash or bank deposits. Reducing the sizeable and unproductive cash reserves of the household sector is another important objective of economic policy that the retail debt program can support. Therefore, in this analysis, it is assumed that demand for retail instruments is the primary factor driving the outstanding retail debt, and ÁKK readily accepts all demand without further consideration.

2.2 The Retail Debt Portfolio

The Hungarian retail debt portfolio consists of several different instruments. The flagship product MÁP+ is a 5-year step-up coupon bond with a 4.95% internal rate of return (if held until maturity) that can be redeemed at face value at interest payment dates. Interest is automatically reinvested into the bond. PMÁP is an inflation-linked bond with 3- and 5-year tenors. They at present pay a 1% and a 1.25% premium over Hungarian CPI, respectively. They have 3- and 5-year EUR-denominated equivalents (“PEMÁP”) as well with a premium over the Euro Area CPI. A 1-year security (“1MÁP”) is also available. It is Pareto-dominated by the MÁP+ from an investor standpoint, and yet, its demand is significant. Three materialized (printed) securities exist as well, one equivalent to the MÁP+ (“NYMÁP+”) as well as a 1- and a 2-year security (“KTJ1, KTJ2”), both largely inferior to the MÁP+ or its printed version. Rounding out the portfolio is the Baby Bond (“BABA”), an inflation-linked bond with a 3% premium that can be bought for children under 18. Approximately 90% of the retail debt portfolio consists of MÁP+, PMÁP, and 1MÁP, with MÁP+ being 53% by itself.

3 The Historical Cost of Retail Debt

By 2012, the share of retail securities in the debt portfolio dropped to 2.3%. Concurrently, the share of government securities in the total assets of households was also 2.3%, down from the 5.4% figure of 2008. It is safe to assume that without an improved retail debt program consisting of higher interest rates, product development and extra marketing activities, the share of government securities in household assets would not have reached the end-2020 figure of 13.8%.

3.1 Methodology

ÁKK conducted a what-if simulation based on these assumptions and historical Hungarian Government Bond auction data. Since the goal of the simulation was to estimate the real extra cost of the retail debt program, ÁKK aimed to compare the realized costs of retail and domestic wholesale debt with the costs of a hypothetically lower (in both outstanding amount and interest rates) retail debt and extra issuances in the domestic wholesale market. The simulation consists of two main phases. First, it was necessary to simulate the retail debt without the improved retail debt program along with its costs. Second, based on the funding gap left by lower retail issuances, ÁKK simulated the modified domestic wholesale issuances considering the new, mostly higher amounts and yields. As the reduction of FX debt was a strategic goal of the government, no extra FX issuance is considered.

ÁKK assumed that in the absence of an improved retail program, starting from 2012, the initial 2.3% proportion of government securities in household assets would have risen linearly to 5.4% by the end of 2014 and would have remained unchanged until the end of 2020. According to this assumption, the amount of government securities owned by households would have reached HUF 3,563 billion by the end of 2020, leaving a gap of HUF 5,555 billion compared to the factual value of HUF 9,118 billion, as shown in Fig. 1.

Assuming also that the retail debt structure would have not changed over the years in the absence of product development, it is possible to estimate the hypothetical costs of retail financing as well.

The difference between the observed and hypothetical retail debt volume represents the surplus that would have had to be financed through the domestic wholesale market. The conducted simulation is based on ÁKK's historical auction data from 2012 to 2020, containing the bid amounts and yields of primary dealers with the accepted amounts and yields for each auctions. In the analysis, ÁKK simulated the possible domestic wholesale issuances that would have been necessary without the enhanced retail financing, also considering the impact of higher issued amounts

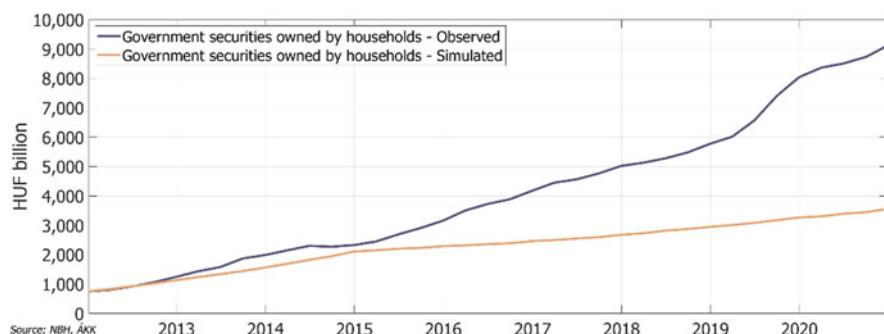


Fig. 1 Volume of government securities owned by households. *Source:* NBH, ÁKK

on the yields. The simulation has three main steps. First, an adjustment was made on the bids by taking into account that a lower retail debt volume would have resulted in a higher volume of cash and bank deposits, increasing the demand for government bonds from the primary dealers' side. Second, the new accepted amounts and the corresponding yields were calculated based on the adjusted bids and increased financing need. Third, an adjustment was added to the new yields, with a consideration that higher auction yields for several consequent auctions could have had an increasing effect on the bid yields of the following auctions, resulting in even higher auction yields.

When estimating the possible volume of cash and bank deposits, it was assumed that similarly to other retail asset types, it could have followed a similar quadratic trend as the total household assets. Given that between 2012 and 2020, the volume of total household assets can be estimated by the polynomial

$$F(x) = 16x^2 + 358.52x + 32,044, \quad (1)$$

where x is the number of quarters starting from 2012 (Fig. 2).

Based on historical data, it was assumed that the share of cash and bank deposit volume could have stayed around 30% of the total household assets. The simulated volume can be given by

$$D^S(x) = 4.8x^2 + 107.56x + 10,681. \quad (2)$$

Denoting the observed cash and bank deposit volume by D and the applicable adjustment by r , the latter is given by

$$r(i) = \frac{D^S(i)}{D(i)} \quad (3)$$

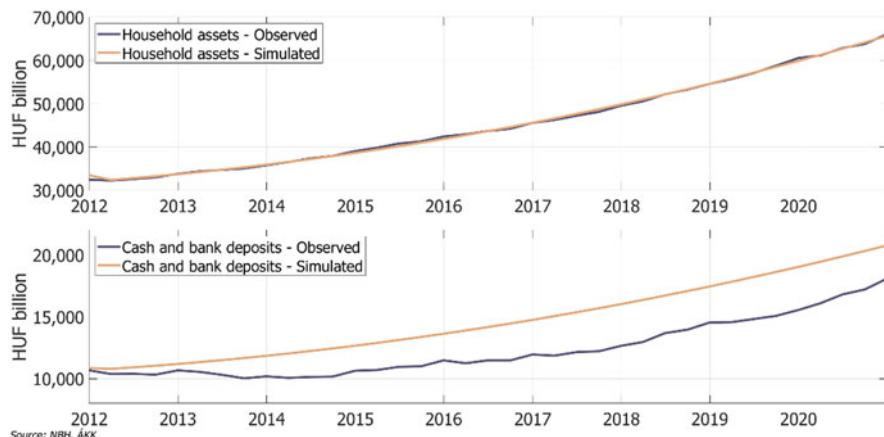


Fig. 2 Observed and simulated household assets, cash and bank deposits. Source: NBH, ÁKK

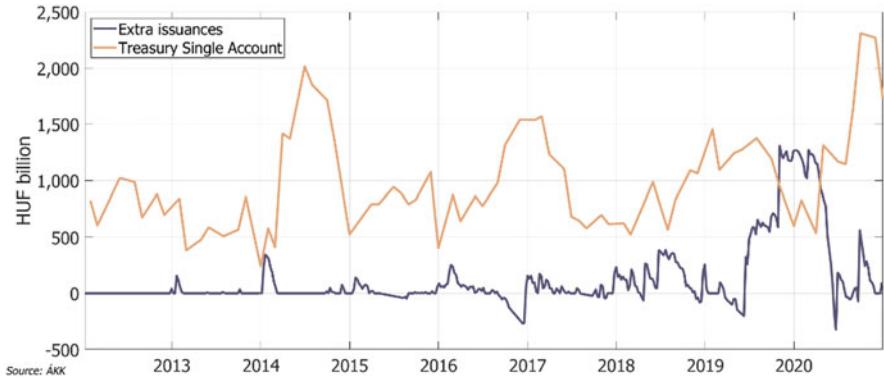


Fig. 3 Extra issuances vs. the Treasury Single Account. *Source:* ÁKK

at the i^{th} auction. Suppose that there are d primary dealers, m papers. Let us denote the original submitted bid amount of primary dealer k for paper j on the i^{th} auction by $A_B(i, j, k)$. Thus, the adjusted bid amounts can be given by

$$\tilde{A}_B(i, j, k) = r(i) * A_B(i, j, k), \quad (4)$$

Let $I_E(i)$ denote the extra amount of bonds that needs to be issued at auction i , $A_C(i, j)$ the original accepted amount, and $I_P(i, j)$ the possible amount that can be issued in addition to paper j according to the adjusted bid amounts. Then let

$$I_P^{Total}(i) = \sum_{j=1}^m I_P(i, j) = \sum_{k=1}^d \tilde{A}_B(i, j, k) - A_C(i, j). \quad (5)$$

The new accepted amounts of paper j at auction i can be calculated as

$$\tilde{A}_C(i, j) = A_C(i, j) + q(i, j) * \min(I_E(i), I_P^{Total}(i)), \quad (6)$$

where $q(i, j) = \frac{I_P(i, j)}{I_E(i)}$.

It is important to note that as the comparison of I_E and the Treasury Single Account (the cash account of the Hungarian State) shows (Fig. 3), the simulated domestic wholesale demand would not have always been enough to cover the gap left by the absence of an improved retail program. However, due to the limitations¹

¹ Due to the success of MÁP+ in mid-2019, ÁKK modified its financing plan. The bids of primary dealers may have been higher without this modification, which is not reflected in the model. Furthermore, the November 2019 spike in I_E is due to a large extra (simulated) issuance of 2016 maturing. In reality, a different maturity profile could have been constructed with a slightly longer maturity bond.

of the methodology, this does not mean that an alternative source of funding (FX) would have been required without an improved retail program.

The new accepted amounts determine the new yields as

$$\tilde{Y}_C(i, j) = \frac{\sum_{k=1}^h Y_B(i, j, k) \tilde{A}_B(i, j, k)}{\sum_{k=1}^h \tilde{A}_B(i, j, k)} \quad (7)$$

where $h = \text{argmin} \left(\sum_{k=1}^m \tilde{A}_B(i, j, k) \geq \tilde{A}_C(i, j) \right)$.

Since issuing bonds with higher yields for several auctions could have also indicated higher bid yields on the following auctions, ÁKK adjusted the new auction yields with this effect. The time series of the yields were modeled by autoregressive processes. As the ADF [1] tests pointed out, both the original and new yields are integrated in the first order. To achieve stationarity, linear trends were removed from the time series by setting a breakpoint to early 2017 when the decreasing trend stopped. Let \tilde{Y}_C^D denote the time series of the detrended original auction yields and \tilde{Y}_C the detrended new auction yields. ÁKK considered the AR(3) to be the best model intuitively but in case of tenors 1 and 5, AR(2) was found to be a better fit according to the Akaike [2] and Bayesian [3] information criteria and the significance of parameters (Table 1).

The AR model parameters represent the innovation dynamics of the yields with respect to the previous auctions, while the $\varepsilon \sim N(0, 1)$ i.i.d. random process represents the effect of market events, which are considered to be independent from the retail debt program. Let $Y_C(i, j)$ denote the original accepted yield of a bond in tenor j at auction i , and $\nabla Y(i, j) = \tilde{Y}_C(i, j) - Y_C(i, j)$. In order to capture the rollover effect of ∇Y on the level of yields and incorporate the innovation dynamics with the observed market events, the yields were adjusted as shown in Table 2, where \hat{Y} denotes the estimated yields after the adjustment.

ÁKK applied the estimated AR parameters and added ∇Y recursively, starting from the initial values of the new auction yields. With this approach, it was possible

Table 1 The selected AR models for the original yields

Tenor	Model	
1Y	AR(2)	$Y_C^D(i, 1) = 1.24Y_C^D(i - 1, 1) - 0.3Y_C^D(i - 2, 1) + \varepsilon(i, 1)$
3Y	AR(3)	$Y_C^D(i, 3) = 0.98Y_C^D(i - 1, 3) - 0.15Y_C^D(i - 2, 3) + 0.12Y_C^D(i - 3, 3) + \varepsilon(i, 3)$
5Y	AR(2)	$Y_C^D(i, 5) = 0.85Y_C^D(i - 1, 5) - 0.07Y_C^D(i - 2, 5) + \varepsilon(i, 5)$
10Y	AR(3)	$Y_C^D(i, 10) = 0.93Y_C^D(i - 1, 10) - 0.17Y_C^D(i - 2, 10) + 0.16Y_C^D(i - 3, 10) + \varepsilon(i, 10)$
15Y	AR(3)	$Y_C^D(i, 15) = 0.9Y_C^D(i - 1, 15) - 0.23Y_C^D(i - 2, 15) - 0.29Y_C^D(i - 3, 15) + \varepsilon(i, 15)$

Source: ÁKK

Table 2 Estimated yield adjustments

Tenor	Estimation
1Y	$\hat{Y}(i, 1) = 1.24\hat{Y}(i - 1, 1) - 0.3\hat{Y}(i - 2, 1) + \nabla Y(i, 1) + \varepsilon(i, 1)$
3Y	$\hat{Y}(i, 3) = 0.98\hat{Y}(i - 1, 3) - 0.15\hat{Y}(i - 2, 3) + 0.12\hat{Y}(i - 3, 3) + \nabla Y(i, 3) + \varepsilon(i, 3)$
5Y	$\hat{Y}(i, 5) = 0.85\hat{Y}(i - 1, 5) - 0.07\hat{Y}(i - 2, 5) + \nabla Y(i, 5) + \varepsilon(i, 5)$
10Y	$\hat{Y}(i, 10) = 0.93\hat{Y}(i - 1, 10) - 0.17\hat{Y}(i - 2, 10) + 0.16\hat{Y}(i - 3, 10) + \nabla Y(i, 10) + \varepsilon(i, 10)$
15Y	$\hat{Y}(i, 15) = 0.9\hat{Y}(i - 1, 15) - 0.23\hat{Y}(i - 2, 15) - 0.29\hat{Y}(i - 3, 15) + \nabla Y(i, 15) + \varepsilon(i, 15)$

Source: ÁKK

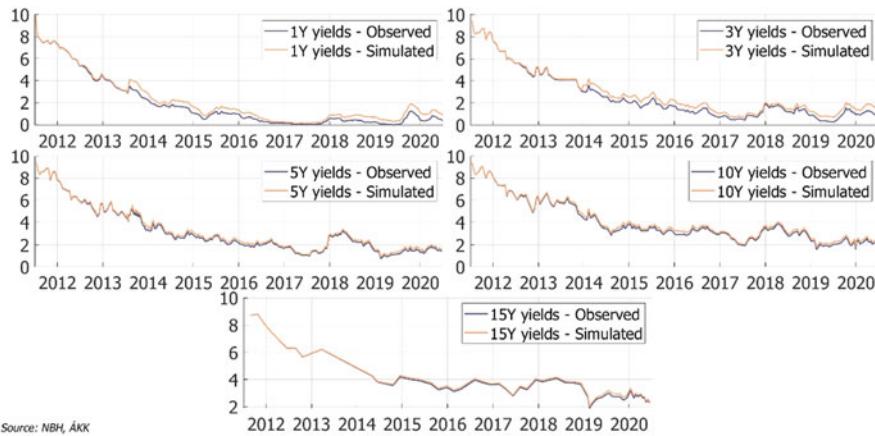


Fig. 4 Observed and simulated historical yields. Source: NBH, ÁKK

to calculate the yields for each auction so that the higher levels had already been incorporated in the previous yields.

Finally, the earlier removed trends were added back, resulting in the simulated yields which are illustrated in Fig. 4 in comparison with the original observed yields.

It can be observed that the difference between the observed and simulated yields are higher in case of short-term bonds than long-term bonds. The reason behind this is that demand for short-term bonds is more flexible, therefore most of the extra financing need was covered by these two tenor segments.

3.2 Results

The annual costs of the domestic wholesale issuances were calculated based on the conducted simulation. In order to compare them appropriately with the retail results, ÁKK considered the cumulative annual domestic wholesale costs, meaning that the

Table 3 Cost effects proportional to GDP

	Fact cost	Simulated cost	Retail effect	Wholesale effect	Total effect
2012	4.56%	4.62%	-0.05%	0.00%	-0.05%
2013	4.52%	4.60%	0.01%	-0.09%	-0.08%
2014	3.97%	4.20%	0.13%	-0.35%	-0.23%
2015	3.44%	3.55%	0.25%	-0.36%	-0.11%
2016	3.09%	3.22%	0.32%	-0.45%	-0.13%
2017	2.65%	2.63%	0.40%	-0.38%	0.02%
2018	2.33%	2.28%	0.46%	-0.41%	0.05%
2019	2.23%	2.03%	0.59%	-0.39%	0.20%
2020	2.36%	2.21%	0.69%	-0.54%	0.15%
Total	3.09%	3.10%	0.35%	-0.35%	0.00%

Source: ÁKK, HCSO

cost calculated for a specific year includes the costs of those bonds as well that were issued in the previous years and have not matured until the year in question.

As Table 3 shows, there is no significant difference in financing costs on this 9-year horizon. It is mostly due to the fact that the yields of domestic wholesale bonds were higher in the first years than retail yields over the past years and the average time to maturity and re-fixing of domestic wholesale debt is also higher than the retail debt. This also means that issuing shorter-term retail bonds with yields adjusting to market changes may have meant less cost than issuing a 10-year bond in 2012 and paying an 8% coupon each year. It is worth noting that this effect is mainly the result of the decreasing yields both in the Eurozone and in Hungary. However, it is also important to note that at the same time, retail cost effect is constantly increasing. It is mostly due to the fact that while the Hungarian economic policy aims to provide positive real interest rates to retail investors, the wholesale real interest rates have been in negative territory over the past few years.

The simulated retail costs are higher in 2012 than the factual costs because ÁKK changed the yields of 1-year retail bonds several times varying between 6.75% and 8%. The simulation, on the other hand, relies on the assumption that yields change only once annually and they are linked to the last 1-year T-Bill yields of the previous year, which was 7.7%.

4 Forecasting the Important Macroeconomic Variables

A macroeconomic model was created with a goal of forecasting the most important variable for retail debt: the total assets of households. A Vector Error Correction (VEC) [4] model was constructed to allow for cointegrating relationships between the variables. Quarterly macroeconomic data (nominal and real GDP, GDP deflator, CPI, inflation target, household assets) provided the input of the model. The sources

of the data are the National Bank of Hungary (NBH) and the Hungarian Central Statistical Office (HCSO). Several assumptions were made regarding the model:

1. Nominal and real GDP as well as household assets have a long-term linear relationship.
2. The GDP deflator is a linear function of the CPI and the inflation target.
3. Quarterly forecasts are ex-post transformed into monthly data using linear interpolation (to serve as inputs for the simulation model).
4. The model does not adjust for quarterly seasonality in GDP (not problematic for a 5-year forecast).
5. The estimated parameters are time-invariant so they do not change in the short-term.

4.1 Methodology

The model has two primary parts. First, a baseline VEC model for GDP, GDP deflator, and household assets. Second, the calculation of CPI from the VEC model.

In the first step, a three-dimensional VEC model was fitted to the quarterly macro data of Hungary for the period 1999Q4–2020Q4. The main variables used for VEC are the following:

- $GDP^{nom}(t)$: GDP at current prices (HUF billion)
- $GDP^{def}(t)$: GDP price deflator based on averages prices of 2015 (%)
- $F(t)$: The total assets of households (HUF billion)

Real GDP is determined by the product of nominal GDP and GDP deflator. The input variables need to be I (1) (integrated of order one) for VEC fitting. To ensure this, and for scaling reasons, the variables were transformed to log-scale with the $y = 100 * \ln(x)$ function. Then the Augmented Dickey-Fuller [1] and Phillips-Perron [5] unit root tests were performed. The results of the statistics are shown in Table 4.

Three different variants of the tests were used, testing for drift-stationarity and trend-stationarity as well. According to the simple stationarity test, the unit root null hypothesis cannot be rejected at the 5% significance level. However, if a drift or trend term is assumed, the value of p is less than 0.05 for each variable. Therefore, by

Table 4 Unit root tests

p -Values	ADF (none)	ADF (drift)	ADF (trend)	PP (none)	PP (drift)	PP (trend)
$\ln(GDP^{nom}(t))$	0.999	0.001	0.004	0.999	0.001	0.004
$\ln(GDP^{def}(t))$	0.999	0.001	0.001	0.999	0.001	0.001
$\ln(F(t))$	0.162	0.001	0.024	0.999	0.001	0.024

Source: HCSO, NBH, ÁKK

Table 5 Trace statistics

r	R	T	c	p
0	True	32.4927	29.7976	0.0077
1	False	10.0656	15.4948	0.3106
2	False	1.7391	3.8415	0.1877

Source: HCSO, NBH, ÁKK

filtering out the drift (for example by differentiation, the unit root can be eliminated. Thus, it can be assumed that the processes are I(1).

The number of cointegrating relationships was determined using Johansen method based on the values of the trace statistics [6]. The minimum number of quarterly lags was determined by the minimum values of Akaike [2] and Bayesian [3] information criteria. Based on these, three quarters was the optimal lag parameter in case of VEC. The following form with deterministic linear trends and intercepts in the cointegrated time series was used:

$$A(B'y(t) + c_0) + c_1, \quad (8)$$

where A is 3 by 1 matrix of adjustment speed, B is a 3 by 1 cointegration matrix, $y(t)$ is the level of the data, and c_0, c_1 are intercepts. The likelihood ratio tests show that the process involves a drift rather than a time trend. After that, the values of the trace statistics can be determined, as shown in Table 5.

The null hypothesis is that $H(r)$ rank of cointegration is less than or equal to r . Since the null hypothesis was first accepted for $r = 1$, a first degree cointegration relationship is identified. Thus, by fixing the lag and rank parameters, the VEC model form is the following:

$$\Delta y(t) = c + AB'y(t-1) + \sum_{j=1}^3 \Phi_j \Delta y(t-j) + \varepsilon(t), \quad (9)$$

where $\Phi_j, j = 1, 2, 3$ are 3 by 3 matrices of short-run coefficients, ε_t is a noise process, and c is the overall constant. The parameters were estimated using the maximum likelihood method.

In the second part, the consumer price index was estimated. Consumer price index also has a relevant effect on the cost of retail debt. CPI estimation is a complex task, it can be influenced by many external factors (net exports, policy changes, global market expectations, etc.) and thus it is difficult to integrate directly into the existing VEC framework. Hence, for simplicity reasons it was considered that it is better to derive CPI from the change in the GDP deflator. This was accomplished by simple linear regression.

$$\pi(t) = \beta_1 X(t) + \beta_2 \pi^{target}(t) + \varepsilon(t), \quad (10)$$

Table 6 Regression coefficients

Parameters	OLS	SE	t-stat	p-value	Adj. R^2
$X(t)$	0.586	0.115	5.092	0.001	0.40
$\pi^{target}(t)$	0.440	0.181	2.429	0.017	

Source: HCSO, NBH, ÁKK

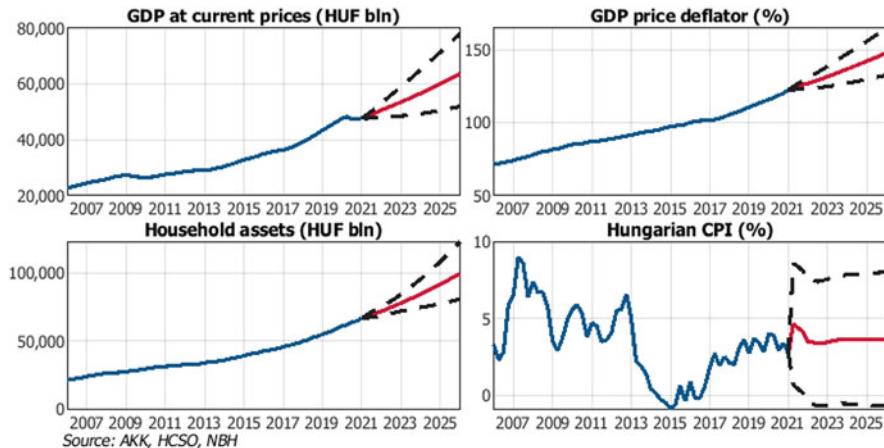


Fig. 5 Results of the VEC macroeconomic model. Source: ÁKK, HCSO, NBH

where $\pi(t)$ is the CPI, $X(t)$ is the inflation proxy calculated from the annual delta of the GDP deflator, $\pi^{target}(t)$ is the inflation target set by the central bank, and $\varepsilon(t)$ is a noise process.

The inflation target set by the National Bank of Hungary did not change since 2006. It is assumed that the 3% target will remain unchanged in the 5-year forecast horizon. The OLS estimates for π are displayed in Table 6.

The estimation was performed for the period of 2000Q4–2020Q4. Based on the adjusted R^2 , a moderately fitting model was obtained with a relatively high noise. The explanatory variables are significant and the error term follows a normal distribution.

4.2 Results

Figure 5 shows the results of the macroeconomic forecast with a 95% confidence interval. During the 5-year forecast horizon, real GDP is expected to grow by 10.4%, reaching its pre-crisis level in Q4 2022.

The level of nominal GDP is expected to increase by 33%, with the assets of households increasing by 49% from the end of 2020 to the end of 2025. In the long run, inflation is expected to slightly increase according to the results of the model.

After a spike of 4.6% in mid-2021, it is expected to stabilize at 3.7%, remaining within the inflation target range of 2–4%.

5 The Future of the Retail Debt Program

5.1 *Estimation of the Factors Driving the Outstanding Amount of Retail Debt*

Several approaches have been considered to calculate the outstanding amount of retail debt for the next 5 years. A bottom-up agent-based approach was rejected due to the lack of individual investor level data. A top-down, macro-driven approach was found problematic as the asset allocation of retail investors, especially between different retail government securities, depends on more than just macroeconomic factors. Therefore, a security-level simulation model was constructed where the most important factors are new money flowing into retail securities, the ongoing buybacks (redemptions before maturity) of debt securities and the reinvestment of maturing debt into newly issued securities by retail investors. These factors depend on historical transactions as well as macroeconomic variables.

The factors were estimated using ÁKK's transaction data. Several adjustments had to be made due to severe structural breaks in the data. First of all, before 2018, non-household investors (e.g., foundations, municipal governments, and churches) were allowed to buy retail securities. Even at the start of 2020, over HUF 1,000 billion outstanding retail securities were not owned by households. The data was corrected to exclude non-household owners to reflect that new retail debt can only be bought by households.

Sales channels for the three main securities (MÁP+, PMÁP, 1MÁP) include both banks and the Hungarian State Treasury. The rest of the dematerialized securities are only available at the Treasury, while the materialized ones can be bought and redeemed at post offices. For the three main securities, separate parameters have been estimated for the two different sales channels, as buyback mechanics are completely different for the two institutions. Buybacks at the Treasury appear immediately as a transaction, while banks can re-sell or hold papers that were bought back by them. They only appear as buybacks in the transaction data if ÁKK exercises a buyback option from the banks.

Reactions of retail investors to changes in retail instruments or the macroeconomic environment are difficult to predict due to the apparent irrational behavior of investors. As mentioned, 1MÁP is Pareto-dominated by MÁP+ from an investor standpoint and there is still a significant demand for this instrument. Also, past changes in PMÁP interest premium had no statistically significant effect on its demand. In addition, changes in CPI had no effects on the demand of the fixed rate MÁP+ versus the demand of the inflation-linked PMÁP.

Another challenge was the structural break caused by the June 2019 introduction of the flagship product MÁP+ that captured over 50% of the market share in less than 2 years. It drastically changed the landscape of retail debt due to a high and predictable yield, automatic reinvestment of coupons, and high liquidity due to favorable redemption conditions. Therefore, meaningful data could only be obtained from the post-MÁP+ period.

The introduction of MÁP+ also coincided with the abolition of interest tax on retail securities that had a negative effect on the sales of the prior months due to retail investors opting to wait for the introduction of more favorable conditions. Unfortunately, there is no way of separating the effects of these two structural breaks.

The concept of new money inflows is somewhat difficult to grasp. Theoretically, it is gross sales minus reinvestment, however reinvestment cannot be calculated easily due to the lack of individual-level data. Maturities cause no statistically significant changes in sales past 1 month, but in reality, it might take more than that for an individual to decide to reinvest. In the model, reinvestments past 1 month are treated as new inflows.

Maturities for 1MÁP occur every week. A robust linear regression of sales on the past 2 weeks of maturities proved to be the best fit, with the constant term being the new money inflows. It was found that 1MÁP series issued before the first MÁP+ were partly reinvested into MÁP+, while those issued afterwards (maturities starting June 2020) were reinvested only into 1MÁP.

PMÁP and PEMÁP maturities occur only a few times per year, while issuances happen on a daily frequency. PMÁP renewal rates were calculated using lagged maturities of the past 9 working days. Both the 3-year and 5-year maturities were significant predictors for both PMÁP instruments, so a full transition matrix had to be estimated. For PEMÁP, no 5-year maturities happened since its introduction, but 3-year maturities had an effect on 5-year issuances. Therefore, a transition matrix similar to PMÁP was estimated.

The introduction of the MÁP+ changed reinvestment behaviors as well. Therefore, MÁP+ sales had to be corrected using PMÁP maturities to calculate the new money inflows as well as transition parameters from PMÁP to MÁP+. The reinvestment of eventual MÁP+ maturities is also an important issue, as 3 trillion HUF of MÁP+ is expected to mature in 2024. Due to the lack of data, the best estimation is that the total reinvestment percentage will be equal to the instrument with the closest similarity, the 5-year PMÁP.

The landscape of the printed securities also underwent some change since the November 2020 introduction of the NYMÁP+. The traditional (and largely inferior) KTJ had high reinvestment rates that plummeted after November 2020. This difference in KTJ reinvestment, under the assumption that it was reinvested into NYMÁP+, was used to calculate the new money inflows into NYMÁP+.

Buybacks were estimated for all securities as the percentage of buybacks compared to the outstanding amount of the previous month.

According to the estimation, almost 2/3 of the new money flows into MÁP+. There is also a significant ongoing transition process, with about half of the

Table 7 ARIMAX model for new money inflows (values in HUF billion)

	Coefficient	SE	t-stat	p-value
Constant	22.56	24.93	0.91	0.37
AR(1)	0.11	0.07	1.68	0.09
AR(2)	0.06	0.08	0.76	0.45
AR(3)	0.23	0.07	3.24	0.00
Beta	0.19	0.07	2.83	0.00
Dof	2.54	0.68	3.75	0.00
Variance	14,488.76	12,237.37	1.18	0.24

Source: ÁKK

maturing PMÁP flowing into MÁP+ as well. Monthly outflows through buybacks are approximately 1% of the total debt portfolio, while outflows due to non-renewed maturities are in the 10–25% range for most instruments, occurring every 1–5 years depending on the tenor.

Historical new money inflows were calculated on a monthly frequency from 2013 using gross sales minus the estimated renewal rates. New money inflows into retail debt can be explained by a change in household assets, one of the variables in the macro model. The new money inflows were modeled and forecasted using an ARIMAX model [7]. The best fit was an ARIMA(3, 0, 0) specification with innovations drawn from the t distribution and change of household assets as an external explanatory variable.

Table 7 shows that changes in household assets are a significant predictor for new money inflows into retail debt. Based on the macroeconomic forecast, new money inflows are expected to increase by approximately 25% by 2025 compared to 2020.

5.2 Simulation and Results

The monthly new inflow, buyback, and reinvestment factor were used to simulate future transactions of retail securities. The simulation uses a weekly frequency with buybacks and issuances occurring on Mondays, the monthly factors divided evenly between the 4 or 5 weeks. Renewal of maturities based on the parameters occurs over a 3-week period in the model, with 60% reinvested in the first week, 30% in the second week, and 10% in the third week after the maturity. The simulation takes the inherent features of the different retail securities into account, including automatic reinvestment of MÁP+ and BABA interest.

The boundary conditions for the simulation include no significant changes in the prevailing interest rate regime, no newly developed retail debt instruments, and no changes in the pricing of retail securities.

Figure 6 shows the quarterly total growth rate of the retail debt portfolio as well as the contribution of different instrument categories to the total growth rate. Most of the increase in outstanding comes from MÁP+. PMÁP is expected to remain steady

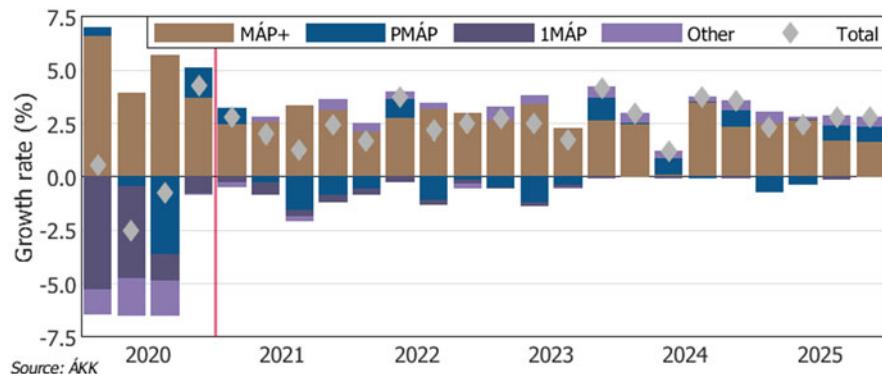


Fig. 6 Quarterly growth rate decomposition of the retail debt portfolio. Source: ÁKK

Table 8 Average interest of the retail debt portfolio

	2021	2022	2023	2024	2025
Average interest (%)	4.20	4.46	4.66	4.65	4.46

Source: ÁKK

in absolute terms, while 1MÁP is expected to dwindle further. The current strategic goal of reaching HUF 11,000 billion will be reached by mid-2022 according to the model. Throughout the 5-year simulation horizon, a yearly growth rate of 10.8% is predicted by the model, with retail debt reaching approximately 15.7% of the forecasted household assets, up from the 14.0% value of end-2020. Compared to the out-of-sample outstanding amounts of January–August 2021, the model oscillated in the [-50 ; +50] range (HUF billion), the maximum percentage deviation being about 0.5% of the factual outstanding amount.

Interest expenditures are calculated using an accrual methodology, with the future interests of the inflation-linked PMÁP calculated using the forecasted inflation, while interest expenditures related to changes in the EUR/HUF exchange rate (only relevant for PEMÁP) are calculated using ÁKK's Markov regime switching model [8].

Table 8 shows the average interest rate of the retail debt portfolio. The variance of the average interest is explained mainly by the large amount of MÁP+ issued in 2019 that is going to reach a 6% interest by 2024 and start over at 3.5% after renewal.

6 Conclusion

This paper has two main results. First, it proves that the retail program was beneficial in reducing the external vulnerability of Hungarian government debt and providing a stable, domestic source of financing.

During the 2011–2020 period, the share of retail debt in the debt portfolio increased from 2% to over 25%, whereas the share of FX debt decreased significantly from nearly 50% to under 20%. Due to the success of the retail debt program, the share of non-retail HUF-denominated debt increased only by 5 percentage points, while foreign ownership in domestic wholesale debt decreased from 41% to 24%. The interest expenditures of Hungary relative to GDP decreased from 4.1% to 2.4%. Thus, the retail debt program helped in creating a stable, reliable investor base and in diversifying funding.

Based on the what-if analysis between 2012 and 2016, the increased retail financing helped to reduce the interest costs of the debt portfolio, albeit at a lower ATM. From 2017 onwards, with negative real yields becoming prevalent and older, higher interest domestic wholesale debt maturing, the extensive retail debt became progressively more expensive compared to the alternative scenario. However, providing positive real interest rates in a negative real yield environment is necessary to further the policy goal of increasing the willingness of households to invest. In total, over the 9-year estimation period, there was no difference between the interest expenditures of the factual and the alternative scenario.

The results of the analysis also show that retail debt program helped in pushing the short end of the HUF yield curve lower by not putting too much pressure on domestic wholesale funding.

Second, if the current favorable conditions for retail investors remain intact, the outstanding amount of retail debt is going to continue to grow by 10.8% per year on average in the next 5 years under the prevailing market conditions and the macroeconomic forecast. However, with retail debt becoming increasingly more expensive, with the forecasting model presented in this paper, ÁKK has an effective tool to reassess and fine-tune the retail debt program in the future.

The methodology used has several limitations. For example, government expenditures were treated as externally given, unaffected by liquidity, in the what-if analysis. In addition, due to recent structural breaks, useful data for forecasting the retail debt is limited to 1.5 years, and assumptions had to be made regarding MÁP+ renewals. Furthermore, due to irrational investor behavior and a lack of data, predicting responses to changes in pricing was out of the scope of this paper.

There are several ways the retail debt forecast can be improved in the future. First, the creation of a more complex macroeconomic model is an avenue for improvement. Combining it with ÁKK's Markov regime switching model [8] would allow for more comprehensive forecasts. Second, an agent-based approach may be possible in the future should individual-level transaction data become available. Third, the retail forecast can be integrated into the Hungarian optimal debt portfolio model [9, 10], with the purpose of having a complex quantitative tool to update the retail debt strategy once its current goals are met.

Disclaimer

The authors of this paper are employees being responsible for the support of decision-making in the course of the development of the debt management strategy of the Hungarian Government Debt Management Agency Pte. Ltd. (in Hungarian:

Államadósság Kezelő Központ Zrt.; “ÁKK”). Therefore, this paper should be construed as how ÁKK implements its respective policies and, for the purposes of this paper, how the authors demonstrate the modeling thereof.

All data provided by ÁKK, as well as formulas, models, and methodologies created by the authors and contained in this paper are the exclusive intellectual property of ÁKK and are protected by copyright and other protective laws. Any use or other exploitation of such data, formulas, models, and methodologies in any manner is subject to express prior written authorization.

All data and formulas contained in this paper are provided solely for the purpose of illustrating the modeling framework for projecting possible outcomes under different economic scenarios. The data contained in this paper is a computer-generated output from respective mathematical models using available statistical and economic data and the output of the models should not be regarded as representative of any current data or forecasts, furthermore, must not be relied upon as an accurate prediction of current or future market performance, etc.

Results published in this paper, including but not limited to macroeconomic forecasts and the composition of the Hungarian government debt, do not represent the official views of ÁKK regarding debt financing. The only purpose of disclosing such results is to illustrate the features and possibilities of the optimal debt portfolio model.

The authors give no warranty and make no representation as to the accuracy, reliability, timeliness, or other features of any data contained in this paper or data obtained from using the model.

All models must be scientifically validated by the user for the strategy for which it is to be used, and for the most appropriate and safe application of models, scientific and expert interpretation and adequate advice are required.

References

1. Fuller, W.A.: *Introduction to Statistical Time Series*. Wiley (2009)
2. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **19**, 716–723 (1974)
3. Schwarz, G., et al.: Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978)
4. Juselius, K.: *The Cointegrated VAR Model: Methodology and Applications*. Oxford University Press (2006)
5. Phillips, P.C., Perron, P.: Testing for a unit root in time series regression. *Biometrika*. **75**, 335–346 (1988)
6. Johansen, S.: *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press on Demand (1995)
7. Box, G.E., Tiao, G.C.: Intervention analysis with applications to economic and environmental problems. *J. Am. Stat. Assoc.* **70**, 70–79 (1975)
8. Bebes, A., Tran, D., Bebesi, L.: Yield Curve Modeling with Macro Factors: An Implementation of the Kim Filter in a Two-Economy Markov Regime Switching State-Space Model. University of Granada (2018)

9. Bebes, A., Tran, D., Bebesi, L.: Optimizing the Hungarian Government Debt Portfolio. International Institute of Social and Economic Sciences (2018)
10. Tran, D., Bebes, A.: How Should Public Debt Management Institutions Develop Medium-term Issuance Strategies? PDM Network (2019)

Predicting the Exchange Rate Path: The Importance of Using Up-to-Date Observations in the Forecasts



Håvard Hungnes 

Abstract Central banks, statistical agencies, and international organizations such as the IMF and OECD typically use information about the exchange rate some weeks before the publication date as the basis for their exchange rate forecasts. This paper tests if exchange rate forecasts can be made more accurate by utilizing information about exchange rate movements closer to the publication date. To this end, we apply recent tests of equal predictability and encompassing for path forecasts. We find that the date on which the exchange rate forecast is based is crucial. Using exchange rate forecasts made by Statistics Norway over the period 2001–2018, we find that the random walk, when based on the exchange rate 1 day ahead of the publication deadline, encompasses the predicted path by Statistics Norway. However, when using the exchange rate 15 days before the publication deadline, the random walk path and the predicted exchange rate path by Statistics Norway have equal predictability.

Keywords Forecast performance · Forecast evaluation · Forecast comparison

1 Introduction

The efficient market hypothesis implies that the current exchange rate reflects all available information. However, the hypothesis does not mean that we cannot predict exchange rate changes. In the absence of risk premiums, the return in two countries—measured in a common currency—must be equal. If the countries have different interest rates, this difference in return must be compensated by an equivalent expected exchange rate change. This relationship between the interest rate difference and the expected exchange rate change is known as the uncovered interest rate parity theory.

H. Hungnes (✉)
Statistics Norway, Research Department, Oslo, Norway
e-mail: havard.hungnes@ssb.no

The theory of uncovered interest rate parity has repeatedly been rejected in empirical studies. A country's currency is often found to appreciate if the country has a higher interest rate than other countries; see, for example, [6, 21].

Random walk models are often better at projecting the exchange rates than other exchange rate models [17]. This finding is probably why several forecasters, like the Bank of Canada and the European Central Bank, assume unchanged exchange rates in their forecasts [1, 8]. The IMF assumes unchanged exchange rates in real terms in its projections [15].

However, even if these forecasters use unchanged (nominal or real) exchange rates ahead, the exchange rate used as the basis for the prediction is not the most recent observed exchange rate. For example, in its forecast from January 2020, the IMF used the average real exchange rates in a period broadly covering the end of October and the beginning of November 2019 as their forecast for the real exchange rate [15]. The European Central Bank used, in their forecast published March 12, 2020, the exchange rates equal to the average in the first half of February [8].

Each quarter, following the publication of new quarterly national account data, Statistics Norway publishes forecasts for the Norwegian economy 3–4 years ahead in annual terms. Among the variables Statistics Norway publishes forecasts for is the krone exchange rate measured against a basket of currencies of Norway's most important trading partners in terms of import value (also known as the Norwegian import-weighted krone). Until 2018, the exchange rate forecast was partly based on judgmental forecasts and partly on an econometric model for the exchange rate. An early version of the exchange rate model used by Statistics Norway applies data from the years 1983–2002 to estimate the model [2]. Since the beginning of 2019, Statistics Norway has forecasted an unchanged exchange rate.

Norway has experienced large exchange rate fluctuations in the period we are considering, which is the period after Norway started its inflation targeting at the beginning of 2001 and until 2020. In this period, the cost of one euro has been as low as 7.22 Norwegian kroner and as high as 12.32 (when considering the official daily rates published by Norges Bank). The cost of one dollar varied between 4.96 and 11.40 Norwegian kroner in the same period. In the analysis, we consider the Norwegian import-weighted krone, which also has fluctuated much in this period. With these large fluctuations, we might get more precise results from our tests of equal predictability and encompassing for path forecasts than one could get from similar studies for other countries with smaller exchange rate variations.

This paper uses the equal predictability test of path forecasts to compare the exchange rate forecasts by Statistics Norway with the exchange rate path that follows from a random walk [12]. We also apply the encompassing test for path forecasts to test if the forecasts of the random walk model based on the exchange rate at two different time points encompass the forecasts by Statistics Norway [11].

Section 2 presents the test of equal predictability for path forecasts and the encompassing test for path forecasts. Section 3 defines the import-weighted krone exchange rate. This section also applies the equal predictability test and the encompassing test to evaluate the exchange rate forecasts by Statistics Norway. Section 4 provides a conclusion.

2 Theory

Let $s_{t+h|t}^i = \log S_{t+h|t}^i$ be the forecast of the log of the exchange rate for period $t + h$, made in period t by forecaster (or forecasting method) i . We assume that the exchange rate for period t is not known in period t ; thus, exchange rate forecasts for the current period—also referred to as now-casting—can be made and is denoted $s_{t|t}^i$. The forecast error for the exchange rate in period $t + h$ for the forecast made in period t by forecaster i is given by

$$e_{t+h|t}^i \equiv s_{t+h} - s_{t+h|t}^i, \quad (1)$$

where s_{t+h} is the log of the actual exchange rate in period $t + h$. As we apply the logarithmic scale, the forecast error is approximately a measure of the percentage error in the forecasts.

The mean squared forecast error (MSFE) of T forecasts with forecast horizon h for forecaster i is

$$T^{-1} \sum_{t=1}^T (e_{t+h|t}^i)^2. \quad (2)$$

Unfortunately, the MSFE is not invariant to linear transformations of the forecasts when $h > 0$ [3, 4]. For example, the MSFE will differ depending on whether the forecasts are measured in levels or first-differences. However, we can avoid this problem by considering the full path of forecasts. We define the vector of exchange rate forecasts by forecaster i for the current and the next 2 years, made in year t by $\mathbf{s}_{t,H|t}^i = (s_{t|t}^i, s_{t+1|t}^i, s_{t+2|t}^i)'$, where the forecast horizon is given by $H = 2$ measured in years. The actual exchange rate path in these years is given by $\mathbf{s}_{t,2} = (s_t, s_{t+1}, s_{t+2})'$, which implies that the vector of forecast errors of forecaster i is given by $\mathbf{e}_{t,2|t}^i = (e_{t|t}^i, e_{t+1|t}^i, e_{t+2|t}^i)'$, where the elements are defined in (1).

A general test of equal predictability for two univariate forecasts is previously suggested [23]. Adjusted to a vector of forecasts, the regression that forms the basis of the test is given by

$$\mathbf{s}_{t,2|t} = (1 - \alpha)\mathbf{s}_{t,2|t}^A + \alpha\mathbf{s}_{t,2|t}^B + \varepsilon_t, \quad (3)$$

where the expectation of the vector ε_t is zero if the forecast of forecasters A and B is unbiased. In (3), the weights of forecaster A and forecaster B sum to unity, where α is the weight on the forecast of forecaster B . If $\alpha = \frac{1}{2}$, the two forecasts have equal weights, and this is the basis of our equal predictability test. If $\alpha = 0$, the forecast by forecaster A is the best, and the additional information in the forecast by B cannot be used to improve the forecast. This is the basis of our test of whether the forecast by forecaster A encompasses the forecast by forecaster B . Similarly, if $\alpha = 1$, forecaster B provides the best forecast and the forecast by forecaster A cannot be used to improve the forecast.

By using the definition of forecast errors, (3) can be reformulated as

$$\mathbf{e}_{t,2|t}^A = \alpha \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right) + \varepsilon_t. \quad (4)$$

The conditional estimators of α and the variance of ε are given by

$$\hat{\alpha}_{(\Omega)} = \frac{T^{-1} \sum_{t=1}^T \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right)' \Omega^{-1} \mathbf{e}_{t,2|t}^A}{T^{-1} \sum_{t=1}^T \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right)' \Omega^{-1} \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right)}, \text{ and} \quad (5)$$

$$\hat{\Omega}_{(\alpha)} = \frac{1}{T} \sum_{t=1}^T \left(\mathbf{e}_{t,2|t}^A - \alpha \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right) \right) \left(\mathbf{e}_{t,2|t}^A - \alpha \left(\mathbf{e}_{t,2|t}^A - \mathbf{e}_{t,2|t}^B \right) \right)'. \quad (6)$$

The conditional estimators above do not account for the likely autocorrelation structure in the residual in (4) due to the overlapping forecast horizons. The quasi-maximum likelihood estimates $\hat{\alpha}_{(\hat{\Omega})}$ and $\hat{\Omega}_{(\hat{\alpha})}$ can be obtained by an iterative procedure until convergence is achieved [19].

The loss-difference function between the forecast errors of forecaster A and forecaster B is

$$d_t = \mathbf{e}_{t,H|t}^{A'} \mathbf{H} \mathbf{e}_{t,H|t}^A - \mathbf{e}_{t,H|t}^{B'} \mathbf{H} \mathbf{e}_{t,H|t}^B, \quad (7)$$

where \mathbf{H} —which is an $(H+1) \times (H+1)$ matrix—represents the parameters in the loss function [20]. Testing the null hypothesis of $\alpha = \frac{1}{2}$ —i.e., when forecast A and forecast B have equal predictability—is identical to testing the population equivalent of \bar{d} being zero, where $\bar{d} = T^{-1} \sum_{t=1}^T d_t$, with d_t defined in (7) where $\mathbf{H} = \hat{\Omega}_{(\hat{\alpha})}^{-1}$ [12].

The test statistic we apply is

$$T^{1/2} w_0^{1/2} \bar{d} \hat{q}^{-1/2}, \quad (8)$$

where w_0 is a small sample correction factor. The estimated variance of the estimator is¹

$$\hat{q} = \frac{1}{T} \left[\sum_{t=1}^T \tilde{d}_t^2 + 2 \sum_{l=1}^{\tau_H} \sum_{t=1}^{T-l} w_l \tilde{d}_t \tilde{d}_{t+l} \right], \text{ with } \tilde{d}_t = \left(\mathbf{e}_{t,H|t}^{A'} - \mathbf{e}_{t,H|t}^{B'} \right) \hat{\Omega}_{(\hat{\alpha})}^{-1} \hat{\varepsilon}_t, \quad (9)$$

¹ We can also include a small sample correction for the heteroskedasticity. In this application, we apply the HC3 correction [16], which we operationalize by defining $\tilde{d}_t = \frac{1}{1-h_t} \left(\mathbf{e}_{t,H|t}^{A'} - \mathbf{e}_{t,H|t}^{B'} \right) \hat{\Omega}_{(\hat{\alpha})}^{-1} \hat{\varepsilon}_t$, where $h_t = D_t' \left(\sum_{j=1}^T D_j D_j' \right)^{-1} D_t$ with $D_t = \mathbf{e}_{t,H|t}^A - \mathbf{e}_{t,H|t}^B$.

where τ_H is the truncation lag, $\tau_H \geq H$. To secure this variance to be positive, we may use $w_i = 1 - \frac{i}{\tau_H+1}$ for $i = 1, \dots, \tau_H$ [18]. The small sample correction $w_0 = T^{-1}[T - 1 - 2\tau_H + T^{-1}\tau_H(\tau_H + 1)]$ is suggested and applied here [10]. The test statistic is asymptotically standard normally distributed when the forecasting models are estimated using a rolling sample [9]. However, the deviation between the actual distribution of the test statistic and the normal distribution can be substantial in small samples. Therefore, we apply a t -distribution with $TK - K$ degrees of freedom [11].

In investigating the exchange rate path forecasts, we will also test if one path forecast encompasses another path forecast. We can conclude that forecast A encompasses forecast B if we cannot reject the hypothesis $\alpha = 0$ but can reject the hypothesis of $\alpha = 1$ [7]. Also, for this type of test, the test statistic in (8) can be applied [11]. For the test of the hypothesis $\alpha = 0$, the definition of d_t is changed to $d_t = (\mathbf{e}_{t,H|t}^{A'} - \mathbf{e}_{t,H|t}^{B'}) \hat{\Omega}_{(\hat{\alpha})}^{-1} \mathbf{e}_{t,H|t}^A$. With this definition of d_t , \bar{d} is identical to the numerator of the estimator of α . For the test of the hypothesis $\alpha = 1$, we use $d_t = (\mathbf{e}_{t,H|t}^{A'} - \mathbf{e}_{t,H|t}^{B'}) \hat{\Omega}_{(\hat{\alpha})}^{-1} \mathbf{e}_{t,H|t}^B$. With this definition of d_t , $\bar{d} = 0$ if the $\hat{\alpha}_{(\Omega)} = 1$. The test statistic in (8) with \hat{q} given by (9) is also used for these tests for encompassing.

3 Results

Statistics Norway publishes macro-economic forecasts of the Norwegian economy each quarter. Since the beginning of 2001, these forecasts have included year-to-year forecasts for the import-weighted krone (I-44) for the same year as the forecast is made and (at least) the two subsequent years.² In this section, we consider the exchange rate forecasts made in the years 2001–2018. With a 2-year horizon of the forecasts, the forecasts made in 2018 include forecasts of the exchange rate in 2020. The data set and the ox code used in this paper are available to download [5, 14].

The I-44 exchange rate index is a geometric weighted average of the exchange rates of 44 countries. The weights are updated annually and based on Statistics Norway's statistics for imports to Norway from the 44 largest countries measured in import value. Norges Bank updates the composition of countries annually.

² There are two exceptions: First, in the publication of the forecast made the first quarter in 2001, Statistics Norway only published a forecast for I-44 for the years 2001 and 2002. In our analysis, we assume that the forecasted value of I-44 for 2003 in the forecast published in the first quarter of 2001 is equal to the forecasted value for 2002, i.e., no change in the I-44 from 2002 to 2003 on a year-to-year basis. Second, Statistics Norway did not publish forecasts in the third quarter of 2013. In this analysis, we have set this forecast equal to the forecast made in the second quarter of that year. For the forecast based on the random walk, we have also used the exchange rate equal to the exchange rate in the market relative to the time the second quarter forecast from Statistics Norway was made.

Table 1 reports the weights used from September 4, 2018. These weights are based on the value of imports to Norway in 2017. The number of currencies in the table is less than 44 since some currencies are used in more than one country (e.g., the euro).

Table 1 Weights used for import-weighted krone exchange rate, I-44, based on import to Norway in the year 2017

Country, currency	Short name	Weight
Bangladesh, Taka	BDT	0.003
Brazil, Real	BRL	0.015
Canada, Dollar	CAD	0.020
Switzerland, Franc	CHF	0.012
China, Yuan Renminbi	CNY	0.101
Colombia, Peso	COP	0.002
Czech Republic, Koruna	CZK	0.011
Denmark, Krone	DKK	0.056
European Union, Euro	EUR	0.325
United Kingdom, Pound	GBP	0.049
Hungary, Forint	HUF	0.004
Indonesia, Rupiah	IDR	0.002
India, Rupee	INR	0.006
Iceland, Krone	ISK	0.004
Japan, Yen	JPY	0.021
South Korea, Won	KRW	0.070
Malaysia, Ringgit	MYR	0.005
Peru, New sol	PEN	0.002
Poland, Zloty	PLN	0.035
Romania, New leu	RON	0.004
Russia, Ruble	RUB	0.019
Sweden, Krone	SEK	0.118
Singapore, Dollar	SGD	0.004
Thailand, Baht	THB	0.009
Turkey, Lira	TRY	0.011
Taiwan, New Dollar	TWD	0.006
United States, Dollar	USD	0.070
Vietnam, Dong	VND	0.007
Coopération Financière en Afrique Centrale, CFA-franc	XAF	0.002
South Africa, Rand	ZAR	0.004
Mexico, Peso	MXN	0.003
Sum		1

Source: Norges Bank



Fig. 1 The Norwegian import-weighted krone exchange rate index, I-44, 2001–2020, quarterly averages (index equal to 100 in 1995)

The calculation of the I-44 index is based on Laspeyres' index formula:

$$S_t = S_{t-1} \prod_{j=1}^N \left(\frac{S_t^{(j)}}{S_{t-1}^{(j)}} \right)^{\alpha_{t-1}^{(j)}}, \quad (10)$$

where S_t is the I-44 index at time t ; $S_t^{(j)}$ is the exchange rate j at time t (where we use the round brackets for the top index to not confuse it with the index of the forecaster); and $\alpha_{t-1}^{(j)}$ is the weight of the exchange rate j from time $t - 1$, where $\sum_{j=1}^N \alpha_t^{(j)} = 1$, with N being the number of currencies in the 44 countries. Figure 1 shows the I-44 index in the years 2001–2020.

Norges Bank publishes the official I-44 index. The published annual figures of I-44 are the arithmetic average of the trading day observations of the index. Hence—if t runs over the major time period, here years, and ς runs over the minor time period, here the trading days within year t —the annual figure of the I-44 period index for year t is

$$S_t = \frac{1}{\varsigma_{max}(t)} \sum_{\varsigma=1}^{\varsigma_{max}(t)} S_{t,\varsigma}, \quad (11)$$

where $\varsigma_{max}(t)$ is the number of trading days in year t .

The random walk forecast of the I-44 index in year t made when ς' is the latest observed trading day in year t is

$$S_{t|t(\varsigma')}^{RW} = \frac{\varsigma'}{\varsigma_{max}(t)} \left(\frac{1}{\varsigma'} \sum_{\varsigma=1}^{\varsigma'} S_{t,\varsigma} \right) + \frac{\varsigma_{max}(t) - \varsigma'}{\varsigma_{max}(t)} S_{t,\varsigma'}, \quad (12)$$

where the term in the round brackets is the average of the index from trading day 1 to trading day ς' in year t and where $S_{t,\varsigma'}$ in the last term reflects that the latest observed value of the index is the best forecast under the random walk hypothesis of the exchange rate for all the remaining trading days of year t . The random walk forecast of the annual value of the exchange rate index in the coming years is equal to the last observation of the index, that is,

$$S_{t+h|t(\varsigma')}^{RW} = S_{t,\varsigma'} \text{ for } h = 1, 2, \dots \quad (13)$$

Tables 2 and 3 compare the forecasts by Statistics Norway with forecasts generated by a random walk. Let $s_{t+2|t(q)}^F = (s_{t|t(q)}^F, s_{t+1|t(q)}^F, s_{t+2|t(q)}^F)'$ be the vector of the exchange rate forecasts by Statistics Norway up to horizon $H = 2$, where the forecasts are made in quarter q of year t . Furthermore, let $s_{t+2|t(\varsigma')}^{RW} = (s_{t|t(\varsigma')}^{RW}, s_{t+1|t(\varsigma')}^{RW}, s_{t+2|t(\varsigma')}^{RW})'$ be the implied forecasts by a random walk where the last observation is $t(\varsigma')$, with elements given by (12) and (13). To test the

Table 2 Equal predictability and encompassing tests for path forecasts ($H = 2$)—random walk based on the exchange rate 1 day before the publication deadline

Projection quarter	Weight F	Weight RW_1	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	-0.28	1.28	1.21 [0.23]	1.98 [0.05]	0.43 [0.67]
Q2	-0.47	1.47	1.39 [0.17]	2.11 [0.04]*	0.67 [0.50]
Q3	-0.27	1.27	1.31 [0.19]	2.17 [0.04]*	0.46 [0.64]
Q4	0.05	0.95	2.51 [0.02]*	5.33 [0.00]**	0.30 [0.77]
All	-0.11	1.11	2.26 [0.02]*	4.10 [0.00]**	0.42 [0.68]

F indicates the forecasts made by Statistics Norway. RW_1 (RW_{15}) indicates the random walk forecasts, which are set equal to the official exchange rate 1 day (15 days) before the publishing deadline for the forecasts by Statistics Norway. “Q1” denotes the forecasts made in the first quarter of the year, in the years 2001–2018. Similarly, for “Q2,” “Q3,” and “Q4.” “All” implies that we consider the forecasts from all quarters in the analysis. The p-values in square brackets are based on a two-sided test; the corresponding p-value for a one-sided test is either $p_{1-sided} = p_{2-sided}/2$ or $p_{1-sided} = 1 - p_{2-sided}/2$, depending on the direction of the one-tailed hypothesis. ** and * indicate significance at the 1% and the 5% level for the two-sided test. In the estimation, we use $\tau_H = 2$ for the forecasts made in Q1, Q2, Q3, and Q4, and $\tau_H = 11$ for “All”

Table 3 Equal predictability and encompassing tests for path forecasts ($H = 2$)—random walk based on the exchange rate 15 days before the publication deadline

Projection quarter	Weight F	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	0.38	0.62	0.25 [0.80]	1.29 [0.20]	0.79 [0.43]
Q2	0.22	0.78	0.47 [0.64]	1.32 [0.19]	0.38 [0.70]
Q3	0.09	0.91	0.68 [0.50]	1.52 [0.14]	0.16 [0.88]
Q4	0.59	0.41	0.73 [0.47]	3.14 [0.01]**	4.61 [0.00]**
All	0.38	0.62	0.50 [0.62]	2.61 [0.01]**	1.61 [0.11]

Note: See Table 2

encompassing and equal predictability hypothesis, we apply (3) with $s_{t+H|t}^A = s_{t+H|t(q)}^F$ and $s_{t+H|t}^B = s_{t+H|t(\zeta)}^{RW}$.

The forecasts by Statistics Norway are published quarterly, usually at the beginning of the third month in the quarter. The publication day is (with a few exceptions) Thursday, with a deadline of preparing the forecasts on Tuesday the same week they are published. The work with the predictions starts (usually) Monday two and a half weeks before the publication. Thus, the forecasts are typically made in about 12 working days. Until the end of 2018, the path for the exchange rate index was usually decided at the beginning of this period, though it could be revised during the process of making the forecasts.

Table 2 reports the results of the equal predictability test and the encompassing test when the random walk forecasts are based on the exchange rate only 1 day before the deadline. This observation of the exchange rate is the most updated official exchange rate it can use in the forecasts as Norges Bank publishes the official quote of the exchange rate approximately at 16:00 CET. In the table, we consider forecasts made in each of the four quarters of the year separately. Thus, in the row marked “Q1,” forecasts made each year in the first quarter from 2001 to 2018 are considered. For this first quarter of the year forecasts, we see that the estimated weight of the forecasts by Statistics Norway is negative. Due to the small number of observations, we cannot reject that the weights are 0.5. However, the hypothesis that the forecasts by Statistics Norway have a weight of at least 1 is rejected at the 5 percent level when applying a one-sided test. (The p-value barely exceeds 5 percent with the two-sided test.) The last test in the row shows that we cannot reject that the random walk forecasts have a weight of 1. The previous two tests imply that the random walk forecast encompasses the forecasts made by Statistics Norway.

For the forecasts made in the second, third, and fourth quarter, we see similar results as for the forecasts made in the first quarter of the year: The hypothesis of equal weights cannot be rejected in two of these three quarters; the hypothesis that forecasts made by Statistics Norway has a weight of 1 is rejected for all of the quarters, and the opposite hypothesis that the random walk has a weight of 1 cannot be rejected in any of the quarters. Therefore, also for forecasts made in these quarters, we find that a random walk-based forecast encompasses the forecast by Statistics Norway.

In the last row, marked “All,” we have stacked all forecasts made by Statistics Norway in these years (2001–2018) after each other (in the order they were made). We have also taken into account that this will lead to an autocorrelation of a higher order, as more forecasts overlap in time. The estimated weight on the forecasts by Statistics Norway is close to zero, indicating that the projected exchange rate path has no value over a path given by a random walk model. The test statistics for $\alpha = 0$ and $\alpha = 1$ confirm this finding. We also reject the hypothesis of equal weights for the two forecasts at the 5 percent level. The overall conclusion is that the exchange rate forecasts by Statistics Norway add no extra information to the future values of the exchange rate beyond what the random walk-based forecasts give.

In Table 3, we examine how important it is for the results that the exchange rate used for the random walk forecasts are as up-to-date as possible. We do this by letting the random walk forecasts be based on the exchange rate 15 days before the publication deadline for the forecasts by Statistics Norway, the exchange rate from the day Statistics Norway usually started its work with the forecasts. Here also, we consider the forecasts made in the four different quarters of the year separately, in addition to considering the forecasts from all quarters jointly. The estimated weights for the forecasts made by Statistics Norway are now positive, no matter what quarter the forecasts were made, and we cannot reject the hypothesis that the weights are 0.5 in any of the considered cases. When considering forecasts from all quarters jointly, see the last row of the table, the hypothesis that the forecasts made by Statistics Norway have a weight of 1 is rejected, whereas the hypothesis that the random walk-based forecasts have a weight of 1 is not rejected. Thus, we cannot reject that the forecasts made by Statistics Norway and the random walk-based forecasts have equal predictability. At the same time, we cannot reject that the random walk-based forecasts encompass the forecasts by Statistics Norway. However, the hypothesis that the forecasts by Statistics Norway encompass the random walk-based forecasts is clearly rejected.

The change in the estimates of the weights from Tables 2 to 3 shows the importance of the additional 2 weeks of exchange rate data for the forecasts. In Table 4, we compare the two random walk forecasts directly. Considering the random walk forecasts for each quarter separately, the weight on the most updated

Table 4 Equal predictability and encompassing tests for path forecasts ($H = 2$)—random walk based on the exchange rate 1 day vs. 15 days before the publication deadline

Projection quarter	Weight RW_1	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	2.51	-1.51	2.54 [0.01]*	1.91 [0.06]	3.16 [0.00]**
Q2	1.36	-0.36	3.68 [0.00]**	1.55 [0.13]	5.81 [0.00]**
Q3	0.89	0.11	2.01 [0.05]*	0.59 [0.56]	4.62 [0.00]**
Q4	1.09	-0.09	2.66 [0.01]*	0.42 [0.67]	4.90 [0.00]**
All	1.68	-0.68	4.73 [0.00]**	2.73 [0.01]**	6.73 [0.00]**

Note: See Table 2

random walk forecast varies from 0.89 to 2.51. We cannot reject that this weight is 1 (i.e., $\alpha = 0$) for any of the four quarters. For all of the four quarters, we can also reject the hypothesis that the weight of random walk forecast based on the exchange rates 15 days ahead of the deadline for Statistic Norway's forecasts is equal to 1 (i.e., $\alpha = 1$). Thus, when considering forecasts made in each quarter separately, we may conclude that the forecasts based on the exchange rate 1 day prior to the publication deadline encompass the forecast based on the exchange rate 15 days prior to the publication deadline.

Table 4 also reports the estimated weights when considering the random walk-based forecasts at the time of all the publication dates of forecasts by Statistics Norway in the years 2001–2018. When considering forecasts made in all four quarters jointly, the estimated weight for the most recent based forecast is 1.68 and exceeds 1 significantly. Based on the estimation results for the individual quarter the forecasts are made, we see that it is for the forecasts made in the first quarter that the estimated weight deviates mostly from 1.

Tables 2, 3, 4 compare different exchange rate forecasts up to horizon 2, i.e., up to 2 years ahead. When comparing the forecast paths, the test statistics applied here weight the nowcast with the forecasts 1 and 2 years ahead (i.e., $H = 2$). The weights are given by the inverse of the estimated covariance matrix in (6). Typically, the variance of the forecast error for the exchange rate in the same year (i.e., the nowcast) will be smaller than the variance of the forecast errors 1 and 2 years ahead. This implies that the test statistics typically will put a higher weight on the nowcasts. Now, we consider this further by repeating the tests in Tables 2, 3, 4 with smaller forecasting horizons. Tables 5, 6, 7 report the results with $H = 1$, i.e., when considering path forecasts consisting of a nowcast and a 1-year-ahead forecast of the exchange rate. Tables 8, 9, 10 report the results with $H = 0$, i.e., when only considering nowcasts of the exchange rate. The tests with $H = 0$ are all univariate tests.

The results in Tables 5, 6, 7, 8, 9, 10 more or less confirm the results with the longer forecast horizon in Tables 2, 3, 4. When considering forecasts made in all quarters jointly, Tables 5 and 8 confirm that the random walk forecasts based on the exchange rate 1 day before the publication deadline encompass the forecasts given

Table 5 Equal predictability and encompassing tests for path forecasts ($H = 1$) — random walk based on the exchange rate 1 day before the publication deadline

Projection quarter	Weight F	Weight RW_1	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	-0.18	1.18	1.35 [0.18]	2.35 [0.02]*	0.36 [0.72]
Q2	-0.48	1.48	1.32 [0.19]	1.99 [0.05]	0.64 [0.52]
Q3	-0.40	1.40	1.88 [0.07]	2.93 [0.01]**	0.84 [0.41]
Q4	0.03	0.97	2.85 [0.01]*	5.87 [0.00]**	0.17 [0.86]
All	-0.12	1.12	1.86 [0.06]	3.35 [0.00]**	0.37 [0.71]

Note: See Table 2

Table 6 Equal predictability and encompassing tests for path forecasts ($H = 1$) — random walk based on the exchange rate 15 days before the publication deadline

Projection quarter	Weight F	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	0.43	0.57	0.16 [0.88]	1.32 [0.19]	1.00 [0.32]
Q2	0.32	0.68	0.33 [0.74]	1.24 [0.22]	0.57 [0.57]
Q3	0.10	0.90	0.73 [0.47]	1.63 [0.11]	0.18 [0.86]
Q4	0.46	0.54	0.22 [0.82]	2.94 [0.00]**	2.49 [0.02]*
All	0.46	0.54	0.18 [0.86]	2.23 [0.03]*	1.87 [0.06]

Note: See Table 2

Table 7 Equal predictability and encompassing tests for path forecasts ($H = 1$) — random walk based on the exchange rate 1 day vs. 15 days before the publication deadline

Projection quarter	Weight RW_1	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	2.47	-1.47	1.79 [0.08]	1.33 [0.19]	2.24 [0.03]*
Q2	1.40	-0.40	2.38 [0.02]*	1.6 [0.29]	3.69 [0.00]**
Q3	0.86	0.14	1.24 [0.22]	0.48 [0.64]	2.96 [0.00]**
Q4	1.09	-0.09	2.07 [0.04]*	0.33 [0.74]	3.81 [0.00]**
All	1.70	-0.70	4.35 [0.00]**	2.55 [0.01]*	6.16 [0.00]**

Note: See Table 2

Table 8 Equal predictability and encompassing tests for path forecasts ($H = 0$) — random walk based on the exchange rate 1 day before the publication deadline

Projection quarter	Weight F	Weight RW_1	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	-0.09	1.09	1.20 [0.24]	2.21 [0.03]*	0.18 [0.86]
Q2	-0.62	1.62	2.19 [0.03]*	3.17 [0.00]**	1.21 [0.23]
Q3	-0.55	1.55	2.53 [0.01]*	3.73 [0.00]**	1.33 [0.19]
Q4	0.04	0.96	3.06 [0.00]**	6.40 [0.00]**	0.28 [0.78]
All	-0.26	1.26	1.91 [0.06]	3.16 [0.00]**	0.66 [0.51]

Note: See Table 2

by Statistics Norway. However, when considering the random walk forecasts based on the exchange rate 15 days before the publication deadline, Tables 6 and 9 confirm that we cannot reject that this forecast path has equal predictability to the forecast path given by Statistics Norway. Finally, Tables 7 and 10 confirm that the random walk forecast based on the exchange rate 1 day before the publication deadline is superior to the random walk forecasts based on the exchange rate 2 weeks earlier.

Table 9 Equal predictability and encompassing tests for path forecasts ($H = 0$)—random walk based on the exchange rate 15 days before the publication deadline

Projection quarter	Weight F	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	0.69	0.31	0.64 [0.52]	1.06 [0.29]	2.35 [0.02]*
Q2	0.47	0.53	0.04 [0.97]	0.63 [0.53]	0.55 [0.58]
Q3	0.11	0.89	0.60 [0.55]	1.38 [0.17]	0.17 [0.87]
Q4	0.49	0.51	0.04 [0.97]	2.51 [0.02]*	2.44 [0.02]*
All	0.61	0.39	0.43 [0.67]	1.43 [0.15]	2.29 [0.02]*

Note: See Table 2

Table 10 Equal predictability and encompassing tests for path forecasts ($H = 0$)—random walk based on the exchange rate 1 day vs. 15 days before the publication deadline

Projection quarter	Weight RW_1	Weight RW_{15}	t-test (two-sided)		
	$1 - \alpha$	α	$H_0 : \alpha = \frac{1}{2}$	$H_0 : \alpha = 0$	$H_0 : \alpha = 1$
Q1	2.57	-1.57	3.38 [0.00]**	2.56 [0.01]*	4.19 [0.00]**
Q2	1.35	-0.35	2.36 [0.02]*	0.97 [0.34]	3.76 [0.00]**
Q3	0.87	0.13	1.03 [0.31]	0.37 [0.71]	2.43 [0.02]*
Q4	1.08	-0.08	2.18 [0.03]*	0.31 [0.76]	4.05 [0.00]**
All	1.80	-0.80	5.73 [0.00]**	3.53 [0.00]**	7.94 [0.00]**

Note: See Table 2

4 Conclusions

We have used new tests for equal predictability and encompassing for path forecasts to compare the predicted exchange rate path by Statistics Norway with a random walk forecast [11, 12]. The date the random walk forecast is based on is shown to be crucial. When the random walk is generated from the exchange rate at the deadline of the publication of the forecasts made by Statistics Norway, the random walk forecast path encompasses the forecasted path by Statistics Norway. However, when the random walk forecast path is generated based on the exchange rate 15 days before the publication deadline, the random walk path and the forecasted exchange rate path by Statistics Norway have equal predictability.

We can draw two lessons. First, there is no indication that the exchange rate forecasts by Statistics Norway were better than a random walk forecast. The exchange rate path forecast from Statistics Norway has equal predictability as the random walk forecast based on the exchange rates 15 days before the publication deadline. Therefore, starting from the forecasts made in the first quarter of 2019, Statistics Norway has forecasted an unchanged exchange rate [22]. Second, using the exchange rate as close to the projection deadline as possible improves the forecasts. Therefore, Statistics Norway updates its exchange rate forecasts until the projection deadline.

Acknowledgments Thanks to Pål Boug, Thomas von Brasch, Terje Skjerpen, participants at the 22nd Dynamic Econometric Conference in 2019 and the 42nd Annual Meeting of the Norwegian Association of Economists in 2020, the referees for and participants at the 7th International Conference on Time Series and Forecasting in 2021 for valuable comments, and the referees for this book chapter. Also, thanks to Trym Kristian Økland for collecting and organizing the data. An earlier version of this paper is available as a working paper from Statistics Norway [13].

References

1. Bank of Canada: Monetary Policy Report, January 2020. <http://www.bankofcanada.ca/wp-content/uploads/2010/02/update120707.pdf>
2. Bjørnland, H.C., Hungnes, H.: The importance of interest rates for forecasting the exchange rate. *J. Forecast.* **25**(3), 209–221 (2006). <https://doi.org/10.1002/for.983>
3. Clements, M., Hendry, D.F.: *Forecasting Economic Time Series*. Cambridge University Press (Oct 1998). <https://doi.org/10.1017/CBO9780511599286>
4. Clements, M.P., Hendry, D.F.: On the limitations of comparing mean square forecast errors. *J. Forecast.* **12**(8), 617–637 (1993). doi: <https://doi.org/10.1002/for.3980120802>
5. Doornik, J.A.: An Object-oriented Matrix Programming Language Ox7. Timberlake Consultants Press, London (2013). <http://www.timberlake.co.uk/shop/ox-7-an-object-orientated-matrix-programming-language.html>
6. Engel, C., Lee, D., Liu, C., Liu, C., Wu, S.P.Y.: The uncovered interest parity puzzle, exchange rate forecasting, and Taylor rules. *J. Int. Money Finance* **95**, 317–331 (2018). <https://doi.org/10.1016/j.jimfin.2018.03.008>
7. Ericsson, N.R.: On the limitations of comparing mean square forecast errors: Clarifications and extensions. *J. Forecast.* **12**(8), 644–651 (Dec 1993). <https://doi.org/10.1002/for.3980120806>
8. European Central Bank: ECB staff macroeconomic projections for the euro area, March 2020, pp. 1–5 (2020). <http://www.ecb.int/pub/pdf/other/ecbstaffprojections201209en.pdf>
9. Giacomini, R., White, H.: Tests of conditional predictive ability. *Econometrica* **74**(6), 1545–1578 (2006). <https://doi.org/10.1111/j.1468-0262.2006.00718.x>
10. Harvey, D.I., Leybourne, S.J., Newbold, P.: Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **13**(2), 281–291 (1997). [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
11. Hungnes, H.: Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations. Discussion Papers 871. Statistics Norway (2018). <http://hdl.handle.net/11250/2560772>
12. Hungnes, H.: Equal predictability test for multi-step-ahead system forecasts invariant to linear transformations. Discussion Papers 931. Statistics Norway (2020). <http://hdl.handle.net/11250/2656482>
13. Hungnes, H.: Predicting the exchange rate path - The importance of using up-to-date observations in the forecasts.pdf:pdf. Discussion Papers 934. Statistics Norway (2020). <http://hdl.handle.net/11250/2663959>
14. Hungnes, H.: Forecasting the Norwegian import-weighted krone exchange rate. Mendeley Data (2021). <https://doi.org/10.17632/7schvgp54p.3>
15. IMF: World Economic Outlook - January 2020. World Economic Outlook Update (January 2020). <http://www.imf.org/en/Publications/WEO/Issues/2020/01/20/weo-update-january2020>
16. MacKinnon, J.G., White, H.: Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econ.* **29**(3), 305–325 (1985). [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
17. Meese, R.A., Rogoff, K.: Empirical exchange rate models of the seventies. *J. Int. Econ.* **14**(1-2), 3–24 (1993). [https://doi.org/10.1016/0022-1996\(83\)90017-X](https://doi.org/10.1016/0022-1996(83)90017-X)

18. Newey, W.K., West, K.D.: A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**(3), 703–708 (1987). <https://doi.org/10.2307/1913610>
19. Oberhofer, W., Kmenta, J.: A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica* **42**(3), 579–590 (1987). <https://doi.org/10.2307/1911792>
20. Pesaran, M.H., Skouras, S.: Decision-based methods for forecast evaluation. In: Clements, M.P., Hendry, D.F. (eds.) *A Companion to Economic Forecasting*, chap. 11, pp. 241–267. Blackwell Publishing (2002). <http://www.wiley.com/en-us/A+Companion+to+Economic+Forecasting-p-9781405171915>
21. Rossi, B.: Exchange rate predictability. *J. Econ. Lit.* **51**(4), 1063–1119 (2013). <https://doi.org/10.1257/jel.51.4.1063>
22. Statistics Norway: Economic Survey, 1/2019 (2019). <http://www.ssb.no/en/nasjonalregnskap-og-konjunkturer/artikler-og-publikasjoner/economic-survey-1-2019>
23. Williams, E.J., Kloot, N.H.: Interpolation in a series of correlated observations. *Aust. J. Appl. Sci.* **4**(1), 1–17 (1953), <http://hdl.handle.net/102.100.100/336863?index=1>

Part III

Time Series Prediction Applications

Development of Algorithm for Forecasting System Software



Mostafa Abotaleb  and Tatiana Makarovskikh 

Abstract Forecast systems related to forecasting infection cases of Covid-19 are based on time series models because they are considered to be highly accurate in forecasting Covid-19 cases due to their accuracy over epidemiological models that are related to forecasting Covid-19 cases. In this paper, we have two tasks. The first task is to improve forecasting and decrease MAPE% errors in forecasting infection cases through the development of the “Epidemic.TA” system. The development of this algorithm will be called the ensembling time series and neural network system (ET-system). The development of the system was completed by adding a cubic smoothing spline model. This system also applies the method of ensembling between these models in the system (neural network autoregression, Box-Cox transformation, ARMA residuals Trend and Seasonality, trigonometric Box-Cox transformation, ARMA residuals Trend and Seasonality, Holt’s linear trend, autoregressive integrated moving average, and cubic smoothing splines). We applied ensembling by using two methods. The first is the aggregation (average) of results from these models, and the second is ensembling by using average weight by using a prioritizer. The prioritizer gives weights to time series models and neural network models and then gets the ensembling model’s average weight and compares the errors between these models to choose the best forecast model. The results of the developed system (ET-system) were more accurate than the “Epidemic.TA.” On the other hand, the second task in this paper is to use the bootstrap aggregating (bagging) methodology for the NNAR model to decrease the error value of the peak of the wave of infection cases.

The work was supported by Act 211 Government of the Russian Federation, contract No. 02.A03.21.0011. The work was supported by the Ministry of Science and Higher Education of the Russian Federation (government order FENU-2020-0022).

M. Abotaleb  · T. Makarovskikh

Department of System Programming, South Ural State University, Chelyabinsk, Russia
e-mail: abotalebmostafa@bk.ru; Makarovskikh.T.A@susu.ru

Keywords Time series models · Neural network model · Cubic smoothing spline model · Holt's linear trend model · BATS model · TBATS model · ARIMA model · Epidemic.TA system · Covid-19

The List of Acronyms

ET-system	<u>E</u> nse <u>m</u> bling <u>t</u> ime <u>s</u> eries and neural network model <u>s</u> ystem
E.W	<u>E</u> nse <u>m</u> bling models by using <u>w</u> eight <u>a</u> verage
E.A	<u>E</u> nse <u>m</u> bling models by using <u>a</u> verage
Bagging	<u>B</u> ootstrap <u>a</u> gg <u>e</u> rating
NNAR	<u>N</u> eural <u>n</u> etwork <u>a</u> uto <u>r</u> egression
BATS	<u>B</u> ox-Cox <u>T</u> ransformation <u>A</u> RMA <u>R</u> esiduals, <u>T</u> rend and <u>S</u> easonality
ARIMA	<u>A</u> uto <u>regressive <u>i</u>ntegrated <u>m</u>oving <u>a</u>verage</u>
NHS 111 calls	<u>N</u> ational <u>H</u> ealth <u>S</u> ervice 111 calls

1 Introduction

Since the beginning of Covid-19 in Wuhan, China, mathematical models and time series models have been powerful tools for modeling and forecasting of Covid-19 infection cases. In [1] we compared two models for forecasting infection, deaths, and recovery in three countries, and we concluded that Holt's linear trend is better than the ARIMA model in these three countries. In [2] we concluded that without periodically updating the model's hyperparameters, it is difficult to obtain a highly accurate forecast of Covid-19 cases. As a result, the development of a dynamic system to automatically select the best forecasting model and its best parameters is critical to improving forecasting. In [3] we developed an "Epidemic.Network" system that has been implemented and includes BATS, TBATS, Holt's linear trend, ARIMA, and SIR models. It was a SIR model with the highest error rate (mean absolute percentage error) MAPE % for forecasting infection cases in Chelyabinsk.

In [4] they conducted an experimental study on the forecasting of the Covid-19 epidemic pattern and compared the actual and predicted values in both principle and practical aspects. The ARIMA model was used to provide an effective linear model for capturing the linear pattern of the Covid-19 series. The ARIMA model can display (1) AR for past values and (2) MA for current and previous residual series historical knowledge. Decomposition methods are most effective when the sequence matches the decomposition hypothesis. The weakness of the model is that only the data from the time series can derive linear relationships. This does not work well with occurrences that are influenced by several elements, such as meteorological and unique societal effects. When used in other cases, the findings based on a particular disease may not be replicable when applied to other cases. Moreover, there are several other theories about the long-term trend in methods

of decomposition, which assume a nonlinear function in the time series, such as support vector machine (SVM) and generalized models.

In [5] the cubic-spline, Holt, and Holt-Winter models performed well in the majority of our experiments. In [6] they used NNAR (1,1) and ARIMA (0,2,1) for forecasting the infection fatality rate of Covid-19 in Brazil. They concluded that the NNAR model is better than the ARIMA model, an error rate of 6.85% for NNAR and 7.11% for ARIMA.

In the Isfahan province of Iran, in [7] they simulated and forecasted the infection cases of Covid-19 from February 14 to April 11. There are three scenarios that differ in terms of the stringency level of social distancing. Although the constructed SIR model was able to forecast at short-term intervals, in the long term, it was unable to forecast the actual infection cases' spread and pattern of Covid-19. Remarkably, most of the published SIR models developed to forecast Covid-19 for other communities suffered from similar features. In addition, the based assumptions of the SIR model do not seem appropriate in the case of Covid-19 [7].

In [8] we developed the “Epidemic.TA” system that included the neural network model NNAR; the time series models BATS, TBATS, and Holt’s linear trend; and the ARIMA model. The results of “Epidemic.TA” are very accurate for forecasting cumulative infection cases, and we excluded the SIR model from this system, since it produced the highest error rates.

In [9] the authors concluded that the multilayer perceptron network (MLP) is the best for forecasting daily infection cases, and the Holt-Winter model is good for death cases. They also forecasted that they would have 2484 infection cases and 114 death cases on September 14, 2020, but actually had 2089 infection cases and 128 death cases. It means that MAPE was equal to 18.91% for infection cases and to 10.94% for death cases, and MAPE for 30 days for infection cases is 17.49% and for death cases is 13.53%. This means that the best models were not able to forecast for a month with minimal errors.

In [10] the authors concluded that the ARIMA model and cubic smoothing spline models have lower forecast errors and narrower forecast intervals compared to the Holt and TBATS models.

In this study [11], the authors have created a simple model for forecasting that can be used to forecast Covid-19 daily infection cases at the local level. The proposed MLR model exploits the relationship between the infection cases and the phone call data (NHS 111 calls) in addition to other patterns, such as trends, the effect of weekends, and autoregressive lags of confirmed cases. They compared the performance of the model with ETS, ARIMA, seasonal naive, Prophet, and an MLR model without using phone call data using an empirical study. The analysis showed that the proposed model could provide accurate and reliable forecasts. It outperforms all benchmarks based on all accuracy measures considered in the study. They also provide evidence that using phone call data is an important predictor of Covid-19 confirmed cases and should be considered in forecasting models. They could propose that this might be due to the connection between phone calls to the health service and the dynamics related to Covid-19. It is very hard to get data about the numbers of calls for each country. That makes that new model very hard

to implement. So we decided to develop a system ensembling time series and neural network system (ET-system) dependent on time series models, implementing initial data of infection cases of Covid-19, but not using calls (NHS 111 calls) data.

In our paper, we developed the “Epidemic.TA” system to improve forecasting and reduce the error of MAPE% for daily infection cases. The newly developed system is called ensembling time series and neural network system (ET-system).¹ In this system, we added a cubic smoothing spline model. On the other hand, to improve the forecast, we used two ensembling methods by using two approaches: (1) the aggregation (average) of results from time series and neural network models (NNAR, BATS, TBATS, Holt’s linear trend, ARIMA, and cubic smoothing splines) and (2) applying the average weight by using a prioritizer, which gives weights to time series models that were previously mentioned, and then getting the ensembling model’s average weight.

In [8] we were able to forecast the date of the occurrence of the third wave peak in both Italy and Spain. We obtained accurate results on the date of the onset of maximum Covid-19 infection cases, which coincided with the actual time. On February 22, 2021, we simulated by using the NNAR model to anticipate the occurrence of the third wave in the Russian Federation, and tested the last 50 days from that date, where the third wave was forecasted on July 19, 2021, but the actual timing of the wave was July 9, 2021. Here comes the second task of this work, which is to improve forecasting to reduce errors in the value of the peak. For example, we will implement a bootstrap aggregating (bagging) NNAR to minimize errors in forecasting in the Russian Federation and Chelyabinsk and on other hand, comparative between its errors to choose minimize the least errors in forecasting.

2 Data and Materials

To hold our experiments, we used the Covid-19 data set from January 1, 2020, to August 15, 2021, for the Russian Federation from the World Health Organization (WHO), and the data set[12] for Chelyabinsk from March 12, 2020, to August 15, 2021, by Yandex DataLens [13]. The lengths of time series data set about Covid-19 infection cases used in our experiments are shown in Table 1.

It is possible to download the source code for the developed ensembling time series and neural network system (ET-system) by using R-programming and data sets from GitHub [14].

¹ Here and later, the acronyms of ensembling time series and neural network system (ET-system) are listed at the end of the Introduction section.

Table 1 Data set used for forecasting waves of infection cases in the Russian Federation

Russian Federation			
Wave number	First value date	Last value date	The peak of the wave date
1	January 1, 2020	April 30, 2020	May 11, 2020
2	January 1, 2020	November 30, 2020	December 24, 2020
3	January 1, 2020	June 30, 2021	July 9, 2021
Chelyabinsk region			
1	March 12, 2020	May 31, 2020	June 23, 2020
2	March 12, 2020	November 30, 2020	December 19, 2020
3	March 12, 2020	August 10, 2021	Unknown

3 The Review of Ensembling Time Series and Neural Network System (ET-System) for Forecasting Covid-19 Cases and Waves for Infection Cases

3.1 *The Algorithm Schema of the Ensembling Time Series and Neural Network System (ET-System)*

From Fig. 1 the following six steps describe how the algorithms of the ensembling time series and neural network system (ET-system) work:

- Step 1.** Insert Covid-19 time series and global variables; see [14] and [8].
- Step 2.** Preprocess data and split the data into training and testing.
- Step 3.** Run TS-System and ensembling time series and neural network system (ET-system).
- Step 4.** Calculate the accuracy of the training data (ME-RMSE-MAE-MPE-MAPE-MASE-ACF1).
- Step 5.** Calculate the accuracy of the testing data (MAPE%).
- Step 6.** Select the best model for forecasting with the least error MAPE%.

3.2 *The Scheme of the Algorithm for Dynamic Prioritizer*

Figure 2 describes how the prioritizer works. The prioritizer gives weights. After obtaining errors of trained data from the time series models and neural networks, they are ensembling and given weights. It was found that by giving a weight of 0.9 to the best model of the time series and neural networks, and distributing (1–0.9) equally over the other models, gives accurate results and improves forecasts with low error. For more details, the prioritizer worked after the data was trained for Covid-19, where the errors obtained were calculated from the trained data incrementally. The weights are distributed based on these errors, so the best model that has the lowest mean absolute percentage error (MAPE) gets 0.9, and the other models are distributed (1–0.9) equally over the other models.

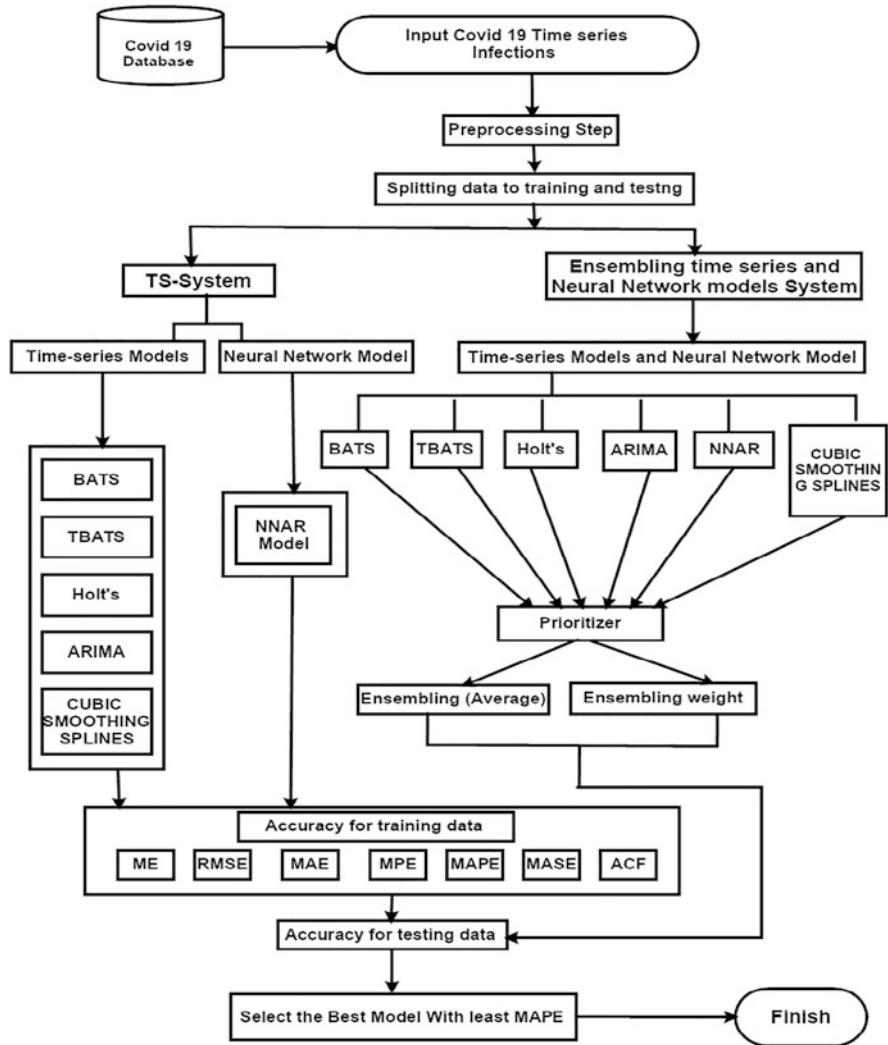


Fig. 1 Scheme of ensembling time series and neural network model system

3.3 The Scheme for Bagging and Bootstrapping the NNAR Model for Improving Forecasting of the Waves of Infection Cases

When we used the NNAR model to forecast the peak of the third wave of infection cases in the Russian Federation, [8], we used the NNAR (8,50) model (Fig. 3).

Let us consider the improvement in forecasting the peaks of infection cases and introduce the new methodology of bootstrapped time series (Tables 6 and 7) to

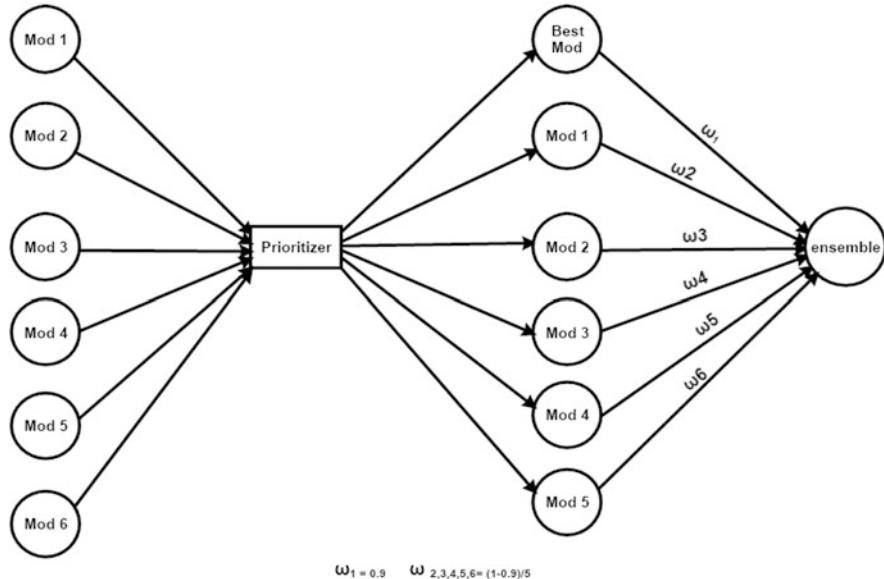


Fig. 2 The scheme of the algorithm for dynamic prioritizer

improve forecast accuracy. So, in our case, we will divide training data into three samples with replacement, as it is possible to divide training data into 100 samples with replacement. The main reason for choosing to divide training data into three samples with replacement is that our Covid-19 data (pattern of data) finds that we get the least error when divided into three samples with replacement. If we produce forecasts from each of the samples and average the resulting forecasts, we get better forecasts than if we simply forecast the training data directly. Figure 4 describes the development of the NNAR model for forecasting waves by using bagging and bootstrapping.

3.4 Design of the Software for Forecasting

Figure 3 describes how this software deals with univariate time series data where the schema describes the dynamic of this software.

- Step 1.** Start software.
- Step 2.** Input global variables and time series data; see [14] and [8].
- Step 3.** Run the (ET-system) algorithm on software.
- Step 4.** Extract results for testing data, accuracy, plots, and graphs, and forecasting in txt and csv format.
- Step 5.** Finish.

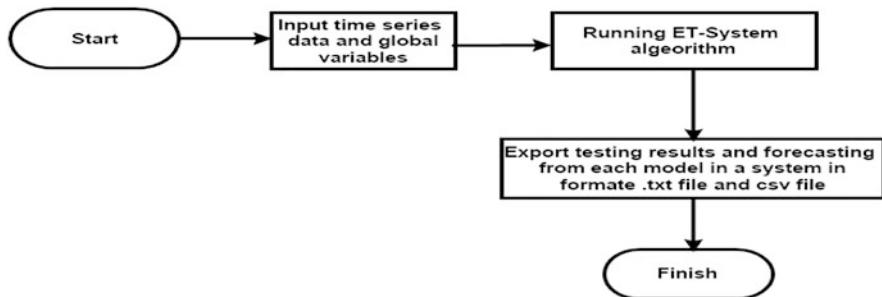


Fig. 3 The scheme of the advanced software system for forecasting Covid-19

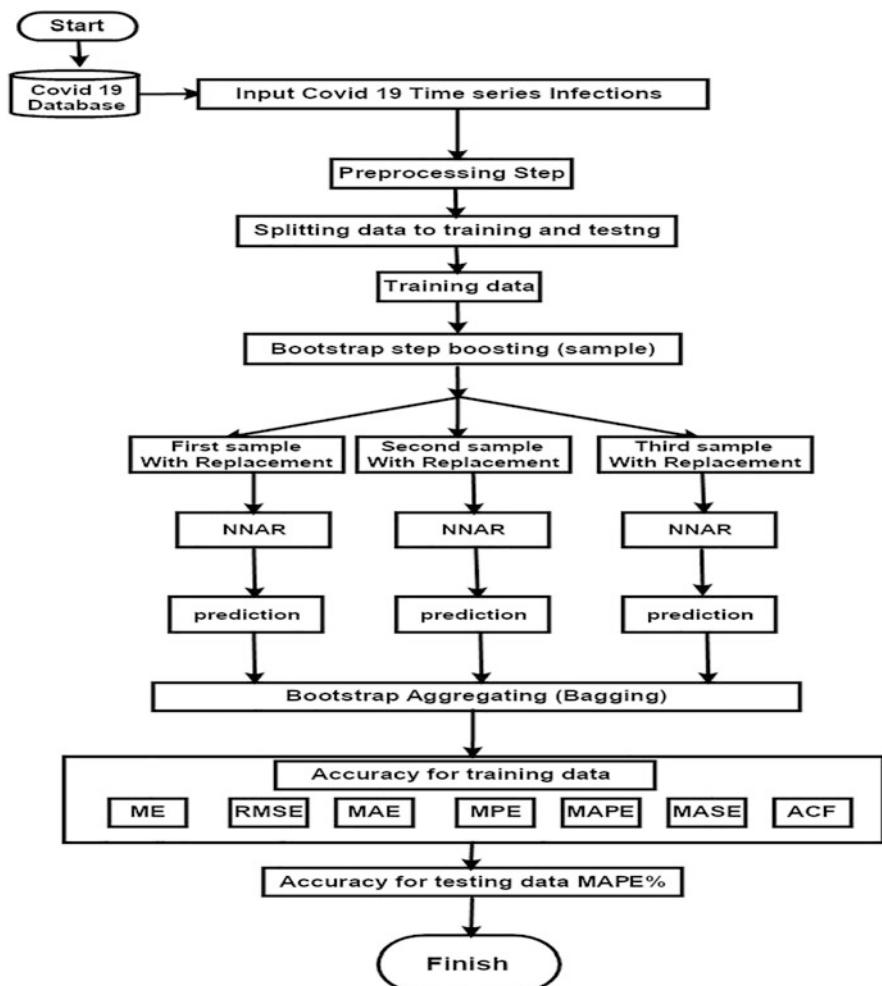


Fig. 4 Developed NNAR for forecasting the waves of infection cases of Covid-19 by using bootstrap aggregating (bagging) for the NNAR model

4 Results

The simulation results are shown in Tables 2, 3, 4, 5, 6, 7. From Tables 2 and 3

The experiment results were performed using four selection or prediction criteria: testing lasted 7 days, 14 days, 21 days, and 30 days, respectively. With the help of these measures, we have found that our proposed system for ensembling average and ensembling average weight performance is better than all other models, which achieved the lowest error rates compared to other models for data used in experiments (Russian Federation and Chelyabinsk).

Tables 4 and 5 show the experiment held for the data, including the period from August 9, 2021, to August 15, 2021 (tested last 7 days), in the Russian Federation and Chelyabinsk region. For the Russian Federation, the range of error is between 0.605 % and 3.032 % and MAPE for the tested period is 1.644 %. For Chelyabinsk, the range of error is between 0.173% and 0.992% and the MAPE for the tested period is 0.580%. As soon as all the obtained MAPE is lower than 1–3%, we can conclude that the obtained results have very high accuracy.

Table 2 Accuracy MAPE% daily Covid-19 infection cases for testing data last (7-14-21-30) days for Russian Federation

Model	7 days	14 days	21 days	30 days
NNAR model	2.468	7.412	4.177	13.105
BATS model	3.952	3.384	1.518	10.425
TBATS model	4.223	3.957	11.739	30.398
Holt's linear trend model	3.937	3.308	11.310	31.794
ARIMA model	4.269	3.950	6.580	9.791
Cubic smoothing spline model	1.838	4.111	6.181	9.276
Ensembling average	2.156	1.573	4.480	14.391
Ensembling average weight	1.644	3.041	1.200	6.439
Best model	E.W	E.A	E.W	E.W

Table 3 Accuracy MAPE% daily Covid-19 infection cases for testing data last (7-14-21-30) days for Chelyabinsk

Model	7 days	14 days	21 days	30 days
NNAR model	2.460	10.613	13.879	11.590
BATS model	1.009	4.318	4.135	11.337
TBATS model	1.202	4.220	3.702	7.894
Holt's linear trend model	1.186	3.822	4.451	7.822
ARIMA model	2.756	4.635	9.837	14.532
Cubic smoothing spline model	5.851	11.519	17.656	27.299
Ensembling average	0.580	1.438	1.003	4.705
Ensembling average weight	0.881	3.536	3.338	7.448
Best model	E.A	E.A	E.A	E.A

Table 4 Accuracy MAPE% daily Covid-19 infection cases for forecasted 1 week ahead by using best model ensembling weight average

Date	Actual	Forecasted	MAPE %
August 9, 2021	22,160	21,820.300	1.533 %
August 10, 2021	21,378	21,755.740	1.767 %
August 11, 2021	21,571	21,705.190	0.622 %
August 12, 2021	21,932	21,653.990	1.268 %
August 13, 2021	22,277	21,601.640	3.032 %
August 14, 2021	22,144	21,550.930	2.678 %
August 15, 2021	21,624	21,493.070	0.605 %
Weekly MAPE % for forecasted daily infection cases by using best model E.W			1.644 %

Table 5 Accuracy MAPE% daily Covid-19 infection cases for forecasted 1 week ahead (Chelyabinsk region) by using best model ensembling average

Date	Actual	Forecasted	MAPE %
August 9, 2021	360	363.572	0.992 %
August 10, 2021	363	365.776	0.765 %
August 11, 2021	365	367.997	0.821 %
August 12, 2021	368	369.696	0.461 %
August 13, 2021	371	371.642	0.173 %
August 14, 2021	375	373.949	0.28 %
August 15, 2021	378	375.847	0.569 %
Weekly MAPE % for forecasted daily infection cases by using best model E.A			0.580%

Table 6 Forecasted peaks of Covid-19 infection cases waves for the Russian Federation

NNAR bootstrap model					
No. peak of wave	Actual value	Forecasted value	MAPE%	Testing days	Bootstrapping
1	11656	10885.7	6.61%	Last 7days	100
2	29935	27250.98	8.97%	Last 25 days	3
3	25766	25885.61	0.46%	Last 120 days	3
NNAR model					
1	11656	5364.59	53.98%	Last 3 days	-
2	29935	18919.69	36.80%	Last 25 days	-
3	25766	23191.05	9.99%	Last 120 days	-

From Table 8 (see Appendix), we implemented the ET-system for forecasting daily Covid-19 infection cases in the Russian Federation and in Chelyabinsk to the end of August 2021 by using the best model which achieved the lowest error. We forecasted for the Russian Federation by using an ensembling average weight and for Chelyabinsk by using an ensembling average, which achieved the least error of MAPE in the (ET-system) for testing in the last 7 days. See Tables 4 and 5 tables.

From Table 6 and 7 we simulated the NNAR bootstrap model for the Russian Federation and Chelyabinsk to forecast the peak value of the first, second, and third waves. All the experiments were performed using three selection or prediction

Table 7 Forecasted peaks of Covid-19 infection cases waves for Chelyabinsk region

NNAR bootstrap model					
No. peak of wave	Actual value	Forecasted value	MAPE %	Testing days	Bootstrapping
1	258	242.13	6.15	Last 7 days	3
2	317	318	0.32	Last 25 days	3
3	386	380.78	1.35	last 3 days	3
NNAR model					
1	258	128.81	50.07	Last 7 days	-
2	317	218.80	30.98	Last 25 days	-
3	386	298.86	22.58	last 3 days	-

criteria. Testing lasts 7 days, 25 days, and 120 days. With the help of these measures, we have found that our bootstrap aggregating (bagging) performance is better than NNAR models. The proposed model achieves the lowest MAPE% throughout the experiment under various selection criteria.

$$MAPE\% = \left| \frac{Actual\ value\ in\ the\ wave - Forecasted\ value\ in\ the\ wave\ date}{Actual\ value\ in\ the\ wave} \right| * 100 \quad (1)$$

5 Conclusions and Further Research

Time series for infection: The cases of Covid-19 are the ones for which classical time series models no longer have sufficient ability to accurately forecast future values. It is obvious, because there are lots of different factors influencing the process. The forecast obtained today is suitable only for the current situation with a fixed number of infected in hospitals, a fixed number of vaccinated, a fixed policy, etc. Surely, it is impossible to fix any of these factors in life. In our paper, we considered the approach in which we combined the existing models by using ensembling aggregation weights and ensembling aggregation results of these models to obtain accurate predictions in the short term, which may extend to 10 days. This indicates the importance of developing and building new models that can detect the pattern of spreading Covid-19 infection.

For the experimental evaluation, we compared the performance of six traditional forecasting models: (1) neural network autoregressive model (NNAR), (2) BATS, (3) TBATS, (4) Holt's linear trend, (5) ARIMA, and (6) cubic smoothing spline model to find out their suitability and correctness. The mean absolute percentage error (MAPE) has been used as a performance measure. The performance of each model has been calculated using these performance measures to determine the best suitable forecasting model among them. All the experiments were performed using four selection or prediction criteria: (1) testing lasts 7 days, (2) 14 days, (3) 21 days,

and (4) 30 days. With the help of these measures, we have found that our proposed model's performance is better than all other models. The proposed model achieves the lowest MAPE throughout the experiment under various selection criteria. Apart from that, we have found the significance of bootstrapping and bagging, as well as the importance of the ensembling average and the ensembling average weight. To obtain highly accurate forecasts, the data must be updated weekly.

We have also used the proposed model to forecast the number of daily Covid-19 confirmed cases at the national level, the Russian Federation, and the Chelyabinsk region. The overall performance is very similar to the local level. The proposed model, ensembling average and ensembling average weight, outperforms others. However, we have not provided the analysis here due to the space limit and the focus of the study at the local level, but it can be provided on request.

In the future, we will again analyze the proposed model with different data sets and find out further boosting techniques which can boost the model efficiency. The bootstrapping and bagging process, method, and forecasting the extreme values of functions is one of the probable solutions. It will be appended to the software, and complete the software.

Appendix

See Table 8.

Table 8 Forecasted infection cases of Covid-19 (with prediction criteria: testing lasts 7 days)

Date	Russian Federation			Chelyabinsk		
	Date	Actual	Forecasted	MAPE%	Actual	Forecasted
August 16, 2021	20,765	21,431.630	3.21	381	377.962	0.80
August 17, 2021	20,958	21,374.780	1.99	384	380.142	1.00
August 18, 2021	20,914	21,311.840	1.90	386	381.691	1.12
August 19, 2021	21,058	21,264.120	0.98	385	383.568	0.37
August 20, 2021	20,992	21,212.800	1.05	383	385.738	0.71
August 21, 2021	21,000	21,154.790	0.74	382	387.592	1.46
August 22, 2021	20,564	21,099.460	2.60	380	389.641	2.54
August 23, 2021	19,454	21,047.150	8.19	379	391.712	3.35
August 24, 2021	18,833	20,984.740	11.43	377	393.250	4.31
August 25, 2021	19,536	20,925.730	7.11	378	394.998	4.50
August 26, 2021	19,630	20,874.220	6.34	377	397.156	5.35
August 27, 2021	19,509	20,819.720	6.72	379	398.940	5.26
August 28, 2021	19,492	20,767.670	6.54	378	400.913	6.06
August 29, 2021	19,286	20,715.250	7.41	377	402.984	6.63
August 30, 2021	18,325	20,657.720	12.73	377	404.410	7.27
August 31, 2021	17,813	20,598.640	15.64	376	406.165	8.02
MAPE%	MAPE% for 16 days		4.35	MAPE% for 16 days		3.10
Best model	Ensembling average weight			Ensembling average		

References

1. Abotaleb, M.S.: Predicting Covid-19 cases using some statistical models: An application to the cases reported in China, Italy and USA. *Acad. J. Appl. Math. Sci.* **6**(4), 32–40 (2020). <https://doi.org/10.32861/ajams.64.32.40>
2. Makarovskikh, T.A., Abotaleb, M.S.A.: Automatic selection of ARIMA model parameters to forecast Covid-19 infection and death cases. *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya Vychislitel'naya Matematika i Informatika* **10**(2), 20–37 (2021). <https://doi.org/10.14529/cmse210202>
3. Abotaleb, M.S.A., Makarovskikh, T.A.: Development of algorithms for choosing the best time series models and neural networks to predict Covid-19 cases. *Bull. South Ural State Univ. Ser. Comput. Tech. Autom. Control Radio Electron.* **21**(3), 26–35 (2021). <https://doi.org/10.14529/ctcr210303>
4. Roy, S., Bhunia, G.S., Shit, P.K.: Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Model. Earth Syst. Environ.* **7**, 1385–1391 (2021). <https://doi.org/10.1007/s40808-020-00890-y>
5. Al-Turaiki, I., Almutlaq, F., Alrasheed, H., Alballa, N.: Empirical evaluation of alternative time-series models for COVID-19 forecasting in Saudi Arabia. *Int. J. Environ. Res. Public Health* **18**(16), 8660 (2021). <https://doi.org/10.3390/ijerph18168660>
6. Ahmar, A.S., Boj, E.: Application of neural network time series (NNAR) and ARIMA to forecast infection fatality rate (IFR) of COVID-19 in Brazil. *JOIV Int. J. Inf. Vis.* **5**(1), 8–10 (2021). <https://doi.org/10.30630/joiv.5.1.372>
7. Moein, S., Nickaeen, N., Roointan, A., Borhani, N., Heidary, Z., Javanmard, S.H., Ghaisari, J., Gheisari, Y.: Inefficiency of SIR models in forecasting COVID-19 epidemic: A case study of Isfahan. *Scientific Reports* **11**(1), 1–9 (2021). <https://doi.org/10.1038/s41598-021-84055-6>
8. Abotaleb, M., Makarovskikh, T.: System for forecasting COVID-19 cases using time-series and neural networks models. In: *Engineering Proceedings* (Vol. 5(1), p. 46). Multidisciplinary Digital Publishing Institute (2021). <https://doi.org/10.3390/engproc2021005046>
9. Talkhi, N., Fatemi, N.A., Ataei, Z., Nooghabи, M.J.: Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. *Biomed. Signal Process. Control* **66**, 102494 (2021). <https://doi.org/10.1016/j.bspc.2021.102494>
10. Gecili, E., Ziady, A., Szczesniak, R.D.: Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy. *Plos one* **16**(1), e0244173 (2021). <https://doi.org/10.1371/journal.pone.0244173>
11. Rostami-Tabar, B., Rendon-Sanchez, J.F.: Forecasting COVID-19 daily cases using phone call data. *Appl. Soft Comput.* **100**, 106932 (2021). <https://doi.org/10.1016/j.asoc.2020.106932>
12. World Health Organization: <https://covid19.who.int/info/>. Accessed 31 July 2021
13. Yandex DataLens: <https://datalens.yandex.ru/>. Accessed 12 Aug 2021
14. Abotaleb, M., Makarovskikh, T.: “E-System” or ensembling time series and neural network-system (ET-System) for forecasting Covid-19 infection casses. <https://github.com/abotalebmostafa11/E-System>. Accessed 11 Aug 2021

Forecasting High-Frequency Electricity Demand in Uruguay



Bibiana Lanzilotta and Silvia Rodríguez-Collazo

Abstract This paper proposes a model for the daily electricity demand in Uruguay, identifying the incidence of special days (calendar effects, holidays, among others) and climatic variables such as temperature, humidity, winds, and heliophany. We propose a non-linear model to represent the association between energy consumption and climate variables. Applying Markov switching models and considering hot and cold months separately, identify breaks in the energy demand function associated with temperature thresholds. Predictive analysis during 2020, the first year of the health emergency, shows that the COVID-19 sanitary crisis did not deteriorate the model performance.

Keywords Non-linear time series models · Daily time series · Electricity · Climatic variables

1 Introduction

The electric sector has traditionally had an intensive use of predictive models. In vertical integration environments, with tariffs centrally fixed by regulators, the punctual forecast of daily domestic demand of energy is an essential requirement to accomplish an efficient generation, given the overrun costs associated with a poor prediction. Overestimating implies generating energy that will not be consumed, and therefore it will be lost or sold to derisory prices. On the other side, underestimating the demand may cause blackouts and high costs [1]. By modeling the demand, it is possible to acquire a more refined knowledge of consumers and markets, as well as a better positioning, by reducing uncertainty when making decisions.

B. Lanzilotta () · S. Rodríguez-Collazo

Facultad de Ciencias Económicas y de Administración, Universidad de la República del Uruguay, Montevideo, Uruguay

Centro de Investigaciones Económicas, Montevideo, Uruguay
e-mail: bibiana.lanzilotta@fceia.edu.uy; silvia.rodriguez@fceia.edu.uy

Our objective is to characterize and model the daily demand for electric energy in Uruguay between 01.01.2010 and 31.07.2020, identifying the incidence of specific events, individually and integrated with the short-term dynamic. Following [2, 3], a daily single model is proposed that accounts for the non-linear association between energy consumption and climatic variables.

In Uruguay, the daily electricity demand has significantly grown in the past years, doubling between the early 1990s and 2020. The Uruguayan electricity generation system is acquiesced by a single state-owned generator (Usinas y Transmisiones Eléctricas, UTE), mainly based on hydraulic, wind, and thermal sources.

As is known, electric energy demand presents different seasonal patterns throughout the year. On one side, the seasonal factor associated with the climate seasons, in general, the highest peaks take place in winter (in Uruguay, between June and August), and more recently, in summer (December to March in Uruguay). In fall and spring, the demand is lower due to the moderation of the temperature and the climatic variables in general. In Uruguay, seasonality associated with the seasons has changed since the '90s, diminishing the gap between winter and summer, but always maintaining higher levels in the cold season of the year. Regarding the weekly pattern, peaks and valleys are repeated with a 7-day frequency, which is mainly explained by the dynamic of the economic activity. This pattern does not suffer significant changes in the last decades.

A wide range of methodologies and models for electricity forecasting are given in the literature. Some methods are based on statistical and econometrics models while other ones are based on computational models (see [4] for a comprehensive systematic review).

This research, within a time-series methodological framework, follows [1, 5, 6] proposals. This approach has the advantage, over the computational methodologies, of providing an interpretable explanation of the behavior of the variable, in addition to its forecast. This paper updates and revises a previous one [7], which considers data till 18.11.2012. The updated results show changes in the demand curve as a function of temperature, probably linked to changes in the uses of electrical energy when temperatures rise above the annual average.

The document is organized as follows. The following section presents the methodological approach and, while in the third section, a brief characterization of the electric energy demand in Uruguay is exposed. The estimated model and its predictive performance in a regular year and during the COVID-19 health emergency are presented in the fourth section. Finally, the fifth section concludes.

2 Methodological Approach

[2, 3], following [1, 5, 6] formulate a forecasting method for the energetic demand in Spain using high-frequency data. In order to do so, they focus on the non-linear link between energetic consumption and climate variables, incorporating a detailed

analysis of intervention on special days (holidays, vacations, strikes), all in a single equation for the whole sample.

2.1 General Model and the Treatment of Special Days

Equation (1) represents daily electricity demand (D_t) as:

$$D_t = FS_t + ES_t + CDE_t + CVC_t + \varepsilon_t, \quad (1)$$

being FS_t a trend associated with socioeconomic factors that influence the energetic demand, ES_t the weekly seasonal pattern, CDE_t the special days contribution to the energetic demand, CVC_t the climate variables contribution, and ε_t random shocks not taken into account in any previous variables. Excluding the most volatile components (the contributions of special days and the climate variables) to the total demand, we have the DB_t (the most stable component) that can be expressed as an ARIMA model:

$$\Delta\Delta_7 DB_t = \eta_t, \quad (2)$$

being η_t a stationary ARMA(p,q) process. Combining Eqs. (1) and (2) we can write:

$$\Delta\Delta_7 D_t = \Delta\Delta_7 CDE_t + \Delta\Delta_7 CVC_t + \eta_t, \quad (3)$$

Expressing CDE_t and CVC_t as vectors of polynomials associated with de L lag operator, $CDE_t = f_1(L)' DE_t$, and $CVC_t = f_2(L)' VC_t$ (with DE_t , an **mxt** a matrix of **m** special days variables, and VC_t an **nxt** matrix of the **n** climate variables) we have:

$$\Delta\Delta_7 D_t = \Delta\Delta_7 f_1(L)' DE_t + \Delta\Delta_7 f_2(L)' VC_t + \eta_t, \quad (4)$$

From Eq. (4) we estimate a function that allows us to forecast the short-term Uruguayan electricity demand. Considering the dependent variable in logarithms does not introduce substantial distortions in the short-term predictions, while it allows improving its variance. Equation (4) models all the effects associated with socioeconomic variables, such as the country's growth, prices, demographic changes, and seasonal effects derived from differences in the series studied. While the first difference aims to eliminate the previously described effects, the difference of order 7 models the differential behavior of the energetic demand between weekdays. The usage of ARIMA models for this kind of modeling, and mainly to forecasting is well documented in [2, 3, 6, 8–12]. Once special day effects are adjusted, the remaining components that include the climate variable effects are denominated, the electricity demand depurated from special days effects (DAD), that is represented in Eq. (5).

$$\Delta\Delta_7 DAD_t = \varphi_t, \quad (5)$$

2.2 A Non-linear Approach to Model the Effect of Climate Variables

One of the unarguable features about the link between climate variables and energetic demand is its nonlinearity. On one side, when temperature is low, a rise in it carries a reduction in the energy demand; this is known as the “heating effect” [13]. On the other side, the “cooling effect” takes place when a rise in the temperature implies an increment of the energy demand (given that strong heats encourage refrigerating electrical appliances).

Most of the existing literature tries to model this functional form by threshold variables, arbitrarily fixed dummies. Another way to tackle this consists of estimating the energy demand function through estimated splines from non-parametric models, which not only seek to find a value for function parameters but also the functional form [14]. Our approach departs from [2, 3] that postulate a non-linear link between the different observed temperatures, approximated by piecewise functions. Our contribution consists of using the Markov Switching Models methodology to estimate the breakpoints of the demand function, using the demand calculated in Eq. (4), sectioning the sample into the hottest and the coldest months of the year. In addition to the variables used by [2], we include different climatic variables such as wind, relative humidity, and sunlight (heliophany) that affect the apparent temperature and therefore the thermal comfort needs.

An additional feature to consider in the modeling of energetic demand is the inertia of the climate effect on it. The environments where we live have the characteristic of keeping ambient temperature, at least for a few days. This is why even if there is a hot day during winter we still need to heat our houses, so energy demand will not decrease despite higher temperatures. For the same reason, a given temperature does not have the same effect in summer as in winter, and different breaking points have to be found for different seasons. To capture this effect we include two qualitative variables: *cold* and *warm*, which reflect the months when the average temperature is higher than the year average temperature and those when it is not, respectively. Considering the average temperature in the last 5 years months May, June, July, August, September, and October are included in the *cold* dummy. *Warm* dummy is defined by difference.

Two-step Procedure for Estimating Breaking Points

Starting from the demand adjusted for the effect of special days (Eq. 5), we apply Markov’s Switching Models methodology to find the breaking points in the link between energetic demand and temperature. In order to do so, we propose a two-stage methodology.

The first step is the estimation of a linear function to determine relevant climatic variables to model energetic demand, as well as its structure and main outliers (Eq. 6).

$$\begin{aligned}
\Delta \Delta_7 DAD_t = & \Delta \Delta_7 Temp_t f_{21}(L)' Warm_t + \Delta \Delta_7 Temp_t f_{22}(L)' Cold_t \\
& + \Delta \Delta_7 Heliophany_t f_{31}(L)' Warm_t + \Delta \Delta_7 Heliophany_t f_{32}(L)' Cold_t \\
& + \Delta \Delta_7 RH_t f_{41}(L)' Warm_t + \Delta \Delta_7 RH_t f_{32}(L)' Cold_t \\
& + \Delta \Delta_7 Wind_t f_{51}(L)' Warm_t + \Delta \Delta_7 Wind_t f_{52}(L)' Cold_t \\
& + \Delta \Delta_7 WinWinter_t + \Delta \Delta_7 Save_t + \Sigma_1^{11} \Delta \Delta_7 Month_{i,t} \\
& + \Sigma_1^S \Delta \Delta_7 Outlier_{i,t} + \theta(L) / \phi(L) a_t,
\end{aligned} \tag{6}$$

being *Temp*, the average observed temperature measured in Celsius degrees, *Heliophany* the number of hours of light during the day, *RH* the relative humidity, *Wind* the wind speed, *WinWinter* represents the sequence of warm days in the middle of winter, the dummy *Save* captures the times when the generating entity imposed saving measures, and *Outlier*, the binary variables used to correct atypical observations. Finally, an ARMA structure is fitted for residuals.

The second step consists of the estimation of breaking points on energy demand related to the observed temperature applying Markov's Switching Model procedure. Once a breaking point candidate is found, Wald tests were applied, in order to confirm that the coefficients above and below the threshold are significantly different. We discard 5% of the lowest and highest observed temperatures on each season ("warm" and "cold"), in order to count with a sufficient number of observations to perform the first and last breaking test. This simplification does not limit the search of breaking points, because it is not expected to find them in the extreme values. The final estimated equation is the following

$$\begin{aligned}
\Delta \Delta_7 DAD_t = & \Sigma_1^v \Delta \Delta_7 W_t^i f_{21a}(L)' Warm_t + \Delta \Delta_7 (Temp_t - W_t^i) f_{21b}(L)' Warm_t \\
& + \Sigma_1^r \Delta \Delta_7 C_t^i f_{22a}(L)' Cold_t + \Sigma_1^r \Delta \Delta_7 (Temp_t - C_t^i) f_{22b}(L)' Cold_t \\
& + \Delta \Delta_7 Heliophany_t f_{31}(L)' Warm_t + \Delta \Delta_7 Heliophany_t f_{32}(L)' Cold_t \\
& + \Delta \Delta_7 RH_t f_{41}(L)' Warm_t + \Delta \Delta_7 RH_t f_{32}(L)' Cold_t \\
& + \Delta \Delta_7 Wind_t f_{51}(L)' Warm_t + \Delta \Delta_7 Wind_t f_{52}(L)' Cold_t \\
& + \Delta \Delta_7 WinWinter_t + \Delta \Delta_7 Save_t + \Sigma_1^{11} \Delta \Delta_7 Month_{i,t} \\
& + \Sigma_1^S \Delta \Delta_7 Outlier_{i,t} + \theta(L) / \phi(L) a_t.
\end{aligned} \tag{7}$$

where W_t^i and C_t^i are threshold variables for warm and cold season, respectively.

3 The Data

The electricity demand series with daily frequency presents different seasonality as well as long-term growth, explained by economic, social, and technologic factors.

In Uruguay, electricity demand has significantly grown in the past years: while in 1992, the average daily demand was around 14.5 thousand MWh, in the late 90s exceeded the 20 thousand MWh daily average, in 2010, reached 25 thousand MWh, and in 2019, 30 thousand MWh.

During this period, seasonality underwent changes (see Fig. 1). While in the period 1992–2000, the energy consumption during summers represented 88% of that of the winter, between 2010 and 2019 that amount rose to 93%. The change in the seasonal pattern is mainly explained by the universalization of electric appliances to guarantee comfort during summers (particularly, air-conditioning equipment). From 2010 to nowadays, the seasonal pattern seems to stabilize. Regarding the weekly pattern, it seems clear that peaks and valleys are repeated with a 7-day frequency, which is mainly explained by the dynamic of the economic activity. The average of consumption on Saturdays and Sundays represents nearly 95% and 85% (respectively) of the business days demand (Monday, Tuesday, Wednesday, Thursday, Friday).

Our study divides the daily data sample into training data and test data. We use demand data measured in MWh from January 1, 2010 to December 31, 2018 as part of the training sample and the testing sample is from January 1 to December 31,

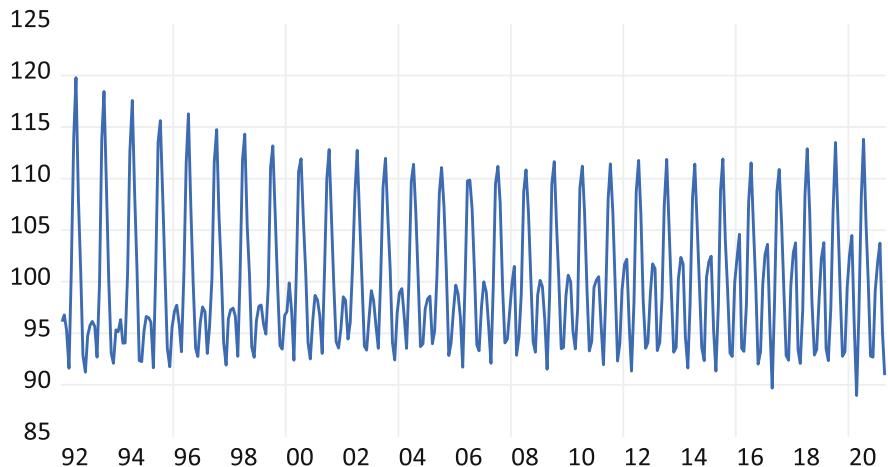


Fig. 1 Seasonal factor of electricity demand (1992–2020). Source: Own elaboration based on UTE data

2019. Subsequently, the predictive performance of the model during the COVID-19 sanitary crisis is analyzed. The out-of-sample evaluation of forecasting accuracy is performed by rolling evaluations [15].

Applying the Augmented Dickey–Fuller (ADF) test, we find that in this period electricity demand presents a regular unit root.¹ As for the seasonal weekly pattern, we apply a seasonal difference to obtain a flexible stationary structure to capture the variations of this component of the series. Therefore, the stationary transformation of the series requires regular and seasonal differencing ($\Delta\Delta_7$). Finally, to stabilize the variability of the series throughout the sample, we took logarithms. This transformation does not distort short-term forecasts. In this way, we will work with a proxy of the energetic demand daily weekly growth rate.

4 Results

4.1 Modelization of Special Days

In the first term, we estimate the impact of special days on electric consumption. There are different methodologies to cope with the problem. [16, 17] model this effect separately, [18] resembles the special days to Sundays, [19] opts to replace them with a similar day of the previous week. An alternative methodological option widely used is to model these days with deterministic variables [3, 8, 20]. This option enables to capture differentially the effects of different special days and then provides better forecasts. In this paper, we follow this last proposal in order to find a stable behavioral pattern that lets us incorporate this information in our forecast models. We consider both workable and not: Easter, Carnival, holidays, and strikes.

Their impact on energetic consumption is different according to which day of the week they occur. They also have lagged and forward effects on the demand. To capture these effects and to reduce the loss of degrees of freedom, we assembled four groups according to the incidence of each holiday in the electricity demand.² Table 1 summarizes the impact of each group and also the coefficients corresponding to Carnival holidays and the Easter effect.

Consequently, for each group, we created seven dummy variables, each one representing a day of the week, in order to capture the different effects of the holidays according to the weekday. 28 variables were included, as well as lagged and forward effects for each one. Variables representing *Easter*, *Carnival*, and strikes were also included in the special day filter equation (Eq. 8).

¹ These results and full estimations are available on request from the authors.

² The individual incidence was estimated in a broader sample (1992–2018).

Table 1 Holiday Groups and effects (average dynamic effects)

Group (G_j)	National holidays	Sample 2010–2019
Group 1	January 1, December 25	-0.097
Group 2	May 1, August 25, March 1	-0.061
Group 3	January 6, July 18, November 2	-0.036
Group 4	April 19, May 18, June 19, October 12	-0.015
Easter		-0.051
Carnival		-0.037

Note: Impacts are on $\Delta\Delta_7\ln D_t$ as specified in [8]. Source: Authors estimations

$$\begin{aligned} \Delta\Delta_7\ln D_t = & \sum_1^4 \Delta\Delta_7 (G_j, t^*Sun_t) f_G(L)' + \sum_1^4 \Delta\Delta_7 (G_j, t^*Sat) f_G(L)', \\ & + \sum_1^4 \Delta\Delta_7 (G_j, t^*Fri_t) f_G(L)' + \sum_1^4 \Delta\Delta_7 (G_j, t^*Thu_t) f_G(L)', \\ & + \sum_1^4 \Delta\Delta_7 (G_j, t^*Wed_t) f_G(L)' + \sum_1^4 \Delta\Delta_7 (G_j, t^*Tue_t) f_G(L)', \\ & + \sum_1^4 \Delta\Delta_7 (G_j, t^*Mon_t) f_G(L)' + \Delta\Delta_7 Easter_t f_{Ea}(L)', \\ & + \Delta\Delta_7 Carn_t f_{Ca}(L)' + \Delta\Delta_7 Strike_t f_{Ca}(L)' + \varphi_t \end{aligned} \quad (8)$$

where G_i is a qualitative variable that represents the holidays on each group i using ones, and *Sun*, *Sat*, *Fri*, *Thu*, *Wed*, *Tue*, *Mon* correspond to representative dummy variables of the days of the week. The interaction of holiday variables (grouped) and day of the week variables allows us to measure the impact of the holidays of the previously defined groups depending on the weekday they fall. *East* indicates Easter Sunday, while *Carn* designates the Carnival holiday. $f_i(L)'$ is a polynomial vector associated with the lag operator L . Finally, $\varphi_t = \Delta\Delta_7 DAD_t$, as is defined in Eq. (5).³

4.2 Nonlinear Modelization of the Effect of Climate Variables

The results are presented in Table 2 and plotted in Fig. 2. We found two breaking points on the energy demand function for each of the defined seasons: *warm* and *cold*. For the warm one, we find the first break at 16 °C. For lower values, no significant effects of the temperature on the energetic demand were found. Given that in the sample the average minimum temperature of a warm-season day is 9 °C, we can establish the zone defined between these two points as a neutral or comfort zone, where the temperature does not affect the electricity demand.

The second estimated break in the warm season occurs at 25 °C. For values between 16 °C and 25 °C, a 1 °C increase in temperature increases the daily growth

³ The estimated coefficients in Eq. (7), for each group of special holidays, are available as Complementary Material.

Table 2 Main results of breaking point estimation on the electric demand function

	Function section	Lag	Coeff.	Σ coeff.
Warm	Between 16 °C and 25 °C	0	0.7424%	
		1	0.1494%	0.8918%
	More than 25 °C	0	0.2942%	
		1	0.0293%	0.3235%
Cold	Less than 10 °C	0	0.5223%	
		1	0.2541%	0.7764%

Source: Authors estimations

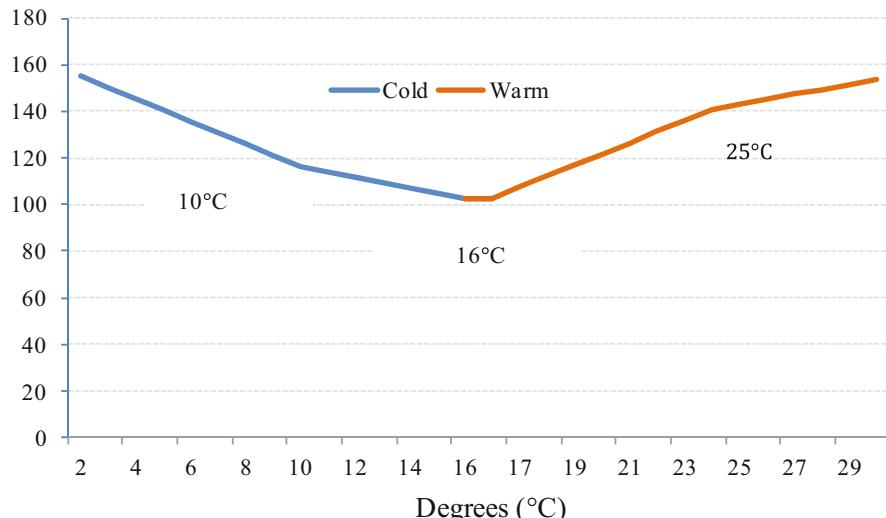


Fig. 2 Graphic representation of temperature impact on the daily weekly growth rate energetic demand, Base Index 100 = 0% growth. Source: Authors estimations

rate compared to that of the previous week by 0.89%, with a contemporaneous effect of 0.74% on the same day and an increase of 0.15% on the following day. For average temperatures observed above 25 °C we found an overall positive effect of the temperature increase of 0.32% on energy demand, which shows that once this threshold is exceeded, each 1-degree increase in temperature has a smaller influence on energy demand than if the same increase were to occur at temperatures below 25 °C. This finding can be explained by the temperature saturation effect on electricity consumption (saturation in the use of cooling equipment).

In the cold season, the first break occurred at 16 °C. Between 10 °C and 16 °C, each degree of temperature drop raises daily growth by 0.37% (compared to the same time the previous week). This contemporaneous effect is 0.23% and 0.13% the following day. The second break is at 10 °C, with 3.7 °C being the average minimum temperature observed in this season during the period analyzed. Between these two values, a 1 °C decrease in temperature increased the daily energy demand

Table 3 Climatic variable effects on electricity demand

	Climatic variable	Lag	Coeff.	Σ coeff.
Warm	Heliophany	0	0.0069%	
		1	0.0357%	0.0426%
	Relative humidity	0	0.0262%	0.0262%
Cold	Heliophany	1	0.0262%	0.0262%
	Wind	1	0.0013%	0.0013%

Source: Authors estimations

growth rate (with respect to the same change in the previous week) by 0.78%, with 0.52% being the contemporaneous effect and 0.25% on the following day. Within this temperature range, the displacement of energy demand per unit degree is the highest in the temperature range between 16 and 10 °C.

Regarding other climatic variables (heliophany, relative humidity, and wind), the results of the estimations are presented in Table 3. They show that their influence is more significant during the warmer months.

Model evaluation and validation results are the following: residuals mean is zero, standard error of the regression is with a 0.0266; no statistical evidence indicating error autocorrelation was found and the null hypothesis of Jarque–Bera normality test was accepted.⁴

4.3 Predictive Evaluation

In order to assess the model's predictive capability we left out the last year of the sample. The testing period runs from 1 January 2019 to 31 December 2019 and the forecast error for each of the following months was calculated. This year was selected because it was the last year prior to the onset of the pandemic (in Uruguay, a health emergency is declared on March 13, 2020).

It is implemented a rolling-origin evaluation, we successively update the forecasting origin and produce forecast from each new origin. In the testing period, the stability of the estimated parameters is maintained. The prediction at 7 and 14 steps is performed for each week of each month of the year.

Models show relatively good performance according to the predictive performance indicators selected (see Table 4): the Mean Absolute Percentage Error (MAPE) and the Mean Relative Error.

The average MAPE of the year 2019 corresponding to the 7-step forecast is 2.7%, with a deviation of 0.7. The minimum value of the monthly MAPE corresponds to April and the maximum is recorded in October. In the 14-step forecast, the annual average MAPE is 3.1%, with a deviation of 1.2; the minimum and maximum take place in the same months.

⁴ Full estimates are available on request from the authors.

Table 4 Mean absolute relative errors and mean relative errors by prediction horizons and months (2019)

Steps	Jan	Feb	March	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Avg	SE
<i>Mean absolute relative errors</i>														
$h = 7$	3.4	3.2	2.0	1.6	2.0	2.8	2.7	3.0	2.0	3.9	2.1	2.9	2.7	0.7
$h = 14$	3.6	4.0	2.3	1.9	2.3	3.5	3.3	3.7	2.6	4.7	2.4	3.2	3.1	1.2
<i>Mean relative errors</i>														
$h = 7$	-0.64	0.63	-0.03	-0.10	-0.31	0.11	0.14	0.63	0.18	-1.60	-1.05	-0.30	-0.19	0.65
$h = 14$	0.02	0.19	0.14	-0.14	-0.04	-0.04	0.35	0.03	-0.26	-0.42	-0.42	-0.03	-0.05	0.23

Source: Authors calculations

Note: Prediction during 4 weeks for each month
Avg average, SE standard error

The worst performance of the model, assessed by these two indicators, is in the month of October, at both the 7-step and 14-step horizons. For the purpose of analyzing whether these errors correspond to atypical events that occurred in October 2019 or whether the models do not adequately represent the seasonal characteristics corresponding to the early spring months, forecast evaluation was performed for two additional years 2018 and 2020. As a result, both the mean relative error and the mean absolute relative error are larger in magnitude in 2019. In any case, the maximum MAPE for the month of October is approximately twice as high as the month with the lowest MAPE. These results suggest that the bad predictive performance of the model in October is not a regular issue.

Finally, note that those errors were estimated from predictions with exogenous variables already observed. Failing to have this information, uncertainty arising from forecasting these variables must be added.

4.4 Evaluation of the Prediction System During the Health Emergency

In Uruguay, the sanitary emergency was decreed on March 13, 2020. At no time was quarantine mandatory, but social isolation was promoted in different ways. The response of the population to this social isolation was very intense during the second half of March and until May.

During this period, on-site classes were suspended in all education and the service sector immediately reduced its activity. Other sectors of activity, such as construction, stopped their activities for a month and a return to activities was organized by establishing a sanitary protocol accompanied by a follow-up of possible contagions. The industry was paralyzed at the beginning, with workers being sent to unemployment insurance; other sectors, such as agriculture, almost did not stop their activities. The majority of public sector workers gradually developed their activities in teleworking mode, as well as teaching activities at all levels, both public and private.

During the second quarter of 2020, Uruguayan GDP contracted by 13%, and in 2020 economic activity fell by 6%. Employment and the activity rate declined while unemployment increased. The majority of informal workers in Uruguay, which is below 25%, were unable to adhere to the social distancing measures promoted by the government.

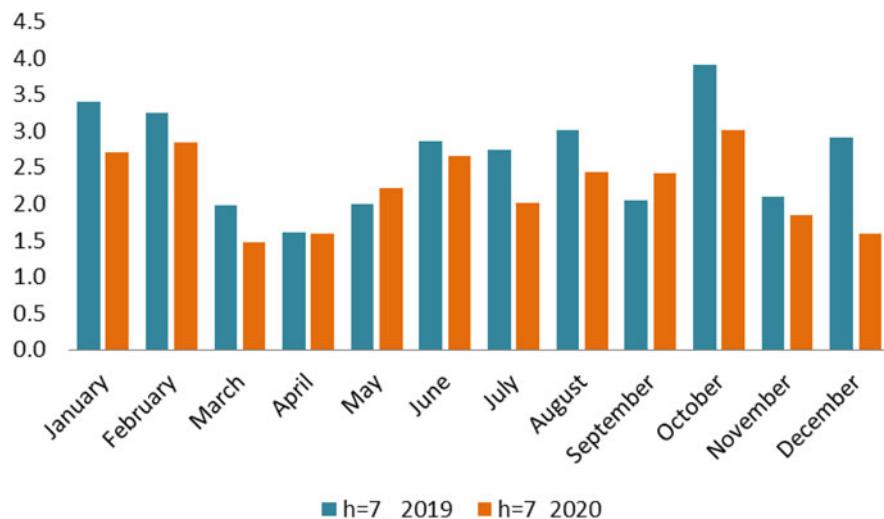
Against this backdrop of profound and unexpected changes in economic activity during 2020, but without the constraints of a mandatory quarantine, we propose to analyze the degree of adaptation of our model. The predictive performance was evaluated, without an adaptation of the models to this break in order to analyze the degree of flexibility they have. Table 5 and Figs. 3 and 4 present the results of the predictive performance assessment through the MAPE and mean relative errors (MPE) for 2020, comparing them with the results during 2019.

Table 5 Mean absolute relative errors (MAPE) and mean relative errors (MRE) by prediction horizons and month

	MAPE				MRE			
	2019		2020		2019		2020	
	h = 7	h = 14						
January	3.40	3.65	2.71	3.25	3.40	3.65	2.71	3.25
February	3.26	4.04	2.84	3.57	3.26	4.04	2.84	3.57
March	1.99	2.30	1.47	1.96	1.99	2.30	1.47	1.96
April	1.60	1.91	1.59	2.16	1.60	1.91	1.59	2.16
May	1.99	2.37	2.21	2.51	1.99	2.37	2.21	2.51
June	2.86	3.50	2.65	3.18	2.86	3.50	2.65	3.18
July	2.75	3.35	2.02	2.41	2.75	3.35	2.02	2.41
August	3.02	3.66	2.43	2.83	3.02	3.66	2.43	2.83
September	2.05	2.60	2.41	2.56	2.05	2.60	2.41	2.56
October	3.90	4.69	3.01	3.12	3.90	4.69	3.01	3.12
November	2.09	2.41	1.85	2.32	2.09	2.41	1.85	2.32
December	2.91	3.26	1.60	2.58	2.91	3.26	1.60	2.58

Note: Prediction during 4 weeks for each month. h—Steps

Source: Authors calculations

**Fig. 3** Mean absolute relative errors at 7 steps for each month (%). 2019–2020

The average absolute relative errors (both of the 7- and 14-step forecasts) are lower in 2020 in 10 of the 12 months of the year. This result suggests the model is flexible enough to adapt to great shocks such as the pandemic meant for Uruguay. The fact that no mandatory quarantines were decreed and that the mobility restrictions that occurred during 2020 were mainly focused between the months of

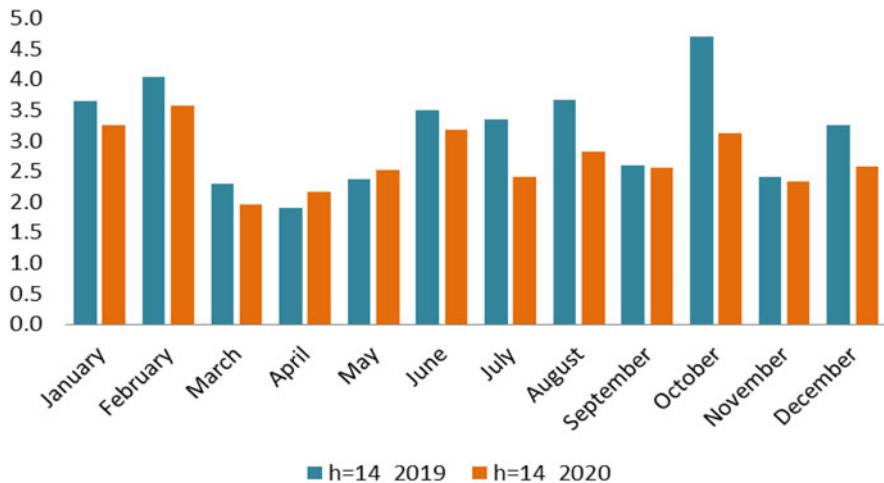


Fig. 4 Mean absolute relative errors at 14 steps for each month (%). 2019–2020

March and May, as well as teleworking in some sectors and the partial return of education to face-to-face work from August 2020, may also explain this result.

In this first stage, no modifications have been made to the specifications of the models that will represent the sanitary emergency during 2020. As a result of the evaluation of the performance of the electric power demand prediction system during the first year of the health emergency, it is concluded that the set of models shows a high flexibility, which allows it to predict the electric power demand during the first year of the COVID-19 pandemic with error levels similar to a non-atypical year.

5 Main Conclusions

We propose a time series non-linear model for daily electricity demand in Uruguay. We took the training sample between January 2010 and December 2018, and a testing sample between January 2019 and December 2019. Subsequently, the predictive performance of the model during the year 2020 (seriously affected by the pandemic of COVID-19) is analyzed.

The method followed has the advantage over computational approaches of providing an interpretable explanation of the variable's behavior in addition to forecasting it.

Our results show, on the one hand, the incidence of special days (calendar effects, holidays) and energy-saving measures. The results show the relevance of capturing these effects with the selected approach, to capture the heterogeneity of the joint impact of public holidays according to the day of the week on which they fall and

their temporal dynamics. Additionally, we represent the association between energy consumption and climate variables (temperature, humidity, winds, and heliophany) with a non-linear model with estimated breaks (estimated by applying Markov switching models). The breaks were identified by considering the division of the sample into warm months (November, December, January, February, March, and April) and cold months (May, June, July, August, September, and October) at 16 °C, 25 °C (in the warm months) and at 10 °C in the cold months.

The estimated coefficients show that the electricity demand function as a function of temperature has been modified concerning [7]. At high temperatures, the demand function increases at a higher rate, and therefore the curve is sharper. In contrast to the previous study for a decade ago, a saturation period is reached. The section of the function corresponding to colder temperatures remains relatively similar. These changes are probably associated with the increased use and availability of cooling equipment by households.

The results of the predictive evaluation show good performance over a 7- and 14-day horizon. Finally, the paper examines the predictive performance of the model during the first year of the health emergency, as a result, it is concluded that the model shows high flexibility, which allows it to predict the electric power demand during the first year of the COVID-19 pandemic with error levels similar to a non-atypical year.

References

1. Bogard, C., George, G., Jenkins, G.M., McLeod, G.: Analyzing a large number of energy time series for utility company. In: Jenkins, G.M., McLeod, G. (eds.) *Case Studies in Time Series Analysis*. Gwilym Jenkins & Partners, Lancaster (1982) chapter 5
2. Bessec, M., Fouquau, J.: The non-linear link between electricity consumption and temperature in Europe: a threshold panel approach. *Energy Econ.* **30**, 2705–2721 (2008)
3. Bunn, D.W., Barnes, E.D.: Economic and operational context of electric load prediction. In: Bunn, D.W., Farmer, E.D. (eds.) *Comparative Models for Electrical Load Forecasting*. Wiley, New York (1985) chapter 1
4. Vivas, E., Allende-Cid, H., Salas, R.: A systematic review of statistical and machine learning methods for electrical power forecasting with reported MAPE score. *Entropy* **22**(12), 1412 (2020)
5. Cancelo, J.R., Espasa, A.: Modeling and forecasting daily series of electricity demand. *Investigaciones Económicas*. **XX**(3), 359–376 (1996)
6. Cancelo, J.R., Espasa, A.: Using high-frequency data and time series models to improve yield management. *Int. J. Serv. Technol. Manag.* **2**, 59–70 (2001)
7. Cancelo, J.R., Espasa, A.: Algunas consideraciones sobre la modelización de series diarias de actividad económica. *Actas de las X Jornadas de Economía Industrial*, 195–201 (1995)
8. Cancelo, J.R., Espasa, A., Grafe, R.: Forecasting the electricity load from one day to one week ahead for the Spanish system operator. *Int. J. Forecasting*. **24**, 588–602 (2008)
9. Darbellay, G.A., Slama, M.: Forecasting the short-term demand for electricity. Do neural networks stand a better chance? *Int. J. Forecasting*. **16**, 71–83 (2000)
10. Engel, R.F., Granger, C.W.J., Rice, J., Weiss, A.: Semi-parametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81**, 310–320 (1986)

11. Smith, M.: Modeling and short-term forecasting of New South Wales electricity system load. *J. Bus. Econ. Statist.* **18**, 465–478 (2000)
12. Soares, L.J., Souza, L.R.: Forecasting electricity demand using generalized long memory. *Int. J. Forecasting.* **22**, 17–28 (2008)
13. Espasa, A.: Modeling daily series of economic activity. In: Proceedings of the Business and Economic Statistics Section, pp. 313–318. American Statistical Association (1993)
14. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. *Int. J. Forecasting.* **22**, 679–688 (2006)
15. Tashman, L.J.: Out of sample test of forecasting accuracy: an analysis and review. *Int. J. Forecasting.* **16**, 437–450 (2000)
16. Hippert, H.S., Bunn, D.W., Souza, R.C.: Large neural networks for electricity load forecasting: are they overidentified? *Int. J. Forecasting.* **21**, 425–434 (2005)
17. Lanzilotta, B., Carlomagno, G., Rosá, T.: Un sistema de predicción y simulación para la demanda de energía eléctrica en Uruguay. Lanzilotta, B. (resp.), Carlomagno, G: Rosá, T. Informe de Proyecto ANII-FMV 2009 (2012)
18. Piggot, J.L.: Short-term forecasting at British Gas. In: Bunn, D.W., Farmer, E.D. (eds.) *Comparative Models for Electrical Load Forecasting*. Wiley, New York (1985)
19. Lanzilotta, B., Collazo, S.R.: Modelos de predicción de demanda de energía eléctrica con datos horarios para Uruguay. *Cuadernos del CIMPAGE.* **18** (2016)
20. Ramanathan, R., Engle, R., Granger, C.W.J., Vahid-Araghi, F., Brace, C.: Short-run forecasts of electricity loads and peaks. *Int. J. Forecasting.* **13**, 161–174 (1997)

Day-Ahead Electricity Load Prediction Based on Calendar Features and Temporal Convolutional Networks



Lucas Richter, Fabian Bauer, Stefan Klaiber, and Peter Bretschneider

Abstract Transmission system operator (**TSO**) have to ensure grid stability economically. This requires highly accurate load forecasts for the transmission grids. The ENTSO-E transparency platform (**ETP**) currently provides a load estimation and a day-ahead load prediction for different **TSO** in Germany. This paper shows a hybrid model architecture of a *feedforward network* based on calendar features to extract the general behaviour of a time-series and a *temporal convolutional network* to extract the relations between short-historical and future time-series values. This research shows a significant improvement of the current day-ahead load forecast and additionally a model robustness while training with a non-optimal data set.

Keywords Grid load · Prediction methods · Artificial neural networks · Temporal convolutional networks

Acronyms

TSO	Transmission system operator
ETP	ENTSO-E transparency platform
X_{load-da}	ENTSO-E transparency platform—day-ahead forecast
CAL	Feedforward neural network based on calendar features
CTCN	Temporal convolutional network with calendar features
TCN	Temporal convolutional network

L. Richter (✉) · S. Klaiber · P. Bretschneider
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Ilmenau,
Germany
e-mail: Lucas.Richter@iosb-ast.fraunhofer.de; Stefan.Klaiber@iosb-ast.fraunhofer.de;
Peter.Bretschneider@iosb-ast.fraunhofer.de

F. Bauer
Ilmenau University of Technology, Energy Usage Optimization Group, Ilmenau, Germany
e-mail: Fabian.Bauer@tu-ilmenau.de

LSTM	Long short-term memory
ARIMA	Autoregressive integrated moving average
ARIMAX	Autoregressive integrated moving average with explanatory variable
x_{load}	ENTSO-E transparency platform—actual electrical load
x_{cal}	Calendar features
y_{cal}	Output of calendar feature network
hol	Holidays
$shol$	School holidays
$bday$	Bridge days

1 Introduction

TSO must keep power generation and consumption in balance at all times to ensure a stable and reliable energy supply. Therefore, an accurate and efficient grid load forecast is needed to plan reliable and cost-optimal grid operation, taking into account the feed-ins of conventional and renewable generation plants. This requires a precise knowledge about the current grid state. The latter one has been integrated in different load forecast models to predict future values of electricity load time-series [8–10]. Load time-series are generally characterised by short-term periodicities and depend on the individual consumption behaviour of consumers. This behaviour is mainly affected by calendar effects concerning different weekdays, holidays and daytime hours [6]. This paper presents a two-step model approach to improve the day-ahead electricity load forecast [1]. First, a normalisation function is applied on a multi-year load time-series to make the values comparable to each other. After this pre-processing step, calendar information is used inside a neural network to extract a generalised behaviour considering date, hour, holidays and school holidays. In the second step, the output of the latter neural network is combined with the actual load inside a temporal convolutional network to adjust to the real values.

2 Data

ETP is an online data platform for European electricity system data [1]. It was established in early 2015 under EU Regulation 543/2013 [1] to support market participants, reduce insider trading and make this electricity data available to various actors. EU Members States are engaged to publish essential information related to electricity load, generation, transmission and balancing [1]. The data has generally a temporal resolution of 15 min and can be officially downloaded for customer usage.

This study uses the ENTSO-E transparency platform—actual electrical load (x_{load}) (see Sect. 2.1) and the ENTSO-E transparency platform—day-ahead forecast ($x_{load-da}$) (see Sect. 2.2) time-series of a German TSO in the period 2015 to

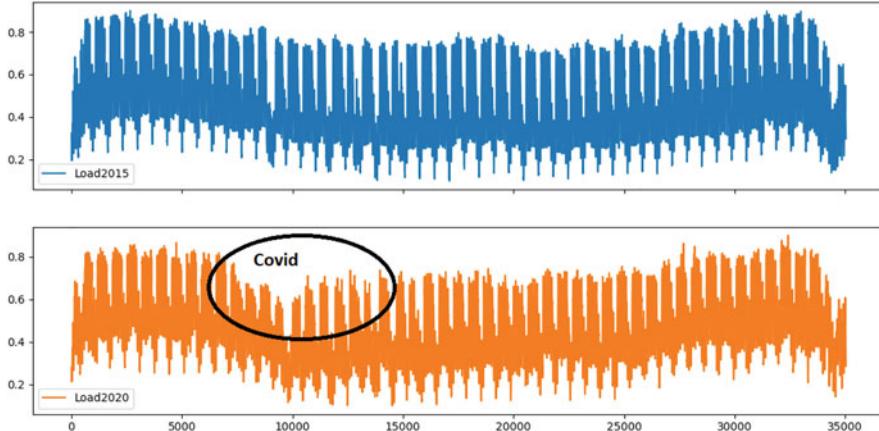


Fig. 1 Electricity consumption of 2015 and 2020 scaled between 0.1 and 0.9

2020 [1]. In Fig. 1, the time-series of 2015 and 2020 are exemplarily illustrated to get an impression of the Covid-19 effect. The data has a temporal resolution of 15 min. Additionally Calendar features (x_{cal}) (see Sect. 2.3) are used. The values of x_{load} are normalised to its minimum and maximum of the last year between 0.1 and 0.9 to allow some variations around its historical boundaries.

2.1 Electricity Load

The electricity load is the total electricity feed-in of all known power stations and imports into the grid minus all exports and the consumption of pumped-storage power plants [2]. In real time the electricity load has to be estimated due to missing or false values of different stations. In consequence, the load estimation comes with uncertainties which have to be considered in the model creation process (see Sect. 4). After a certain period of time, all stations have to correct their missing or false values and publish them to the TSO. These values sum up to the billed time-series.

2.2 Electricity Load Prediction

Current time-series values are mostly used to tune the forecast. In consequence, load prediction models mainly depend and are trained on its estimated values to provide day-ahead predictions for the TSO several times a day. The uncertainty of the measurement estimation has to be considered properly by the prediction model (see Sect. 3).

Table 1 Monday holiday from 23.00 to 23.45 o’clock in January

Jan	...	Dec	Mon	...	Sun	hour ₀	...	hour ₂₃	min ₀	...	min ₄₅	hol	shol	bday
1	0	0	1	0	0	0	0	1	1	0	0	1	0	0
1	0	0	1	0	0	0	0	1	0	1	0	1	0	0
1	0	0	1	0	0	0	0	1	0	1	0	1	0	0
1	0	0	1	0	0	1	0	0	0	0	1	1	0	0

2.3 Calendar Data

Electricity load time-series depend strongly on calendar features (see Sect. 3). To extract this dependence considering Holidays (*hol*), Bridge days (*bday*) and School holidays (*shol*), it is strongly recommended to use these calendar features. In the first step, calendar information is taken from a Python API [3] and [4]. Based on this data, *bday* are extracted. To make the data usable for a neural network, the format and values have to be adapted. x_{cal} has the same temporal resolution as x_{load} and is divided into one-hot encodings (see Table 1) of month, weekday, hour, minutes, *hol* and *shol* per German province.

3 Data Analysis

Figure 1 shows the first and the last year of x_{load} representing its annual cycle with and without the Covid-19 effect. Both years contain local minima during Easter and Christmas time. Concerning this effect, it will be interesting to analyse the robustness of the model (see Sect. 4.3) by choosing a training set which is strongly affected by Covid-19 (see Sect. 5). Additionally, the load time-series possess a strong annual cycle with a summer minimum and a winter maximum due to the fact that human behaviour depends strongly on daylight hours. During winter times, humans switch the lights earlier on and also devices such as television and stove are used more.

The weekday-behaviour considering the entire daytimes is shown in Fig. 2 by applying a principal component analysis on x_{load} . While workdays Monday to Thursday possess quite similar characteristics, Friday differ to the latter ones and the weekend-days Saturday and Sunday are very distinctive to all other days. This effect is also detected on *hol*, *shol* and *bday* in an additional dimension and is considered by using one-hot encodings (see Table 1) as calendar features inside the prediction model (see Sect. 4.3).

As already mentioned, the estimated values possess an inherent uncertainty. Figure 3 depicts the values of x_{load} and the y_{cal} (see Sect. 4.1) for a certain time interval in July 2019. Besides a strong correlation between both time-series, a varying deviation is detectable, especially for minimal and maximal values. This

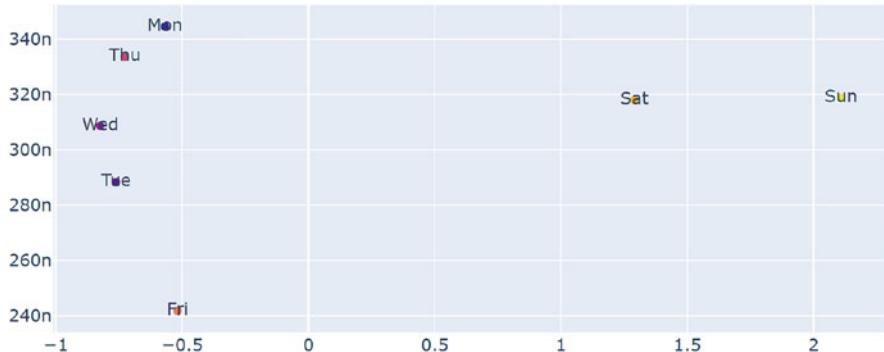


Fig. 2 PCA scatterplot of different weekdays Monday to Sunday

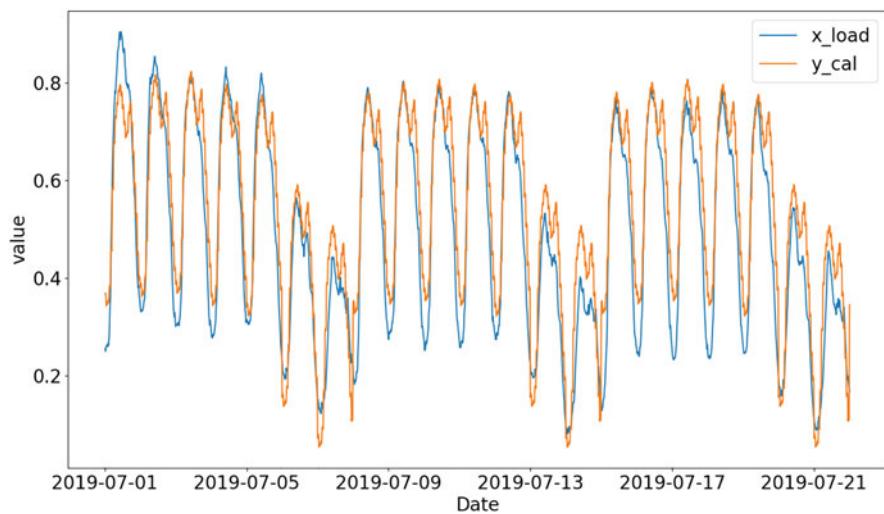


Fig. 3 Example time-series of x_{load} vs. y_{cal} (see Sect. 4.3)

deviation seems not to change randomly, but approximately stays constant over a proper time interval considering local minima.

To handle this deviation inside the model and to make it robust against this changing behaviour, a new score concerning the standard deviation (sd) of the mean difference between x_{load} and y_{cal} over n-days is introduced:

$$sd_{mean_{diff}}(n) = sd \left(\sum_{i=0}^k \left| \frac{x_{load}[i] - y_{cal}[i]}{x_{load}[i]} \right| \right) \quad (1)$$

where

n: number of days per time interval

i: the ith time interval of n-days of the time-series

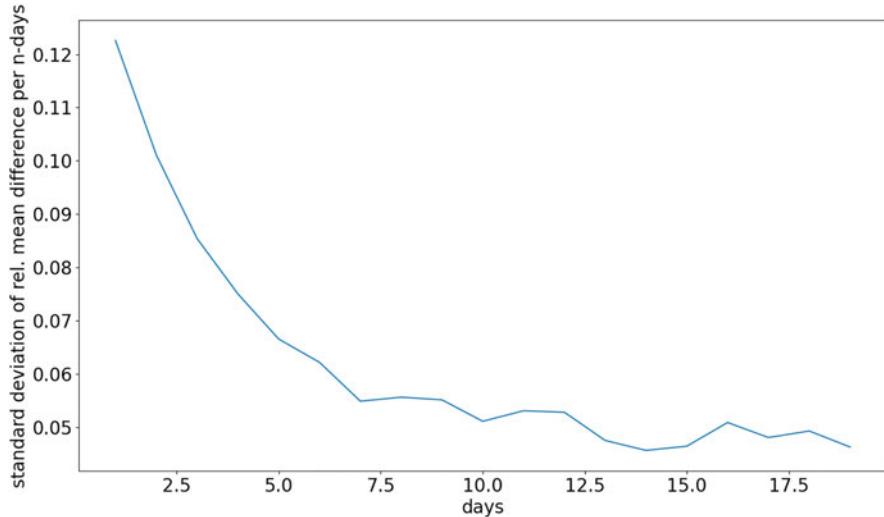


Fig. 4 Standard deviation of rel. mean difference per n-days

Figure 4 depicts Eq. (1) applied for different number of n-days. While $sd_{mean_{diff}}$ decreases within the first 7 days, it stays nearly constant afterwards. This can be explained by adjusting ratios between x_{load} and y_{cal} over a specific time interval. Based on this analysis, 7 days as historical input data for x_{load} and y_{cal} are chosen inside the prediction model to prevent overfitting due to the increasing parameter space with more input data while training data is limited.

4 Model Architecture

Time-series possess a trend, a seasonal and a cyclic component which are considered in the classical time-series decomposition methods. Electricity load time-series generally possess a seasonal and a trend component. Classical prediction methods are linear stochastic time-series models such as Autoregressive integrated moving average ([ARIMA](#)), and, if exogenous variables are included, there are various extensions such as Autoregressive integrated moving average with explanatory variable ([ARIMAX](#)). These models are well suited if the structure of the data is well understood and a sufficient amount of data is available. For problems with a large number of variables and an increasing influence of non-linear or unknown dependencies, machine learning methods promise a significant advantage over existing methods [5, 12].

4.1 Feedforward Network Based on Calendar Features

Calendar information was already used to predict electricity prices for short and long terms [11], to extract the general behaviour of a time-series and to use the latter one as stabilisation feature inside a neural network. Rementol et al. constructed a neural architecture with calendar information and renewable energy generation as input variables which are processed separately. In the first step, calendar information is featured into embedding layers (Emb) which can be seen as state vectors of the this exogenous variable. In the next step, Emb and renewable energy generation are concatenated and flattened to process them in several dense layers to fit finally to the output variable (see Fig. 5). This simple architecture was able to outperform conventional Long short-term memory (**LSTM**) prediction models clearly and can be applied to various time-series which mostly depend on calendar effects (see Sect. 3) and less on other exogenous variables. In this paper, a calendar network Feedforward neural network based on calendar features (**CAL**) is separately trained to fit x_{load} . Finally the y_{cal} can be used as an exogenous variable inside a Temporal convolutional network (**TCN**).

4.2 Temporal Convolutional Network

TCNs were originally developed to identify and time segment patterns in signals [7]. **TCNs** are able to capture and model long-range dependencies and relationships in historical observations. The **TCN** architecture uses a hierarchy of temporal convolution filters, with the special feature that input sequences of any length can be mapped to any length output sequence (see Fig. 6). The convolution operations are performed using residual blocks to extract all information from the historical observations to achieve higher forecast accuracy. This also offers the possibility of

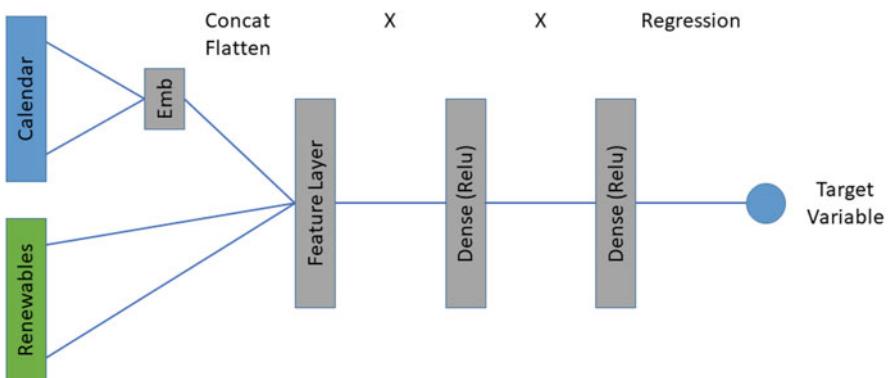


Fig. 5 Hybrid model to predict electricity price [11]

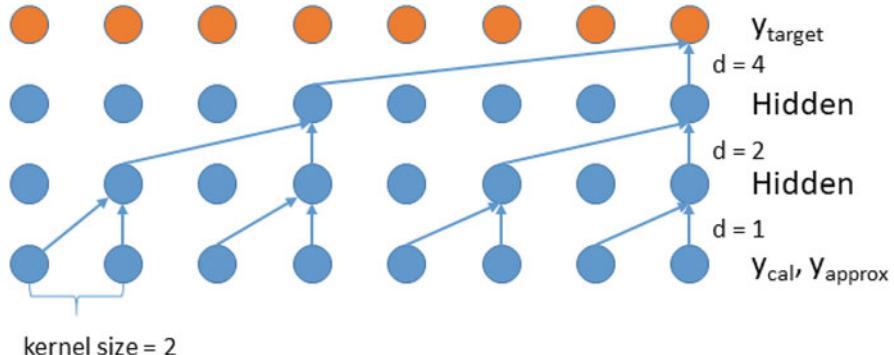


Fig. 6 Scheme of a temporal convolutional network with d for dilation rate

including exogenous variables in addition to the past observations when training the model. In the research, it was shown that TCNs achieve a higher prediction quality in time-series prediction compared to competing LSTM-based recurrent neural networks and reduce the configuration effort as well as the training time.

4.3 Hybrid Model

This paper now combines two neural architectures **CAL** and a **TCN** (see Sects. 4.1 and 4.2) to a Temporal convolutional network with calendar features (**CTCN**) to predict the day-ahead electricity load: (i) For generalisation of a model based on calendar information, **CAL** uses x_{cal} as input in conjunction with multiple dense layer to fit to x_{load} . (ii) **CTCN** uses y_{cal} as an exogenous variable as well as seven historical days of x_{load} inside a **TCN** (see Fig. 6). In short, a **TCN** is able to consider a very long history of input values by using causal dilated convolutional layers with a dropout inside a residual block wherein the input is added to the output of each layer. This architecture helps to combine elder and more recent values of x_{load} and y_{cal} to predict day-ahead values of x_{load} as well as to consider inputs of different lengths. Additionally, this architecture should address the uncertainty of the load estimation of the past n-days in comparison to its mean values y_{cal} (see Fig. 7).

5 Training and Evaluation Set

The data set is divided into three parts: (i) The years from 2015 to 2018 are used to train a generalised **CAL**. This is very important due to the fact that Covid-19

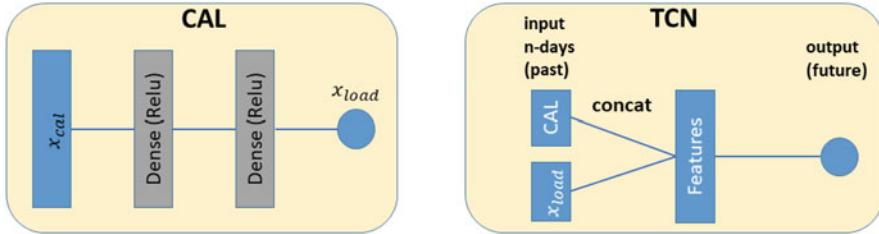


Fig. 7 Temporal convolutional network with calendar features

Table 2 RMSE and MAPE for $x_{load-da}$ and CTCN

	RMSE in GW	MAPE in %
$x_{load-da}$	1.046	4.50
CTCN	0.552	2.88

changed the characteristics of x_{load} . (ii) The time-series of x_{load} and y_{cal} from mid of 2019 to mid of 2020 are used to train the CTCN. (iii) The trained model is evaluated to the time-series from mid of 2020 to end of 2020. In the last two data sets, the impact of the Covid-19 pandemic can be seen with the temporary shutdown of the German economy. This circumstance can be considered as a stress test to the CTCN.

6 Results

The CTCN clearly outperforms $x_{load-da}$ in terms of the *root mean squared error* (RMSE) and *mean absolute percentage error* (MAPE) (see Table 2, Fig. 8), despite the model being trained with a data set of changepoints due to the Covid-19 effect. While the deviations of $x_{load-da}$ depend more on its absolute ones, CTCN fits very well to x_{load} . In addition, the CTCN has a lower variance and adjusts better to external influences, such as the temporary shutdown of the German economy during 2020. This can be explained by the optimised model architecture and the fact that CTCN can handle long- and short-term dependencies of recent time-series.

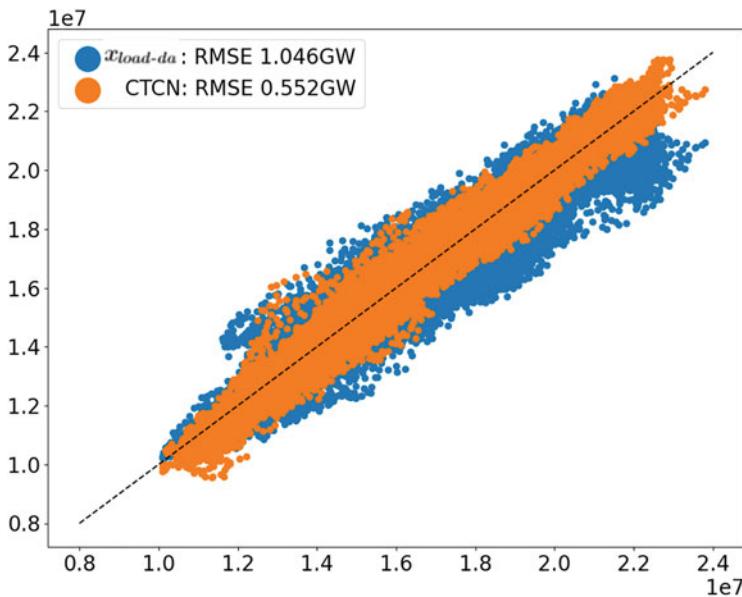


Fig. 8 $x_{load-da}$ and CTCN plotted against x_{load}

7 Conclusion

This paper shows a hybrid network architecture of **CAL** and **TCN** that greatly improves the day-ahead electricity load forecast and its robustness during the Covid-19 crisis compared to $x_{load-da}$. Furthermore, the electrical load behaviour depends strongly on human behaviour in addition to the given variables. In the next step of further research, weather data will be included in addition to calendar features to further improve the results.

Acknowledgments The work was financially supported by BMWi in Germany (Bundesministerium für Bildung und Forschung) under the project “Bauhaus.MobilityLab”.

References

1. <https://transparency.entsoe.eu/> (visited on 03/10/2021)
2. <https://www.tennet.eu/electricity-market/transparencypages/transparency-germany/network-figures/system-load-systemload-forecast/> (visited on 05/20/2021)
3. <https://pypi.org/project/holidays/> (visited on 03/10/2021)
4. <https://www.schulferien.org/deutschland/ferien/> (visited on 03/10/2021)
5. Deb, C., et al.: A review on time series forecasting techniques for building energy consumption. Renew. Sustain. Energy Rev. **74**, 902–924 (2017). <https://doi.org/10.1016/j.rser.2017.02.085>

6. Klaiber, S.: Analyse, Identifikation und Prognose preisbeeinflusster elektrischer Lastzeitreihen. PhD thesis. Technische Universität Ilmenau, 2020
7. Lea, C., et al.: Temporal Convolutional Networks for Action Segmentation and Detection (2016). arXiv:1611.05267 [cs.CV]
8. Mourshed, M., Kuster, C., Rezgui, Y.: Electrical load forecasting models: A critical systematic review. *Sustain. Citi. Soc.* **35**, 257–270 (2017). <https://doi.org/10.1016/j.scs.2017.08.009>
9. Nti, I.K., et al.: Electricity load forecasting: a systematic review. *J. Electr. Syst. Inf. Tech.* (2020). <https://doi.org/10.1186/s43067-020-00021-8>
10. Nyarko-Boateng, O., Nti, I. K., Teimeh, M., Adekoya, A. F.: Electricity load forecasting: a systematic review. *J. Electr. Syst. Inf. Tech.* **7**, Article number: 13 (2020). <https://doi.org/10.1186/s43067-020-00021-8>
11. Ramentol, E., Schirra, F., Wagner, A.: Short- and Longterm Forecasting of Electricity Prices Using Embedding of Calendar Information in Neural Networks (2020). arXiv:2007.13530 [stat.AP]
12. Raza, M. Q., Khosravi, A.: A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renew. Sustain. Energy Rev.* **50**, 1352–1372 (2015). <https://doi.org/10.1016/j.rser.2015.04.065>

Network Security Situation Awareness Forecasting Based on Neural Networks



Richard Staňa , Patrik Pekarčík , Andrej Gajdoš , and Pavol Sokol 

Abstract The increasing number of cybersecurity threats affects the security situation of organisations. The maintenance of the operational picture of the organisation, which integrates all relevant information for selecting appropriate countermeasures, becomes a vital role for organisations. In this paper, we focus on network security situation awareness forecasting. The paper aims to answer two questions—the influence of loss function in neural networks on network security situation awareness forecasting and a comparison of statistical methods and neural networks in network security situation awareness forecasting. For this purpose, we used two-time series representing cybersecurity alerts collected by system Warden. This paper shows an analysis according to which the MAE and MASE loss functions give better results than MSE. Also, we can state that neural networks are more accurate for network security situation awareness forecasting.

Keywords Cybersecurity · Network security · Network security situation awareness · Forecasting · Time series

1 Introduction

Nowadays, the number of new cybersecurity threats and cybersecurity incidents is on the rise. The main goal of organisations' security teams is to prevent cybersecurity incidents or minimise their impact. For example, the organisations' network administrators or security teams may prevent these incidents by disallowing the specific network protocols or updating systems to address security vulnerabilities. In this respect, we observe a trend of transition from reactive activities to proactive activities [1].

R. Staňa · P. Pekarčík · A. Gajdoš · P. Sokol (✉)

Pavol Jozef Šafárik University in Košice, Faculty of Science, Košice, Slovakia
e-mail: richard.stana@upjs.sk; patrik.pekarckik@upjs.sk; andrey.gajdos@upjs.sk;
pavol.sokol@upjs.sk

An important element in ensuring the proactive activities of the organisation is the maintenance of the operational picture of the organisation, which integrates all relevant information for identifying attacks and selecting appropriate countermeasures [2]. This operational picture can be defined as network security situation awareness (NSSA). Bass et al. introduced the origin, concept, target and characteristics of NSSA in more detail in [3].

According to a different perception of an object, NSSA can be divided into the network security situation assessment and network security situation forecasting [4]. Forecasting the security situation is an essential part of the NSSA and allows anticipating cybersecurity attacks and cybersecurity threats. It provides network administrators and security teams time to make adequate decisions on their next steps. Overall, this allows better analysing security threats and management of cybersecurity incidents.

Researchers have proposed and used various approaches to forecast network security situation awareness in recent years, such as statistical methods, game theory methods or neural networks. In the following section, we focus on state of the art in statistical methods and neural networks in more detail. At the same time, there are some problems in these methods, such as the loss of network data information caused by situation assessment and the low forecasting accuracy of the neural network model used for the NSSA forecasting [5]. To improve the accuracy of the NSSA forecasting, this paper aims to (I) analyse the influence of loss function in neural networks on the NSSA forecasting and (II) compare statistical methods and neural networks in NSSA forecasting.

This paper is based on previous research [6, 7]. Within this paper, we assume the fact that in the NSSA forecasting, there is a lot of time series forecasting with neural networks that look like naive forecasting with drift [8]. Definition of the mean absolute scaled error (MASE) shows that it compares forecasting with naive forecasting. Using MASE as a loss function, we can “punish” neural network when its forecasting looks like naive forecasting with drift.

This paper is organised into six sections. Section 2 reviews state of the art in network security awareness forecasting. Section 3 is devoted to research methodology and outlines the dataset and methods used for the analysis. Section 4 states the experimental evaluation. Section 5 discusses the results. The last section concludes the paper and discusses the challenges for future research.

2 Related Works

This section overviews papers and research groups’ activities related to network security situation awareness forecasting. This section is divided into two parts: the statistical time series approach and the neural network approach. Most of the papers focus on the detection of attacks rather than a prediction of attacks or NSSA forecasting [9].

In the field of the NSSA, the autoregressive integrated moving average (ARIMA) models are a very frequently used approach. Examples of research work using these forecasting methods are [10–12]. Above-mentioned ARIMA models are often used in combination with other models. For example, ARIMA models are used with the Bayesian networks to predict future cyberattack (malware, malicious URL and malicious e-mail) occurrences [13]. Another example is a combination of ARIMA models and grey-box models. In the paper [14], the authors responded to the disadvantages of the separate usage of these models. ARIMA models require strict inputs, and the grey-box models do not consider the system's randomness. This combination is used in the extreme-value phenomenon analysis [15]. An exciting combination of methods for forecasting purposes is used in several research papers. The analysis of the fitting of ARMA and GARMA models to the cyberattack process is an objective of paper [16].

Neural networks are commonly used in the field of time series prediction in cybersecurity. There are a lot of papers that use older types of smaller feedforward networks or wavelet neural networks trained by backpropagation or genetics algorithm (and its variants) to forecast network security situations (e.g. [17, 18]). On the other hand, modern approaches like recurrent neural networks (GRU, LSTM) were used in the paper [19] for forecasting the network security situation. In the paper [20], authors compare the ARIMA approach, LSTM and GRU neural networks for cyberattack prediction. This prediction is based on the combination of time series and external signals. Another research paper [21] predicts time series based on data collected by the honeypot. For this prediction, the authors used a bidirectional LSTM neural network. Several research groups have been working with recurrent neural networks like LSTM and GRU to predict cyberattacks based on time series created from industrial data [22–24].

3 Methodology

3.1 Dataset

Our research used a dataset collected and preprocessed by a Warden system [25]. This system was created for sharing cybersecurity alerts between hosts connected to this sharing system. Security alerts are stored in a descriptive data model using a key-value JSON extensible structure called IDEA (Intrusion Detection Extensible Alert) format [26]. Primary data sources for the Warden system may include honeypots, intrusion detection systems, network flow probes, system log records and other sensors and data sources. The data used in the research are collected from real operation in the computer networks of the Czech National Research and Education Network and other Czech commercial organisations.

Security alerts in the IDEA format contain several mandatory fields (form, ID, detect time, category) [26] and many optional fields. The fields we used in this

experiment are the category of security alert, source and destination IP addresses, source and destination ports, network protocol and detection time. The Warden system collected the data we used in this research for 1 year (from 2017-12-11 to 2018-12-11). Our dataset contains approximately one billion security alerts from various data sources (mainly honeypots).

In our research, we used time series with 30-min time period. We deal with the creation of time series and selection of periods in more detail in the papers [6, 27]. Also, we used two selected time series, such as time series representing the total number of alerts and time series representing alerts related to the services running on port 445/TCP. These time series are representatives of two categories of time series for the area of NSSA forecasting (well-predictable time series and unpredictable time series) [8].

3.2 Method Description

There is a wide range of quantitative forecasting methods, and their usage often depends on the specific disciplines, the nature of data or specific purposes. Our research compared the accuracy of three different loss functions' mean absolute error (MAE), mean squared error (MSE) and MASE, by implementing five different neural networks to obtain predictions. After that, we compare the best methods with usually used statistical methods for time series forecasting. From neural networks, we employ five types of neural networks: dense network, LSTM, GRU, convolutional neural networks, and encoder-decoder networks. From the statistical method, we choose the following: ARIMA models, exponential smoothing models (state-space models), the naive approach (with drift), and combination (average) of ARIMA and exponential smoothing models. A complete description of the mentioned architectures can be found below.

3.3 Neural Networks

There is a lot of work done in time series forecasting with neural networks, for example, in the field of stock prediction [28–30], traffic prediction [31, 32], etc. We developed five multilayer neural networks, most of them were inspired by Brownlee [33], and similar networks were previously used in our work [8].

In the following text, we provide a description of the architectures (abbreviations, used later and denoting individual architectures are in parentheses):

- Dense network (DN)—four dense layers (1024, 512, 256, 128 units, activation relu) and one dense layer (1 unit, activation linear)
- Long short-term memory (LSTM)—three LSTM layers (256, 256, 256 units, default parameters) and one dense layer (1 unit, activation linear)

- Gated recurrent unit (GRU)—three GRU layers (256, 256, 256 units), one SimpleRNN layer (128 units) and one dense layer (1 unit, activation linear)
- 1D convolution (Conv1D)—three Conv1D layers (256, 256, 256 filters, 3, 3, 3 kernel size, activation relu, padding same), one dense layer (64 unit, activation relu) and one dense layer (1 unit, activation linear)
- Encoder-decoder LSTM (e1d1)—one LSTM encoder layer (512 units encoder, return state True), RepeatVector layer, one LSTM decoder layer (512 units encoder, return state True) and TimeDistributed (1 dense unit, activation linear)

3.4 Statistical Methods

The choice of statistical methods for this research is based on our previous research activity [8, 27]. ARIMA and exponential smoothing (ETS) [34] are the most commonly used statistical models in the modelling and time series prediction classes.

ARIMA models represent a generalisation of the ARMA model class, including a wide range of non-stationary series. These models ensure the stationarity of the time series by a finite number of differentiations. ARMA models are a combination of automatic regression (AR) and moving average (MA) [35].

The ETS class provides additional access to time series modelling and forecasting. Prediction using models in this class is characterised by a weighted combination of older observations with new ones. The new observations have a relatively higher weight compared to the older observations. Exponential smoothing reflects that weights decrease exponentially with the age of the observations. On the one hand, ETS models are based on trend descriptions and seasonality in the data. On the other hand, ARIMA models aim to describe autocorrelations in the book [34] data.

In the research, we also use the naive methods [35, 36] as a benchmark for statistical methods. These methods can process large datasets and, at the same time, do not have high computational demands. We also added a combination (average) of ARIMA and ETS methods to the experiments to compare standard methods with their diversity. The idea of averaging or increasing is currently nothing new [43].

4 Experiment Evaluation

We consider only one-step ahead predictions.

For forecast accuracy evaluation, we employ two commonly used metrics—MASE used [37] and MAE.

MASE is a preferred metric as it is less sensitive to outliers, more easily interpreted and less variable on small samples. MASE is defined as [34]:

$$\text{MASE} = \text{mean}(|q_j|) \quad (1)$$

where q_j is:

$$q_j = \frac{e_j}{\frac{1}{T-1} \sum_{i=2}^T |y_i - y_{i-1}|}, \quad (2)$$

where y_i represents observed value and T is the length of time series.

For a better view of accuracy in both time series, we take into account also MAE, which is defined as follows [34]:

$$\text{MAE} = \text{mean}(|e_j|) \quad (3)$$

In both cases, e_j is forecast error, i.e. the difference between an observed value and its forecast.

Both time series we used consist of 17,473 values. We did not use the first 27 and last 14 values because there were primarily zeros or missing values. Due to missing values between 15,550 and 15,601 in the whole dataset, we split the dataset into three parts. In the first part, there were values between 28 and 15,549 (15,522 values), the second part included values between 15,602 and 16,601 (1000 values), and the last part contained values between 16,602 and 17,458 (857 values).

The first and the second parts were used for training neural networks. The third part was used for testing. During neural network training, we employed a window containing 384 values (8 days) for every model and time series. We trained all five neural networks in 40 epochs. From every type of network, we trained four instances with Adam optimiser, two with fixed learning rate (lr) to 0.001 and two with decreasing lr from 0.001 to 0.0001 decreasing by two when testing loss did not decrease in four epochs. If testing loss of a particular neural network had a decreasing tendency (at the end of training), we trained it for more than 20 epochs. After training, we choose the best model from four instances based on MASE metrics.

It is essential to prepare dataset before training a neural network. To our time series, we applied standardisation (subtraction of mean and division by standard deviation—mean and standard deviation were calculated using the first part of the dataset). For neural network training, we employed three different loss functions, MAE, MSE and MASE. For both, we used standard implementation, which is in TensorFlow pages. In our implementation of the MASE loss function, we first describe the predicted value and real value as inputs. This was done because we wanted to calculate MASE according to unscaled data. Then MASE was implemented as described by Eqs. 1 and 2.

GPU NVidia GTX 1080 and 1060, Keras and TensorFlow [38] version 2.4 were used to train neural networks. The batch size was set to 128. To make model comparisons easier, we used the tool Weights & Biases [39]. In total, we trained more than 120 neural networks (five networks described in Methodology x three loss functions x four instances x two-time series).

Additionally, we compared predictions based on neural networks with forecasts based on statistical approaches described in the previous section. Due to

missing data in the dataset, long training time when using the extensive dataset and weak impact of older data on statistical methods, we used only values from the second part of the dataset for fitting statistical methods (as described in the previous section). We used values from the third part of the dataset to test their forecasting accuracy. On the other side, we train neural networks on both (first and second) parts of the dataset because, generally, more data means better results from neural networks. With more data, neural networks can find more patterns in data, generalise them better and get better results, even with older data.

The methods were evaluated according to principles and implementations presented in our previous work [6, 8, 27]. For our research, we used R functions from one of the most common R-packages for time series predictions called *forecast* [40]. This package contains valuable features when working with large datasets or potentially in real-time prediction. In addition, these functions are used to adjust ARIMA and ETS model classes automatically. These functions are designed to automatically select the best model from the considered class under the given conditions, for example, considering the information criterion [40].

In the next part of our research, we focused on two ways of adapting statistical models: the classical model and the “rolling window” approach. With the classical model, we kept the entire training datasets. Step by step, we added one more observation to the training set in each round of evaluation. In the second method, we focused on the “rolling window” approach (“one in, one out” approach). As in the previous method, we added one new observation from the test set to the training set. The difference was that in each round of evaluation, we removed the oldest observation from the training set.

Seasonality was not taken into account due to its minimal impact on forecasting performance as shown in paper [27] where we used the same dataset.

At this place, it is essential to note that we have modified the denominator in MASE. The reason was the difference between the size of the training dataset in the case of statistical methods and neural network models. The aim was to achieve comparability of forecasting for both approaches. For this reason, the denominator calculations in the Eq. 2 were based on the 1000 training values used to adjust the statistical models.

5 Results and Discussion

In this section, we compared the results which were obtained according to the description in the previous section. Because MAE was used as a metric, we present some statistical information about the dataset:

- Time series of the total number of alerts: minimum 22, maximum 155,818 and mean 34,594.25
- Time series of the alerts related to the services running on port 445/TCP: minimum 0, maximum 16,168 and mean 5972.56

Table 1 MASE and MAE comparison for three loss functions on all neural network forecasting for the total number of alerts on testing dataset. Every bold number is the best result for the actual neural network from three loss functions

Test metrics	Loss function	DN	LSTM	GRU	e1d1	Conv1D
MASE	MAE	0.9950	0.9213	0.9286	0.9166	0.9254
	MSE	1.0245	0.9442	0.9550	0.9430	0.9567
	MASE	1.0147	0.9192	0.9352	0.9178	0.9362
MAE	MAE	2645.9203	2449.9389	2469.3886	2437.3081	2460.6580
	MSE	2724.2048	2510.7505	2539.4539	2507.7080	2543.9238
	MASE	2698.3200	2444.1826	2486.8879	2440.6539	2489.3861

Table 2 MASE and MAE comparison for three loss functions on all neural networks forecasting port 445/TCP on the testing dataset. Every bold number is the best result for the actual neural network from three loss functions

Test metrics	Loss function	DN	LSTM	GRU	e1d1	Conv1D
MASE	MAE	0.6972	0.6633	0.6307	0.6408	0.7080
	MSE	0.7215	0.7118	0.7321	0.7038	0.8201
	MASE	0.7020	0.6617	0.6210	0.6582	0.7208
MAE	MAE	1186.1808	1128.4426	1072.9371	1090.1418	1204.4519
	MSE	1227.4236	1210.9351	1245.5377	1197.3586	1395.2064
	MASE	1200.1366	1125.6894	1056.5102	1119.8305	1226.3334

Tables 1 and 2 show the results of comparison of neural network models for selected time series (the total number of alerts –Table 1– and security alerts related to the service running on network port 445/TCP, Table 2). In the analysis, we used MASE and MAE metrics to evaluate the results. Each neural network was used with a specific loss function. According to the results shown in the given tables, it can be stated that the MSE loss function shows the worst results in all investigated neural networks. The MAE loss function achieves the best results or approaches them. The MASE loss function implemented by us is comparable to the MAE loss function.

At the same time, we analysed statistical methods in the research. Their comparison according to MASE and MAE metric may be seen in Table 3. As may be seen from the results, the value of the MASE metric for NSSA forecasting in the time series of the number of cybersecurity alerts is bigger than 1. It means that the given forecasting method is worse than the average naive forecast. The time series of alerts associated with services running on port 445/TCP has another result. As may be seen from the table, the used models have a MASE metric value below 1. Exponential smoothing appears to be the best method in both cases. These results confirm the findings from previous research [6]. Similar time series were used in the current article, but with a different period.

Finally, we compared the best statistical method (exponential smoothing) and the best neural network (e1d1 MAE, respectively GRU MASE). As may be seen from Table 4, in both cases, neural networks have better MASE and MAE metrics.

Table 3 Performance comparison of statistical models. Notes: A, ARIMA model; E, exponential smoothing; N, naive model; AE, ARIMA + exponential smoothing (average); w, rolling window

Time series	The total number of alerts		Port 445/TCP	
Test metrics	MASE	MAE	MASE	MAE
A	1.0536	2801.7281	0.7950	1352.5694
Aw	1.0569	2810.3664	0.8046	1368.8168
E	1.0319	2744.0770	0.7661	1303.3641
Ew	1.0374	2758.7118	0.7741	1317.0408
AE	1.0411	2768.4691	0.7661	1303.3641
AEw	1.0450	2778.8436	0.7741	1317.0408
N	1.1851	3151.4376	0.9910	1686.0467
Nw	1.1854	3152.1949	0.9912	1686.3547

Table 4 Comparison of best models based on neural networks and statistical models on both time series. Notes: NN, neural network; E, exponential smoothing; AE, ARIMA + exponential smoothing (average)

Time series	The total number of alerts		Port 445/TCP	
Best NN/statistical model	e1d1 MAE	E	GRU MASE	AE
MASE	0.9166	1.0319	0.6210	0.7661
MAE	2437.3081	2744.0770	1056.5102	1303.3641

Figure 1 shows one-step predictions with the best statistical approach and neural network approach for the total number of alerts that are similar to the naive forecasting with drift. In the same way, in Fig. 2 are predictions for port 445/TCP that are way more accurate in the case of the neural network approach.

In addition to the above, we also compared the accurate predictions of the best methods from statistical and neural network approaches. For this purpose, we used the Diebold-Mariano test [41] and its implementation in the R package *multDM* [42]. If two forecasts have the same accuracy, it represents the null hypothesis (H0). On the other hand, the alternative hypothesis (H1) had the setting w = “less” (the first forecast is less accurate than the second forecast). Since we leave a 5% uncertainty rate, the p-value to confirm the null hypothesis (H0) should be higher than 0.05.

In our evaluation, we compared all combinations of the statistical and neural methods used in this paper for two cases—“count of all alerts” and “port 445/TCP.” We tested the situation that forecasts have the same accuracy (null hypothesis) against the situation that a forecast based on the statistical method is less accurate than the forecast based on the neural network method (alternative hypothesis). For example, in the Diebold-Mariano test, the p-value of comparison of ARIMA and dense with MAE loss is 0.003633, which is less than 0.05. In this case, an alternative hypothesis was accepted (the null hypothesis was rejected). It means that for the time series marked “count of all alerts,” the forecast based on the statistical method (ARIMA) is less accurate than the forecast based on the neural method (dense with MAE loss).



Fig. 1 Graphical comparison of best models based on neural networks (e1d1 with MAE loss) and statistical models (exponential smoothing) for the total number of alerts

Results of Diebold-Mariano test of statistical methods and neural network methods for one-step forecasting of the “count of all alerts” time series are shown in Tables 5 and 6. In these tables, in the first column, there are neural network methods and the other columns contain p-value for the Diebold-Mariano test for a couple of the statistic and neural methods. The forecasts based on statistical methods are less accurate than a forecast based on neural network methods in almost all cases. In two cases, the combination of the forecasting methods has the same accuracy. These cases are highlighted (bold font) in Table 6.

Results of Diebold-Mariano test of statistical methods and neural network methods for one-step forecasting of the port 445/TCP are shown in Tables 7 and 8. In these tables, in the first column, there are neural network methods, and the other columns contain p-value for the Diebold-Mariano test for a couple of the statistic and neural methods. The forecasts based on statistical methods are less accurate than a forecast based on neural network methods in all cases.

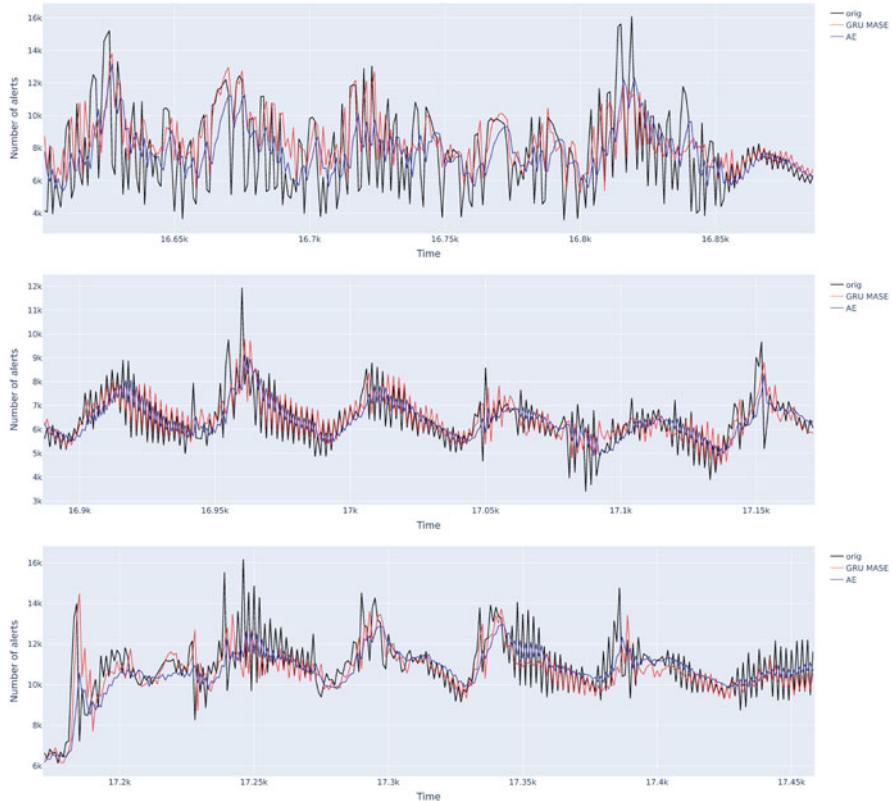


Fig. 2 Graphical comparison of best models based on neural networks (Gru with MASE loss) and statistical models (combination of ARIMA and exponential smoothing) for port 445/TCP

These calculations confirm our results expressed by MAE and MASE measures described above.

6 Conclusion and Future Works

Within the paper, we focused on NSSA forecasting. For this purpose, we used two-time series (the total number of alerts and alerts related to the services running on port 445/TCP). These time series represent two categories of time series for the area of NSSA forecasting (well-predictable time series and unpredictable time series) [8]. This paper aimed to analyse the impact of loss function on the accuracy of NSSA forecasting based on neural networks. According to the obtained results, we found that the loss function has an effect and the MAE and MASE loss function give comparable results. At the same time, we compared the best neural networks

Table 5 Results of Diebold-Mariano test for one-step forecasting of the count of all alerts (Part I). Notes: A, ARIMA; AE, ARIMA + exponential smoothing (average); w, rolling window

	A	Aw	AE	AEw
cnn MAE	3.331×10^{-7}	1.004×10^{-7}	2.355×10^{-6}	8.033×10^{-7}
cnn MASE	2.87×10^{-5}	1.379×10^{-5}	0.0001242	6.342×10^{-5}
cnn MSE	3.794×10^{-7}	1.338×10^{-7}	2.427×10^{-6}	9.099×10^{-7}
dense MAE	0.003633	0.002393	0.01623	0.009665
dense MASE	0.009035	0.005458	0.04873	0.02757
dense MSE	0.0055	0.003181	0.02018	0.0121
e1d1 MAE	1.509×10^{-7}	6.281×10^{-8}	9.142×10^{-7}	3.617×10^{-7}
e1d1 MASE	4.465×10^{-8}	1.45×10^{-8}	3.009×10^{-7}	1.011×10^{-7}
e1d1 MSE	2.301×10^{-8}	7.747×10^{-9}	1.509×10^{-7}	5.148×10^{-8}
gru MAE	2.303×10^{-7}	9.031×10^{-8}	1.721×10^{-6}	6.213×10^{-7}
gru MASE	6.323×10^{-8}	2.475×10^{-8}	5.847×10^{-7}	1.976×10^{-7}
gru MSE	7.88×10^{-8}	3.027×10^{-8}	5.738×10^{-7}	2.036×10^{-7}
lstm MAE	7.155×10^{-8}	2.748×10^{-8}	4.925×10^{-7}	1.827×10^{-7}
lstm MASE	2.239×10^{-8}	8.27×10^{-9}	1.679×10^{-7}	5.845×10^{-8}
lstm MSE	2.393×10^{-8}	1.022×10^{-8}	1.887×10^{-7}	6.967×10^{-8}

Table 6 Results of Diebold-Mariano test for one-step forecasting of the count of all alerts (Part II). The cases where the p -value is greater than 0.05 are highlighted (bold font). Notes: E, exponential smoothing; N, naive method; w, rolling window

	E	Ew	N	Nw
cnn MAE	1.124×10^{-5}	4.263×10^{-6}	1.644×10^{-6}	1.629×10^{-6}
cnn MASE	0.0003741	0.0001893	5.355×10^{-6}	5.323×10^{-6}
cnn MSE	1.059×10^{-5}	4.168×10^{-6}	1.061×10^{-6}	1.048×10^{-6}
dense MAE	0.04673	0.02514	5.046×10^{-5}	4.973×10^{-5}
dense MASE	0.1431	0.07876	0.0003169	0.0003122
dense MSE	0.04932	0.02911	0.0001261	0.0001246
e1d1 MAE	3.876×10^{-6}	1.491×10^{-6}	6.964×10^{-7}	6.9×10^{-7}
e1d1 MASE	1.435×10^{-6}	5.127×10^{-7}	5.35×10^{-7}	5.297×10^{-7}
e1d1 MSE	7.057×10^{-7}	2.531×10^{-7}	3.218×10^{-7}	3.183×10^{-7}
gru MAE	8.847×10^{-6}	3.151×10^{-6}	2.218×10^{-6}	2.19×10^{-6}
gru MASE	3.693×10^{-6}	1.202×10^{-6}	2.027×10^{-6}	1.993×10^{-6}
gru MSE	3.028×10^{-6}	1.051×10^{-6}	1.006×10^{-6}	9.934×10^{-7}
lstm MAE	2.402×10^{-6}	8.777×10^{-7}	7.2×10^{-7}	7.123×10^{-7}
lstm MASE	8.959×10^{-7}	3.085×10^{-7}	4.627×10^{-7}	4.575×10^{-7}
lstm MSE	1.059×10^{-6}	3.633×10^{-7}	6.262×10^{-7}	6.194×10^{-7}

and the best statistical methods. According to the MASE and MAE metrics, we can state that neural networks are more accurate for NSSA forecasting. As part of future works, we would like to focus on NSSA forecasting on time series created from other security alerts (obtained by a platform other than the Warden system).

Table 7 Results of Diebold-Mariano test for one-step forecasting of the port 445/TCP (Part I). Notes: A, ARIMA; AE, ARIMA + exponential smoothing (average); w, rolling window

	A	Aw	AE	AEw
cnn MAE	6.793×10^{-07}	7.495×10^{-08}	0.0004185	8.341×10^{-05}
cnn MASE	6.963×10^{-06}	5.428×10^{-07}	0.00301	0.0005788
cnn MSE	4.434×10^{-06}	7.059×10^{-07}	0.003503	0.0008352
dense MAE	2.788×10^{-13}	3.497×10^{-14}	1.722×10^{-08}	9.922×10^{-10}
dense MASE	6.11×10^{-12}	3.202×10^{-13}	1.73×10^{-07}	7.755×10^{-09}
dense MSE	5.697×10^{-14}	1.508×10^{-15}	3.285×10^{-09}	5.847×10^{-11}
e1d1 MAE	2.055×10^{-13}	6.905×10^{-15}	8.538×10^{-09}	7.248×10^{-10}
e1d1 MASE	1.624×10^{-14}	2.595×10^{-16}	8.628×10^{-10}	4.84×10^{-11}
e1d1 MSE	8.302×10^{-15}	1.028×10^{-15}	2.572×10^{-10}	3.39×10^{-11}
gru MAE	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	6.626×10^{-13}	2.885×10^{-14}
gru MASE	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
gru MSE	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.712×10^{-12}	1.165×10^{-13}
lstm MAE	4.18×10^{-11}	1.474×10^{-12}	2.657×10^{-07}	2.998×10^{-08}
lstm MASE	1.245×10^{-10}	2.13×10^{-11}	6.396×10^{-07}	1.287×10^{-07}
lstm MSE	8.967×10^{-11}	1.966×10^{-11}	6.882×10^{-07}	1.398×10^{-07}

Table 8 Results of Diebold-Mariano test for one-step forecasting of the port 445/TCP (Part II). Notes: E, exponential smoothing; N, naive method; w, rolling window

	E	Ew	N	Nw
cnn MAE	0.0004185	8.341×10^{-05}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
cnn MASE	0.00301	0.0005788	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
dense MAE	1.722×10^{-08}	9.922×10^{-10}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
dense MASE	1.73×10^{-07}	7.755×10^{-09}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
dense MSE	3.285×10^{-09}	5.847×10^{-11}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
e1d1 MAE	8.538×10^{-09}	7.248×10^{-10}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
e1d1 MASE	8.628×10^{-10}	4.84×10^{-11}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
e1d1 MSE	2.572×10^{-10}	3.39×10^{-11}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
gru MAE	6.626×10^{-13}	2.885×10^{-14}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
gru MASE	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
gru MSE	1.712×10^{-12}	1.165×10^{-13}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
lstm MAE	2.657×10^{-07}	2.998×10^{-08}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
lstm MASE	6.396×10^{-07}	1.287×10^{-07}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
lstm MSE	6.882×10^{-07}	1.398×10^{-07}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$

Acknowledgments This research is funded by the VVGS projects under contract No. VVGS-PF-2020-1423, VVGS-PF-2020-1427 and VVGS-PF-2021-1792 and Slovak Research and Development Agency project under contract No. APVV-17-0561.

References

- Cho, J.H., Sharma, D.P., Alavizadeh, H., Yoon, S., Ben-Asher, N., Moore, T.J., Kim, D.S., Lim, H., Nelson, F.F.: Toward proactive, adaptive defense: a survey on moving target defense. *IEEE Commun. Surv. Tutor* **22**(1), 709–745 (2020)
- Carle, G., Dressler, F., Kemmerer, R.A., Koenig, H., Kruege, C., Laskov, P.: Network attack detection and defense. In: *Manifesto of the Dagstuhl Perspectives Workshop*, pp. 2–6 (2008)
- Bass, T., et al.: Multisensor data fusion for next generation distributed intrusion detection systems. In: *Proceedings of the IRIS National Symposium on Sensor and Data Fusion*, vol. 24, pp. 24–27. Citeseer (1999)
- Jiang, Y., Li, C.h., Yu, L.s., Bao, B.: On network security situation prediction based on RBF neural network. In: *2017 36th Chinese Control Conference (CCC)*, pp. 4060–4063. IEEE, Piscataway (2017)
- Shang, L., Zhao, W., Zhang, J., Fu, Q., Zhao, Q., Yang, Y.: Network security situation prediction based on long short-term memory network. In: *20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 1–4. IEEE, Piscataway (2019)
- Husák, M., Bartoš, V., Sokol, P., Gajdoš, A.: Predictive methods in cyber defense: current experience and research challenges. *Futur. Gener. Comput. Syst.* **115**, 517–530 (2021)
- Sokol, P., Gajdoš, A.: Prediction of attacks against honeynet based on time series modeling. In: *Proceedings of the Computational Methods in Systems and Software*, pp. 360–371. Springer, Berlin (2017)
- Sokol, P., Staňa, R., Gajdoš, A., Pekarčík, P.: Network security situation awareness forecasting based on statistical approach and neural networks. *Log. J. IGPL* (In press)
- Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., Liu, M.: Cloudy with a chance of breach: forecasting cyber security incidents. In: *24th USENIX Security Symposium*, vol. 15, pp. 1009–1024 (2015)
- Okutan, A., Werner, G., McConky, K., Yang, S.J.: Poster: cyber attack prediction of threats from unconventional resources (capture). In: *24th ACM Conference on Computer and Communications Security*, pp. 2563–2565 (2017)
- Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*, pp. 1–3. ACM, New York (2017)
- Werner, G., Yang, S., McConky, K.: Leveraging intra-day temporal variations to predict daily cyberattack activity. In: *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 58–63. IEEE, Piscataway (2018)
- Werner, G., Okutan, A., Yang, S., McConky, K.: Forecasting cyberattacks as time series with different aggregation granularity. In: *2018 IEEE International Symposium on Technologies for Homeland Security (HST)*, pp. 1–7. IEEE, Piscataway (2018)
- Qi, Y., Shang, W., He, X.: A combined prediction method of industrial internet security situation based on time series. In: *Proceedings of the 2019 the 9th International Conference on Communication and Network Security*, pp. 84–91 (2019)
- Zhan, Z., Xu, M., Xu, S.: Predicting cyber attack rates with extreme values. *IEEE Trans. Inf. Forensics Secur.* **10**(8), 1666–1677 (2015)
- Pillai, T.R., Palaniappan, S., Abdullah, A., Imran, H.M.: Predictive modeling for intrusions in communication systems using garma and arma models. In: *2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW)*, pp. 1–6. IEEE, Piscataway (2015)
- Zhang, H., Huang, Q., Li, F., Zhu, J.: A network security situation prediction model based on wavelet neural network with optimized parameters. *Digit. Commun. Netw.* **2**(3), 139–144 (2016)
- He, F., Zhang, Y., Liu, D., Dong, Y., Liu, C., Wu, C.: Mixed wavelet-based neural network model for cyber security situation prediction using modwt and hurst exponent analysis. In: *International Conference on Network and System Security*, pp. 99–111. Springer, Berlin (2017)

19. Feng, W., Wu, Y., Fan, Y.: A new method for the prediction of network security situations based on recurrent neural network with gated recurrent unit. *Int. J. Intell. Comput. Cybernet.* **13**(1), 25–39 (2020)
20. Goyal, P., Hossain, K., et al.: Discovering signals from web sources to predict cyber attacks (2018). Preprint. arXiv:1806.03342
21. Fang, X., Xu, M., Xu, S., Zhao, P.: A deep learning framework for predicting cyber attacks rates. *EURASIP J. Inform. Secur.* **2019**(1), 1–11 (2019)
22. Lavrova, D., Zegzhda, D., Yarmak, A.: Using gru neural network for cyber-attack detection in automated process control systems. In: 2019 IEEE International Black Sea Conference on Communications and Networking, pp. 1–3. IEEE, Piscataway (2019)
23. Filonov, P., Kitashov, F., Lavrentyev, A.: RNN-based early cyber-attack detection for the tennessee eastman process (2017). Preprint. arXiv:1709.02232
24. Filonov, P., Lavrentyev, A., Vorontsov, A.: Multivariate industrial time series with cyber-attack simulation: fault detection using an lstm-based predictive data model (2016). Preprint. arXiv:1612.06676
25. Kacha, P., Kostenec, M., Kropacova, A.: Warden 3: security event exchange redesign. In: 19th International Conference on Computers: Recent Advances in Computer Science (2015)
26. Kacha, P.: Idea: security event taxonomy mapping. In: 18th International Conference on Circuits, Systems, Communications and Computers (2014)
27. Pekářík, P., Gajdoš, A., Sokol, P.: Forecasting security alerts based on time series. In: International Conference on Hybrid Artificial Intelligence Systems, pp. 546–557. Springer, Berlin (2020)
28. Pang, X., Zhou, Y., Wang, P., Lin, W., Chang, V.: An innovative neural network approach for stock market prediction. *J. Supercomput.* **76**(3), 2098–2118 (2020)
29. Chen, K., Zhou, Y., Dai, F.: A LSTM-based method for stock returns prediction: a case study of China stock market. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2823–2824. IEEE, Piscataway (2015)
30. Kim, T., Kim, H.Y.: Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. *PloS One* **14**(2), 1–23 (2019)
31. Fu, R., Zhang, Z., Li, L.: Using LSTM and GRU neural network methods for traffic flow prediction. In: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 324–328. IEEE, Piscataway (2016)
32. Zhao, Z., Chen, W., Wu, X., Chen, P.C., Liu, J.: LSTM network: a deep learning approach for short-term traffic forecast. *IET Intell. Trans. Syst.* **11**(2), 68–75 (2017)
33. Lim, B., Zohren, S.: Time-series forecasting with deep learning: a survey. *Philos. Trans. R. Soc. A* **379**(2194), 20200209 (2021)
34. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. (2018)
35. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken (2015)
36. Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting. Springer, Berlin (2016)
37. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**(4), 679–688 (2006)
38. Abadi, M., Agarwal, A., et al.: Tensorflow: large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015). <https://www.tensorflow.org>
39. Biewald, L.: Experiment tracking with weights and biases (2020) Software available from <https://www.wandb.com>
40. Hyndman, R.J., Khandakar, Y., et al.: Automatic time series for forecasting: the forecast package for R. Number 6. Monash University, Department of Econometrics and Business Statistics (2007)
41. Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. *J. Bus. Eco. Stat.* **20**(1), 134–144 (2002)

42. Clements, M.P., Hendry, D.F.: A Companion to Economic Forecasting. John Wiley & Sons (2008)
43. Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P.: Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Commun. Surv. Tutor.* **21**(1), 640–660 (2018)

Part IV

Advanced Applications in Time Series

Analysis

Modeling Covid-19 Contagion Dynamics: Time-Series Analysis Across Different Countries and Subperiods



Zorica Mladenović, Lenka Glavaš, and Pavle Mladenović

Abstract This study offers two sets of empirical results to model the daily COVID-19 contagion time series. The Markov-switching models with ARMA structure are implemented assuming that time-series dependence is nonlinear, whereas regimes are data-driven. The first set of results consists of models estimated for the following European countries: Italy, Germany, the United Kingdom, and Russia during the first epidemic wave. The second set of results deals with modeling time series for Italy over the second and the third epidemic waves. Given the empirical findings reached, we have distinguished among several regimes during the epidemic wave. The persistence of time series over each regime is also discussed.

Keywords COVID-19 · Count data · Markov-switching models · ARMA models · Persistence

1 Introduction

The Markov-switching (MS) models were introduced and developed by Hamilton [9–11]. They have been implemented extensively in different areas of economics because economic time series are often subject to shifts from one type of behavior to another and back again, where the causes of the regime shifts are unobservable. The application of MS models in epidemiology was first advocated by Strat and Carrat [20] and further elaborated in [15]. Most empirical studies are based on the two-state model to detect epidemic and non-epidemic regimes. Flexible Bayesian version of the MS model [15] enables identification of several phases within the epidemic that

Z. Mladenović (✉)
University of Belgrade, Faculty of Economics, Belgrade, Serbia
e-mail: zorica.mladenovic@ekof.bg.ac.rs

L. Glavaš · P. Mladenović
University of Belgrade, Faculty of Mathematics, Belgrade, Serbia
e-mail: lenka@matf.bg.ac.rs; paja@matf.bg.ac.rs

would have clear implications (pre-epidemic, epidemic growth, epidemic plateau, epidemic decline, and post-epidemic). The same idea can be followed in time-series analysis of daily new cases of COVID-19 contagion.

This study examines the daily dynamics of COVID-19 new cases in those European countries that have faced the largest numbers of positive cases. The MS models are used as the main framework. The goal of the study is twofold. First, we compare modeling results for the period that contains the first peak of the epidemic for the following countries: Italy, Germany, the United Kingdom, and Russia. Second, for the case of Italy, we provide results of MS modeling for the subperiods of the second and third peaks. Data are taken from <https://ourworldindata.org/covid-cases>.

Our models have an ARMA structure. However, the model's key characteristic is that constant error variability and ARMA coefficients change across different regimes. We associate identified phases with the estimated level of persistence. For stationary time series, persistence is represented by the infinite sum of weights in its linear representation [17]. Persistence is commonly measured according to the values of parameters in corresponding autoregressive (AR) representation. As a baseline case, if the autoregressive parameter in AR (1) model is statistically not different from 1, then we are dealing with high persistence or a unit-root presence. The lower the value of the autoregressive parameter, the smaller the persistence and higher the probability that undertaken measures would be effective. If the autoregressive parameter is greater than 1, then the time series is explosive, which implies an extremely high level of unpredictability. If the AR model of order greater than 1 is chosen and estimated, then the level of persistence may be assessed by the absolute values of roots of the corresponding characteristic equation.

Time series considered in this study represent count time series. Statistical literature offers a variety of models for such time-series data. Among them, Poisson autoregression and log-linear Poisson autoregression models defined in [8] and [7] have been frequently used. The log-linear Poisson autoregression model has been employed by Agosto and Giudici [1], Turasie [22], and Agosto et al. [2] to capture the spread of the COVID-19 virus. However, our previous empirical findings in [18] clearly suggest that this model estimated for several countries does not exhibit satisfactory statistical properties because it fails several diagnostic tests.

The MS approach has been employed in modeling COVID-19 data for several countries, for example, [3, 16], and [19], but not as often as standard epidemiological models.

The paper has the following structure. In Sect. 2, the MS models are briefly described. Section 3 contains modeling results of COVID-19 new cases for the selected countries (Italy, Germany, the United Kingdom, and Russia) over the subperiod of the first peak. Findings for Italy over proceeding peaks are given by Sect. 3. The empirical results for Italy related to the second and the third peaks are given in Sect. 4. Conclusions are offered in Sect. 5.

2 The Markov-Switching Time-Series Models

In this study, we shall use the approach of modeling time series with changes in a regime that will be referred to as Markov-switching models. These models were introduced by Hamilton [9, 10, 12]. For the presentation of the theory, see also [11].

Suppose that the observations x_1, x_2, \dots, x_n are realizations of the random process X_1, X_2, \dots, X_n . For every $t \in \{1, 2, \dots, n\}$, an unobserved random variable S_t is associated with X_t and determines the state of the process, that is, determines the conditional distribution of X_t given S_t .

We suppose here that random variable S_t takes values from some finite set $\{1, 2, \dots, m\}$ and the probability that the random variable S_t takes some particular value j depends on the past only through the value S_{t-1} , that is,

$$P\{S_t = j | S_{t-1} = i, S_{t-2} = i_{t-2}, S_{t-3} = i_{t-3}, \dots\} = P\{S_t = j | S_{t-1} = i\}. \quad (1)$$

Let us denote

$$p_{ij} = P\{S_t = j | S_{t-1} = i\}, \quad i, j \in \{1, 2, \dots, m\}, \quad (2)$$

that is, p_{ij} is the probability that state i will be followed by state j . The sequence (S_t) is then referred to as an m -state homogeneous Markov chain with stationary transition probabilities $\{p_{ij}\}_{i,j=1,2,\dots,m}$. Obviously, for every $i \in \{1, 2, \dots, m\}$, the following equality holds:

$$p_{i1} + p_{i2} + \dots + p_{im} = 1. \quad (3)$$

The general autoregressive moving average models of order (K,L) for the sequence (X_t) , notation MS-ARMA(K,L), are given by

$$X_t = c_{S_t} + \sum_{k=1}^K \phi_{kS_t} X_{t-k} + \sum_{l=1}^L \theta_{lS_t} \varepsilon_{t-l} + \varepsilon_{S_t}, \quad (4)$$

where the constants c_{S_t} , ϕ_{kS_t} and θ_{lS_t} depend on the regime, and ε_i , $i \in \{1, \dots, m\}$ are independent random variables with the variance that depend only on the regime, that is, for $S_t = i$ we have $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. The special case of the general model (4) is the first-order autoregressive model (AR) that is given by

$$X_t = c_{S_t} + \phi_{S_t} X_{t-1} + \varepsilon_{S_t}, \quad (5)$$

where the constants c_{S_t} and ϕ_{S_t} depend on the regime and ε 's are independent zero mean and normally distributed random variables with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. If $S_t = i$, the conditional distribution of X_t is assumed to be given by the density

$$f(x_t | S_t = i, X_{t-1}, X_{t-2}, \dots, \alpha). \quad (6)$$

In the case of the first-order AR model, the components of α to be estimated are c_i , ϕ_i , σ_i^2 where $1 \leq i \leq m$ and the transition probabilities p_{ij} , $i, j \in \{1, \dots, m\}$. In the case of the general ARMA model, the components of α are

$$c_i, \phi_{1i}, \dots, \phi_{Ki}, \theta_{1i}, \dots, \theta_{Li}, \sigma_i^2, \quad (7)$$

where $1 \leq i \leq m$ and the transition probabilities p_{ij} , $i, j \in \{1, \dots, m\}$.

Several methods were considered from the literature for obtaining the estimates of the unknown parameters of the models considered here. The number of states (regimes) is also subject to estimation. In practical work, it is chosen according to the maximum value of the sample likelihood function or equivalently to the minimum values of information criteria. Additionally, we consider several diagnostic statistics when determining the number of states.

The likelihood of the Markov-switching model can be evaluated using the filtering procedure in [10] followed by the smoothing algorithm in [13] and [11], Ch. 22. The log-likelihood, which is a function of the components of α , can then be maximized subject to the constraint that the probabilities lie between 0 and 1 and sum to unity. Most of the literature suggests using the EM algorithm of [5], following [10]. All estimations are performed using OxMetrics 8; see [6]. The feasible nonlinear programming approach of [14] is used to maximize the log-likelihood of the MS model. As emphasized in [6], it converges more quickly and is more robust than other available techniques.

The essential part of MS modeling consists of investigating whether it outperforms ARMA linear model with constant parameters. Defining a linearity test is a demanding task because parameters are not identified under the null hypothesis that there is no difference between the MS and ARMA. Thus, the likelihood-ratio test does not have the standard asymptotic χ^2 distribution, [21].

We provide results of a test for linearity available in OxMetrics 8, derived from the likelihood-ratio statistic between the estimated and implied linear models. Two p-values are reported. The first one is based on the standard χ^2 distribution. The second one is the so-called approximate upper bound for the significance level of the likelihood-ratio statistic; see [4].

3 Empirical Results for Early Dynamics

In this section, modeling results are described separately for each of the following countries: Italy, Germany, the United Kingdom, and Russia. At the end of this section, summary findings are reported.

3.1 Italy (Sample: February 22–May 31, 2020)

The four-state MS-AR (1) model fits well the dynamics of COVID-19 new cases in Italy. Identified regime three is detected as the most extreme episode. It covers 31% of the sample. This regime is the most persistent one, given the estimate of autoregressive parameter 0.995. Variability is estimated to be the highest during regime three. This episode has been preceded and shortly interrupted by regime one that lasted on average 2 days. Regime one is associated with strong upward and downward trends in the data, estimated to take approximately 15% of the sample. The persistence is estimated to be low. Identified regime four is also characterized by high persistence, 0.942. It marks the subperiods at the beginning and the end of the sample covering half of it. This regime describes time intervals during which the number of new cases was relatively low compared to other regimes. Nevertheless, the finding of the high persistence at the end of the sample indicates that the epidemic episode was far from being over at the beginning of June 2020. Regime two is the shortest one because it includes only 3 days spread randomly over the sample. These dates are characterized by very high jumps and low persistence.

Of all estimated transition probabilities, the highest one, 0.95, is the probability of staying in regime four of relatively high persistence.

The estimated model is provided in Table 1. Together with the estimated parameters $\hat{\phi}_i$ and \hat{c}_i , we provide the p -value in the brackets. With estimated parameters $\hat{\sigma}_i$, the corresponding standard errors are reported. The associated transition probabilities are given in Table 2. The estimated model performs statistically well, as shown by diagnostic tests in Table 3. Figure 1 captures actual and estimated data along with detected regimes.

Table 1 $m = 4$: Constant c_i , autoregressive ϕ_i , and st. error σ_i parameters

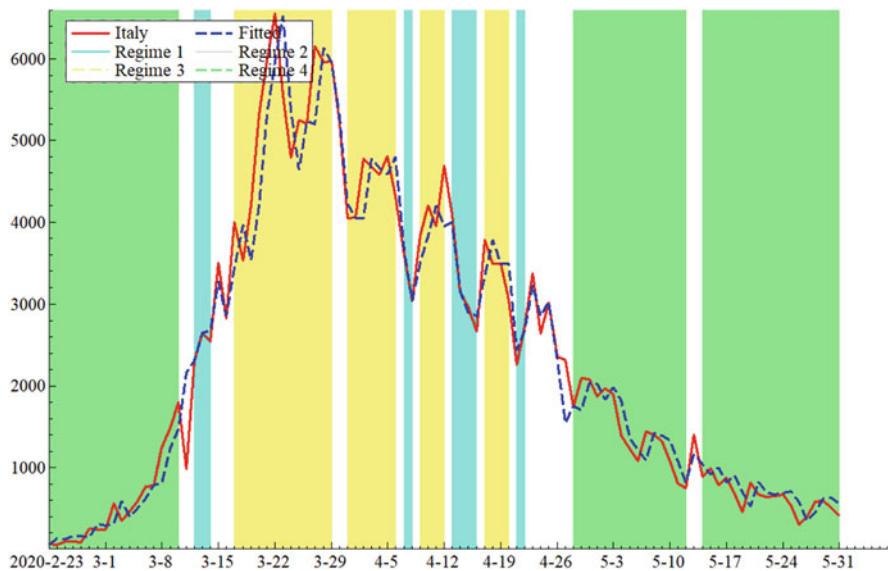
State	$\hat{\phi}_i$	\hat{c}_i	$\hat{\sigma}_i$
1	0.296 (0.00)	2749 (0.00)	83.33 (19.37)
2	0.741 (0.00)	1567 (0.00)	829.42 (250.3)
3	0.995 (0.00)	3453 (0.00)	536.91 (75.49)
4	0.942 (0.00)	1159 (0.00)	187.07 (20.68)

Table 2 $m = 4$: Transition probabilities

Probability	$State_{1,t}$	$State_{2,t}$	$State_{3,t}$	$State_{4,t}$
$State_{1,t+1}$	0.45	0.25	0.21	0.00
$State_{2,t+1}$	0.08	0.00	0.00	0.06
$State_{3,t+1}$	0.47	0.00	0.79	0.00
$State_{4,t+1}$	0.00	0.74	0.00	0.94

Table 3 Specification tests

Linearity	Normality	ARCH 1-2	Autocorr. Q
χ^2_{14}	χ^2_2	F(2,78)	χ^2_{10}
63.5 (0.00, 0.00)	1.5 (0.48)	0.1 (0.93)	8.0 (0.63)

**Fig. 1** Italy: actual and estimated numbers of daily COVID-19 new cases, and estimated regimes

3.2 Germany (Sample: January 28–May 31, 2020)

The four-state MS-AR (1) model also performs statistically well for German daily data of new COVID-19 cases. These states, along with the actual and estimated new COVID-19 cases, are provided by Fig. 2. The most extreme period is detected as regime three. During regime three, the variability is the highest, but the autoregressive parameter is estimated to be the lowest (0.366). Therefore, high persistence was not a key feature of regime three. Relatively low persistence indicates that although being at a high level, the number of new cases was not unpredictable. Regime three is followed by regime one, during which the number of cases started to fall. This downward trend in the data is estimated by the autoregressive parameter 0.565 (moderate persistence), whereas variability remained high.

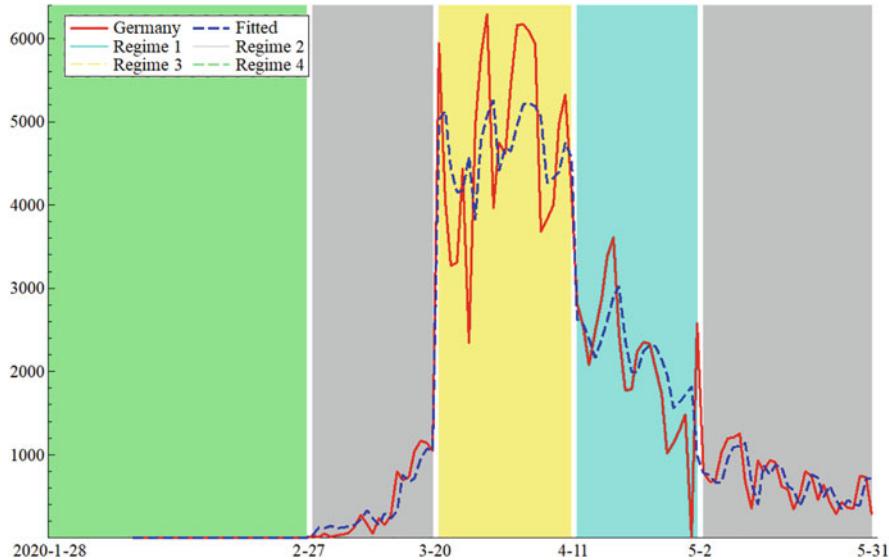


Fig. 2 Germany: actual and estimated numbers of daily COVID-19 new cases, and estimated regimes

Table 4 $m = 4$: Constant c_i , autoregressive ϕ_i , and st. error σ_i parameters

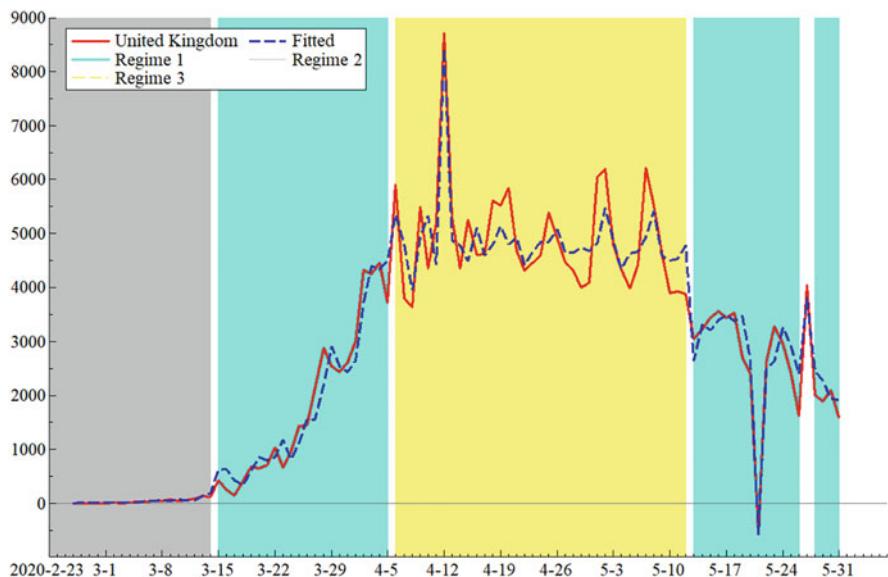
State	$\hat{\phi}_i$	\hat{c}_i	$\hat{\sigma}_i$
1	0.565 (0.01)	2268.3 (0.00)	716.81 (135)
2	0.823 (0.00)	601.6 (0.00)	204.87 (20.53)
3	0.366 (0.07)	4655.5 (0.00)	1004.46 (171.2)
4	0.999 (0.00)	579.2 (0.00)	1.13 (0.14)

This regime one covers approximately 17% of the sample. The remaining part of the sample is identified as regime two, including the time interval before the most extreme episode of regime three. Approximately 40% of the sample is covered by regime two. Variability decreased, although the persistence increased (0.823). The very beginning of the sample with the lowest number of cases is found as regime four that lasted 30 days with 20% of the sample. During this regime, the persistence was substantial (0.999), but the variability was estimated to be low. Therefore, such a combination of persistence and variability did not result in an explosion of the new cases. The estimated model is given in Table 4, and diagnostic tests are reported in Table 5. The highest transition probability is estimated to be 0.98. It refers to the probability of staying in regime two of lower variability and moderate persistence. The second highest transition probability is estimated to be 0.97. This is the probability of switching from the most extreme regime three to the regime of the lowest number of cases.

The transition probabilities are the following: $p_{11} = 0.95$, $p_{12} = 0.05$, $p_{13} = 0.00$, $p_{14} = 0.00$, $p_{21} = 0.00$, $p_{22} = 0.98$, $p_{23} = 0.02$, $p_{24} = 0.00$, $p_{31} = 0.04$,

Table 5 Specification tests

Linearity	Normality	ARCH 1-2	Autocorr. Q
χ^2_{14}	χ^2_2	F(2,102)	χ^2_{11}
489.9 (0.00, 0.00)	0.4 (0.81)	0.6 (0.58)	19.2 (0.06)

**Fig. 3** United Kingdom: actual and estimated numbers of daily COVID-19 new cases, and estimated regimes

$p_{32} = 0.00$, $p_{33} = 0.96$, $p_{34} = 0.00$, $p_{41} = 0.00$, $p_{42} = 0.03$, $p_{43} = 0.00$, $p_{44} = 0.97$.

3.3 The United Kingdom (Sample: February 24–May 31, 2020)

The three-state MS-AR (2) is estimated to describe well new daily cases of COVID-19 in the United Kingdom (see Fig. 3). The most extreme episode is captured by regime three, covering approximately 40% of the sample. During this regime, data are estimated to have the highest variability. Persistence is determined to be moderate since the largest absolute value of root is 0.57. Before and after this regime, data are estimated to belong to regime one, which includes approximately 41% of the sample. Regime one is associated with a strong upsurge of new cases in the second half of March and the beginning of April and a mild downward trend at the end of the sample. This regime one exhibits strong persistence given that the estimated AR (1) parameter is 0.975 (parameter AR (2) is insignificant). The variability is estimated to be the second highest. This finding highlights the

Table 6 $m = 3$: Constant c_i , autoregressive ϕ_{1i} , and ϕ_{2i} parameters

State	\widehat{c}_i	$\widehat{\phi}_{1i}$	$\widehat{\phi}_{2i}$
1	2268.30 (0.00)	0.975 (0.00)	-0.054 (0.75)
2	2118.53 (0.00)	1.067 (0.00)	-0.070 (0.74)
3	4797.55 (0.01)	0.350 (0.14)	-0.327 (0.04)

Table 7 Standard error σ_i parameter

Parameter	$\widehat{\sigma}_1$	$\widehat{\sigma}_2$	$\widehat{\sigma}_3$
Estimate	404.09	16.67	636.96
Standard error	60.80	3.05	94.32

Table 8 Specification tests

Linearity	Normality	ARCH 1–3	Autocorr. Q
χ^2_{10}	χ^2_2	F(3,72)	χ^2_9
158.1 (0.00, 0.00)	2.7 (0.26)	0.4 (0.79)	12.7 (0.18)

severeness of the augmentation of new COVID-19 cases over this subperiod. At the beginning of the time series, the remaining data are detected as part of regime two, which takes 19% of the sample. Although this regime is associated with a subperiod when the number of new cases has just started to increase, the dynamics of time series are found to be extremely persistent, given that the greatest root of the characteristic equation is 1.067 (parameter AR (2) is insignificant).

Estimated parameters are given in Tables 6 and 7, with diagnostic tests provided by Table 8. The highest transition probability is estimated to be 0.95, being the probability of staying in the episode during which time series followed the explosive path. Slightly lower is transition probability 0.94 representing the probability of staying in the regime with the highest number of new cases.

The transition probabilities are the following: $p_{11} = 0.93$, $p_{12} = 0.01$, $p_{13} = 0.06$, $p_{21} = 0.05$, $p_{22} = 0.95$, $p_{23} = 0.00$, $p_{31} = 0.06$, $p_{32} = 0.00$, $p_{33} = 0.94$.

Note: Two impulse dummy variables are included. The first one is designed to take only nonzero value 1 for April 12, 2020, whereas the second is 1 for May 21, 2020, and 0 otherwise.

3.4 Russia (March 3–May 31, 2020)

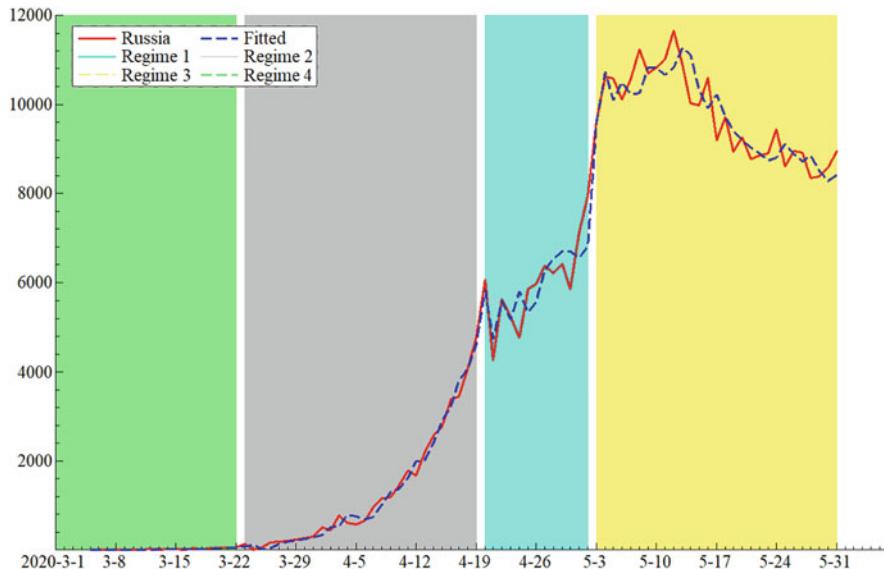
The four-state MS-AR (2) model was chosen for the number of new COVID-19 cases in Russia. Tables 9 and 10 contain estimation results and statistical tests. Figure 4 captures actual and estimated data along with detected regimes. Four regimes are clustered throughout the sample. The most extreme episode is given as regime three, with persistence estimated to be high, 0.99, and variability the second highest. This regime takes one-third of the data. The lowest level episode covers approximately 20% of the data. This is given as regime four. The persistence also appears to be high, 0.966. Between regimes three and four, regimes one and

Table 9 $m = 3$: Constant c_i , AR parameters ϕ_{i1} and ϕ_{i2} , and st. error σ_i

i	\hat{c}_i	$\hat{\phi}_{1i}$	$\hat{\phi}_{2i}$	$\hat{\sigma}_i$
1	488.78 (0.11)	0.470 (0.00)	0.599 (0.00)	571.29 (118.5)
2	110.40 (0.30)	0.847 (0.00)	0.347 (0.02)	155.50 (21.41)
3	2121.00 (0.00)	0.555 (0.00)	0.433 (0.00)	491.69 (68.41)
4	93.52 (0.26)	0.585 (0.01)	0.368 (0.06)	8.85 (1.49)

Table 10 Specification tests

Linearity	Normality	ARCH 1–3	Autocorr. Q
χ^2_{10}	χ^2_2	F(3,63)	χ^2_9
187.2 (0.00, 0.00)	0.3 (0.84)	0.2 (0.92)	8.5 (0.48)

**Fig. 4** Russia: actual and estimated numbers of daily COVID-19 new cases, and estimated regimes

two are detected. Regime two continues after regime three up to April 19, covering 32% of the sample. During this subsample, we observe a strong upward trend that is estimated to describe explosive behavior, given that the largest root is 1.15 (also sum of AR (1) and AR (2) estimates is 1.19). Mild explosiveness is also found during regime one (the largest root is 1.04 and the sum of autoregressive coefficients is 1.07), characterized by the highest variability among all four episodes. During regime one, which lasted approximately 15% of the period, the series still exhibits an upward trend, but with a smaller slope than in regime two.

The transition probabilities are the following: $p_{11} = 0.92$, $p_{12} = 0.00$, $p_{13} = 0.08$, $p_{14} = 0.00$, $p_{21} = 0.04$, $p_{22} = 0.96$, $p_{23} = 0.02$, $p_{24} = 0.00$, $p_{31} = 0.00$,

$p_{32} = 0.00$, $p_{33} = 1$, $p_{34} = 0.00$, $p_{41} = 0.00$, $p_{42} = 0.05$, $p_{43} = 0.00$, $p_{44} = 0.95$.

All transition probabilities of staying in a given regime are greater than 0.9. The probability of staying in the regime of the highest number of cases is even 1, indicating that similar behavior was expected.

3.5 Results Summary

This subsection briefly summarizes key modeling results. We first consider findings for the most extreme episode given the level of persistence estimated (Table 11).

The most extreme episode is identified as regime three in each country, but it differs significantly across them. It is given as one cluster in all cases, except for Italy. Its shortest duration is found in Germany. A much longer duration is detected in the United Kingdom and Russia. The case of Italy is specific because this regime is estimated for several time-series blocks. Variability is found to be the highest in almost all countries during this epidemic plateau regime. Meanwhile, different time-series dynamics are revealed, given the estimated parameters and values of the roots of an AR characteristic equation. The persistence runs from low in Germany to moderate in the United Kingdom and high in Italy and Russia. A combination of low/moderate persistence and high uncertainty suggests that no inertial behavior was found during the epidemic peak in the United Kingdom and Germany.

Results for the pre-peak and post-peak subsamples are presented in Table 12. Except for Germany and Russia, in all other countries, pre-peak and post-peak episodes are identified by the same regime (regime one). In Russia, in the period covered, the epidemic curve did not fall; therefore, we cannot talk about the post-peak episode. However, the persistence is estimated to be extreme in Russia before the epidemic plateau. It is found to be high in the United Kingdom and moderate in Italy. In each case, this regime lasted shorter than the most extreme regime three. Different results are reached for Germany: regime one covers only pre-peak episodes that were found to be highly persistent.

Table 11 Regime three: the most extreme episode

Country	Italy	The United Kingdom	Germany	Russia
Persistence	0.99	0.57	0.37	0.99
Share	31%	40%	19%	33%

Table 12 Regime one: pre-peak and post-peak episodes

Country	Italy	The United Kingdom	Russia
Persistence	0.296	0.975	1.040
Share of the sample	15%	13%	15%

4 Empirical Results for Italy: The Second and the Third Peak

We further explore the time series for Italy so the data that includes the second and the third peak of the COVID-19 pandemic are considered. Given that AR order three is selected, the magnitude of the persistence is assessed from the largest absolute value of the roots of the corresponding characteristic equation.

4.1 *The Second Peak (Sample: October 1, 2020–January 31, 2021)*

The three-state MS-ARMA (3,1) model was chosen for the second peak data in Italy. The peak subsample is contained by regime three, covering 18% of the sample (first 22 days of November 2020). Persistence is estimated to be moderate to high, given that the largest absolute value of roots of the characteristic equation is 0.69. Variability is estimated to be the highest during regime three. Regime one is identified to cover several data just before epidemic peak and just after it, taking only 14% of the sample. Persistence is measured by the absolute value of the root 1.25. Most of the sample (68%) belongs to regime two, representing pre-epidemic and post-epidemic subperiod. Persistence is estimated to be the second highest, given the root of 0.91. Therefore, one may argue that the MS model distinguishes pre-peak, peak, and post-peak regimes. We take regimes one and three to explain the same peak interval, whereas regime two captures both pre-peak and post-peak data (Fig. 5).

Table 13 presents the estimation of constant and AR parameters, whereas Table 14 presents the MA parameter and error variability estimates and gives roots of the AR corresponding characteristic equation. According to several diagnostic tests, the model is not misspecified (Table 15).

4.2 *The Third Peak (Sample: February 1–May 15, 2021)*

The three-state MS-ARMA (3,1) model is again found to explain well the third peak data in Italy. The distribution of episodes across the sample differs from the previous peak period. Regime three is detected to be associated with the largest number of COVID-19 data and includes 32% of the sample (the end of February and March 2021). Persistence is again estimated to be moderate, with the largest absolute value of the characteristic equation root 0.83. Variability is estimated to be the highest. Regime one takes 36% of the sample covering one cluster before and one cluster after regime three. It is characterized by moderate persistence—measured again by the value 0.83. The rest of the sample (32%) is described by regime two, containing

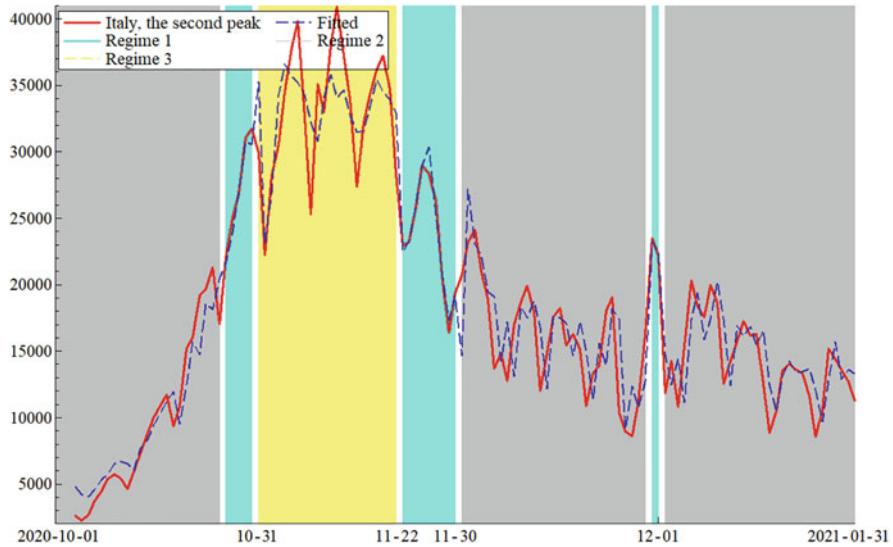


Fig. 5 Actual and estimated numbers of daily COVID-19 new cases, and estimated regimes in Italy: the second peak

Table 13 $m = 3$: Constant c_i , AR parameters ϕ_{i1} , ϕ_{i2} , and ϕ_{i3}

i	\widehat{c}_i	$\widehat{\phi}_{1i}$	$\widehat{\phi}_{2i}$	$\widehat{\phi}_{3i}$
1	28676.5 (.00)	1.035 (.00)	-0.809 (.00)	-0.787 (.00)
2	22576.1 (.00)	1.301 (.00)	-0.768 (.00)	0.376 (.00)
3	37865.9 (.00)	0.242 (.07)	-0.038 (.76)	-0.272 (.01)

Table 14 $m = 3$: MA parameter θ_{i1} , σ_i , and roots of AR char. equation

i	$\widehat{\theta}_{i1}$	$\widehat{\sigma}_i$	Roots
1	-0.117 (0.00)	790.4 (298.5)	-0.50, $0.77 \pm 0.99i$
2	-0.269 (0.00)	2522.5 (230)	0.91, $0.20 \pm 0.61i$
3	-0.117 (0.00)	3424.5 (348)	-0.58, $0.41 \pm 0.55i$

Table 15 Specification tests

	Normality	ARCH 1–2	Autocorr. Q
χ^2_2	F(2,94)		χ^2_{20}
1.1 (0.59)	0.33 (0.72)		29.8 (0.07)

two clusters: at the beginning and the end of the sample. Its persistence is not high in magnitude, as the largest root is 0.61. Again, as in previous cases, we underline different time-series dynamics within the MS model estimation, which depend on the regime found. Regime three is the peak episode, whereas regimes one and two capture pre-peak and post-peak behavior (Fig. 6).

Tables 16 and 17 provide estimates of constant, ARMA parameters, variability, and report roots of the AR corresponding characteristic equation. The model has good statistical properties, although it fails to explain the autocorrelation at higher

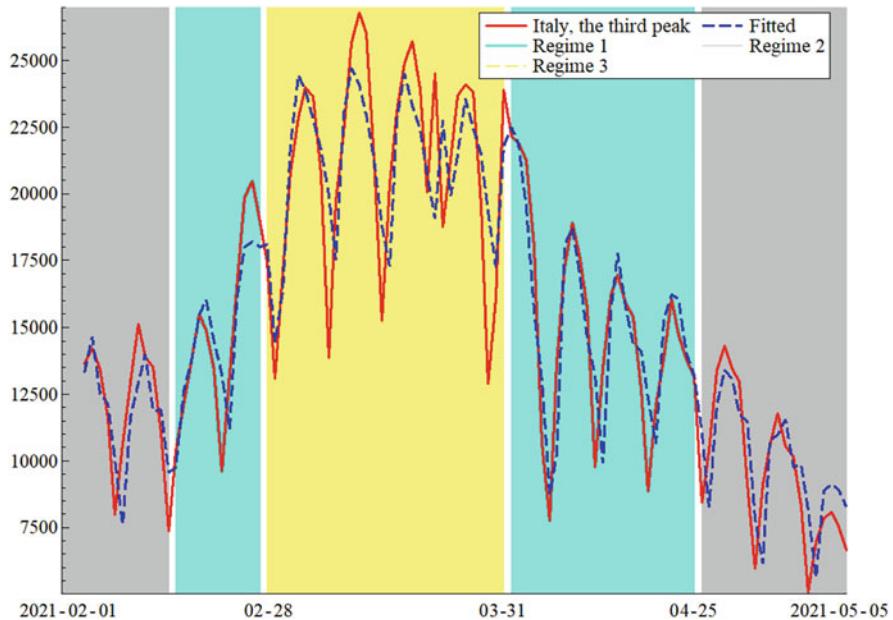


Fig. 6 Actual and estimated numbers of daily COVID-19 new cases, and estimated regimes in Italy: the third peak

Table 16 $m = 3$: Constant c_i , AR parameters ϕ_{i1} , ϕ_{i2} , and ϕ_{i3}

i	\widehat{c}_i	$\widehat{\phi}_{1i}$	$\widehat{\phi}_{2i}$	$\widehat{\phi}_{3i}$
1	1100.1 (.00)	0.466 (.01)	-0.140 (.40)	-0.43 (.00)
2	6986.7 (.00)	0.613 (.01)	0.008 (.97)	-0.193 (.30)
3	22, 544.0 (.00)	0.532 (.00)	-0.284 (.10)	-0.279 (.00)

Table 17 $m = 3$: MA parameter θ_{i1} , σ_i , and roots of AR char. equation

i	$\widehat{\theta}_{i1}$	$\widehat{\sigma}_i$	Roots
1	0.336 (0.00)	1834.2 (250.2)	-0.63, $0.55 \pm 0.62i$
2	0.413 (0.00)	1730.3 (406.7)	0.61, 0, 0
3	-0.04 (0.06)	2673.4 (334)	-0.41, $0.47 \pm 0.68i$

Table 18 Specification tests

	Normality	ARCH 1–2	Autocorr. Q
χ^2_2	F(2,75)	χ^2_{10}	
1.2 (0.55)	1.19 (0.31)	24.3 (0.01)	

lags (Table 18). A model with AR order four solves the problem, but basic results remain unchanged. Therefore, we keep the model with fewer lags to provide an easier comparison for two peak periods.

4.3 Comparison of Two Subsamples

For both subsamples, the same MS specification is chosen. The distinction among regimes is similar: extreme data are contained by regime three, lower values before and after the extreme by regime one, and relatively long episodes of much lower data by regime two. Durations of peak, post-peak, and pre-peak episodes appear to be similar. The peak episode is given by regimes three and one during the second subsample, which covers 32% of the sample. For the third subsample, only regime three is a peak interval, again with 32% of a given sample. Pre-peak and post-peak episodes are proportionally identical (regime two, 68%, and regimes one and two, 68%). The relative size of the regime variability is the same for two subsamples: the highest variability of regime three is followed by the variability of regime one and then regime two.

However, estimated ARMA parameters do not indicate the same level of persistence nor the same influence of isolated shocks, as measured by MA (1) parameter. Persistence is found to be moderate for the most extreme period—regime three (0.69 and 0.83) and much higher for the associated regime one (1.25 and 0.83). The similarity in the persistence is not found for regime two (0.91 vs. 0.61).

During the second peak subsample, the relevance of some adverse shocks is determined for all three regimes, given the negative estimate of the MA (1) parameter. Meanwhile, during the third subsample peak, positive isolated shocks are found according to the same parameter estimates in regimes one and two. All model estimates reflect differences in epidemic measures taken during these two subperiods, including the vaccination process that started during the third wave.

For both models, the highest transition probabilities are found for staying in extreme regime three, and also for remaining in regime two.

5 Concluding Remarks

Our study presents the empirical modeling of daily dynamics of COVID-19 new cases for the following European countries: Italy, Germany, the United Kingdom, and Russia. MS models with ARMA structure are implemented as a key methodological framework. Two aspects are considered. First, for all four countries, the sample includes the first peak. The MS model is a useful approach because it successfully describes the daily dynamics of COVID-19 new cases. However, a different pattern of behavior has been identified across countries. Second, MS models are estimated for Italy's second and third waves. A similar duration of peak, pre-peak, and post-peak episodes is found, and the same specification is chosen. Nevertheless, estimated parameters indicate different dynamics over the second and third peak periods. These empirical findings differ to a greater extent when compared with the results for the first peak.

In summary, MS specification represents a data-driven approach that allows for enough flexibility in modeling pandemic data and extracting episodes of epidemic peak, growth, and decline. It provides valuable information on data dynamics and persistence. It also highlights the necessity of careful data examination for each country separately, given that different epidemiological measures are implemented and that these measures have been subject to change within the same country over time.

References

1. Agosto, A., Giudici, P.: A poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks* **8**(3), 77 (2020)
2. Agosto, A., Campmas, A., Giudici, P., Renda, A.: Monitoring COVID-19 contagion growth. *Stat. Med.* **40**(18), 1–11 (2021)
3. Chakladar, S., Liao, R., Landau, W., Gamalo, M., Wang, Y.: Discrete time multistate model with regime switching for modeling COVID 19 disease progression and clinical outcomes. *Stat. Biopharmaceut. Res.* (2021). <https://doi.org/10.1080/19466315.2021.1880966>
4. Davies, R.B.: Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43 (1987)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977)
6. Doornik, J.A.: Econometric Analysis with Markov-Swithing Models, PcGive 15, OxMetrics 8. Timberlake Consultants Ltd., Richmond (2018)
7. Fokianos, K., Tjøstheim, D.: Log-linear Poisson autoregression. *J. Multivar. Anal.* **102**, 563–578 (2011)
8. Fokianos, K., Rahbek, A., Tjøstheim, D.: Poisson autoregression. *J. Am. Stat. Assoc.* **104**, 1430–1439 (2009)
9. Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384 (1989)
10. Hamilton, J.D.: Analysis of time series subject to changes in regime. *J. Econom.* **45**, 39–70 (1990)
11. Hamilton, J.D.: Time Series Analysis. Princeton University Press, Princeton (1994)
12. Hamilton, J.D.: Specification testing in Markov-switching time series models. *J. Econom.* **70**, 127–157 (1996)
13. Kim, C.-J.: Dynamic linear models with Markov-switching. *J. Econom.* **60**, 1–22 (1994)
14. Lawrence, C.T., Tits, A.L.: A computationally efficient feasible sequential quadratic programming algorithm. *SIAM J. Optim.* **11**, 1092–1118 (2001)
15. Lytras, T., Gkolfinopoulou, K., Bonovas, S., Nunes, B.: FluHMM: a simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Stat. Methods. Med. Res.* **28**, 1826–1840 (2019)
16. Marfak, A., Achak, D., Azizi, A., Nejjari, C., Aboudi, K., Saad, E., Hilali, A., Youlyouz-Marfak, I.: The hidden Markov chain modelling of the COVID-19 spreading using Moroccan dataset. *Data Brief.* **32**, 106067 (2021)
17. Mills, T.C., Markellos, R.N.: The Econometrics Modelling of Financial Time Series, 3rd edn. Cambridge University Press, Cambridge (2008)
18. Mladenović, Z., Glavaš, L., Mladenović, P.: Modeling early COVID-19 contagion dynamics by hidden state autoregressive Markov models. The Conference ITISE2021, pp. 1–12 (2021)
19. Shiferaw, Y.A.: Regime shifts in the COVID-19 case fatality rate dynamics: a Markov-switching autoregressive model analysis. *Chaos, Solitons Fractals X* **6**, 100059 (2021)

20. Strat, Y. L., Carrat, F.: Monitoring epidemiologic surveillance data using hidden Markov models. *Stat. Med.* **18**, 3463–3478 (1999)
21. Teräsvirta, T.: Univariate nonlinear time series models. In: Mills, T., Patterson, K. (eds.). *Palgrave Handbook of Econometrics*, pp. 396–424 (2006)
22. Turasie, A.A.: Temporal dynamics in COVID-19 transmission: case of some African Countries. *Adv. Infect. Dis.* **10**, 110–122 (2020)

Diffusion of Renewable Energy for Electricity: An Analysis for Leading Countries



Alessandro Bessi, Mariangela Guidolin, and Piero Manfredi

Abstract Many countries are undertaking their energy transition process, by investing in renewable energy technologies, in order to face climate change and energy security problems. This paper investigates the temporal trends of the diffusion process of renewable energies, namely, wind and solar, in leading countries for their consumption. In doing so, a bivariate diffusion model is employed to investigate the possibly competitive dynamics between renewables and the top source for electricity production in each country. The obtained results confirm a significant competitive pressure enacted by renewables on the top source. A notable exception is represented by the USA, where renewables appear to reinforce the dominant position of gas.

Keywords Renewable energy · Multivariate diffusion models · Competition · Electricity · Energy transition

1 Introduction

In recent years, in order to face climate change and energy security issues, many countries have undertaken a process of energy transition, characterized by the progressive substitution of nonrenewable energy sources, namely, coal, oil and gas, with renewable energy technologies (RETs), such as photovoltaic, wind, hydroelectric, and biomass. Essentially, energy transition implies a large-scale process of decarbonization [7, 26].

A. Bessi

Department of Economics, University of Messina, Messina, Italy

M. Guidolin (✉)

Department of Statistical Sciences, University of Padova, Padova, Italy
e-mail: guidolin@stat.unipd.it

P. Manfredi

Department of Economics and Management, University of Pisa, Pisa, Italy

Although the switch to renewables is a multidimensional phenomenon, involving a wide variety of social, economic, cultural, and human factors [7], and therefore a multilevel perspective should be adopted in order to capture the complexity of these dynamics [26], there is a certain agreement on the idea that an energy transition primarily involves a transformation of the electricity system [9, 10]. As observed by Geels et al. [7], electricity systems have some specificities that facilitate the integration of RETs.

Focusing on electricity markets, the COVID-19 pandemic has posed extraordinary challenges, or, using the words of the BP's chief economist Spencer Dale, "the global pandemic was the mother of all stress tests" [3], calling the attention on how power systems behave under extreme pressure. At the same time, the COVID-19 pandemic has given the opportunity to learn some important lessons, in order to ensure *secure, flexible, and resilient* power systems in the future [15–17, 20].

According to [18] and [19], the main phenomena observed during the first COVID-19 pandemic year have been a decline in the consumption of coal and gas, with a relative more resilient pattern of gas, and an extraordinary growth of renewables, despite the critical phase suffered by all world economies. Again quoting Spencer Dale, "these trends are exactly what the world needs to see as it transitions to net zero emissions: strong growth in renewable generation crowding out coal." At the same time, it must be recognized that the strong growth in RETs registered in recent times appears not yet sufficient to phase out nonrenewables, and coal in particular, so that a real transformation of the power sector will require a long time to be realized [3].

Stimulated by the strong growth of renewables within the overall electricity mix in many countries, in this paper, we aim to study this process of transformation, by focusing on RETs diffusion in the countries representing top absolute consumers of renewables, as reported in [3]. In detail, the renewable energy sources considered are the "new" RETs, namely, wind and solar, that, according to BP and IEA, have been the major responsible for the recent success of renewables. As already pinpointed, the countries analyzed are the top consumers of new RETs in *absolute* terms, namely, Australia, Brazil, Canada, China, France, Germany, India, Italy, Japan, Spain, the UK and the USA. Moreover, eight of them are among the 12 major energy consumers worldwide. This choice was motivated by the fact that, though many of these countries lie somewhat behind in the per-capita ranking of RETs adopters, meaning that we are leaving out some "virtuous countries" in terms of RETs adoptions, nonetheless they currently represent at the same time massive RETs adopters as well as major carbon dioxide emitters worldwide. It is also reasonable to expect that they will likely keep these roles in the forthcoming decades, thereby playing a central role in the battle against climate change.

To perform the analysis, the paper compares the temporal trend of new (wind and solar jointly considered) RETs consumption with those of the "top source" for electricity production in each of the countries considered. Such comparison has been considered crucial to understand the dynamics of substitution and integration of renewables within power systems, by detecting possible competition effects between energy sources. In particular, the "top source" was defined as *the one*

currently holding the largest share of the energy consumption mix. The latter definition has the advantage of being simple while giving –in view of the simple temporal trend of the top source in most countries considered– the same result of more refined ones.

From the methodological viewpoint, the paper employs a well-accepted approach for the analysis of energy dynamics, based on diffusion models in a competitive setting. Specifically, a bivariate diffusion model is applied to capture –for each country considered– the dynamic interplay between new renewables and the top energy source and detect possibly significant relationships between different energy sources.

The paper has the following structure: Sect. 2 provides the motivation of the research, Sect. 3 resumes some relevant background literature, Sect. 4 illustrates the model employed for the analysis, Sect. 5 describes the obtained results by the selected model, and Sect. 6 proposes some discussion and concluding remarks.

2 Motivation: Energy Trends

As illustrated in the Introduction, the purpose of the paper is to study the diffusion of renewables by considering their dynamic relationship with the major energy source employed for electricity provision. To this end, the set of countries considered is Australia, Brazil, Canada, China, France, Germany, India, Italy, Japan, Spain, the UK, and the USA. This selection has appeared reasonable, because these are currently the global leaders in terms of absolute consumption of the new RETs; additionally, they represent a large subset of major energy consumers worldwide, thus possibly major contributors to current and prospective carbon emissions.

The data for each country are displayed in the various panels of Fig. 1 and cover the period 1965–2020. The black dots represent the time series of RETs, while the blue ones are referred to the time series of the top energy source. The top energy source for electricity provision is represented by natural gas in six countries (Germany, Italy, Japan, Spain, UK, USA), by coal in three (Australia, China, India), by nuclear in France, and by hydropower in Brazil and Canada. In each graph, the scale is determined by the top source trend, leading to significant inter-country differences as absolute energy consumption levels are reasonably influenced by, other things being equal, GDP and population size. As for the trends of (new) RETs, the only countries that showed a robust growth already in the late 1990s were Germany and Spain, while the remaining countries showed their take-off not before 2005. Nevertheless, the trajectories of renewables show a similar structure in most of the countries analyzed, being characterized by an initially flat behavior followed by a phase of marked growth. Such flat dynamics is arguably connected to higher upfront costs and a number of systemic problems [25], which hindered the process in its early stages, thus justifying the need of incentive policies and subsidies. Although most of the countries show a substantially growing trend, one may observe noteworthy exceptions. In particular, in Canada, Italy, and Spain, renewables have

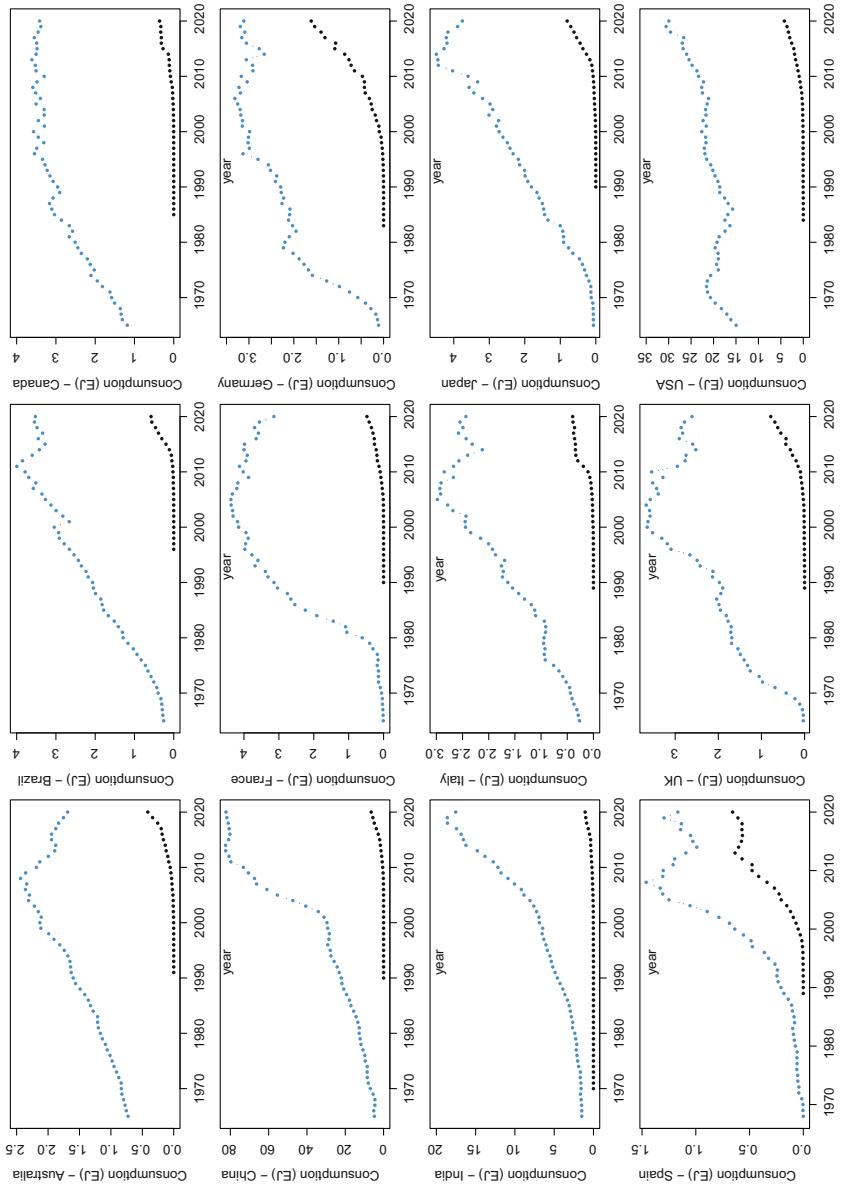


Fig. 1 Time series of yearly absolute consumption (in Exajoules) of new RETs (wind and solar) and the top source for electricity generation in the 12 countries considered, 1965–2020

recorded a fast growth phase, likely due to effective incentives, and a slowdown phase after incentives faded off. Brazil, instead, seems to have suffered from the pandemic breakout, registering minimal changes in 2020.

On the other hand, in most cases, the temporal shape of the top source presents the typical behavior of a mature technology, though at different stages of the energy cycle. In particular, in most countries, it may be observed a phase of sustained growth up to a peak, followed afterward by a recession or by a plateauing epoch (e.g., Brazil, Canada, and China). Notable exceptions to this are represented by India and the USA, where the top source for electricity has been experiencing a continuing increase in recent years, with the only exception of 2020, the first year of the pandemic crisis.

3 Background

The impact of RETs in energy markets has been at the center of a large branch of literature, aimed at modeling the technological, economic, social, and human aspects that may stimulate or hinder their diffusion. For a recent review on some of these streams of research, especially focusing on modeling and forecasting aspects, see [27]. A well-known approach for studying the temporal patterns of energy sources and the related transitions has relied on growth curves such as the logistic equation; see [23, 24], and [8]. Following the lines of research opened by Marchetti and collaborators, in the early 2000s, many contributions applied growth curves for modeling the diffusion of renewables. For a review, especially focused on the diffusion of renewables, see [28]. In particular, some of these works used suitable extensions of the logistic equation, namely, the Bass model [1] and the generalized Bass model [2], to describe the evolution in time of energy sources, both nonrenewables and renewables, and to capture the effects of external shocks, such as ad hoc incentive measures set to accelerate market growth. Among others, we recall [5, 11, 14], and [4]. However, the use of Bass-type models fails to account for the complexity of energy environments that are typically characterized by competition and substitution effects. A first answer to this problem was provided by employing multivariate diffusion models under duopolistic conditions. For example, [10] analyzed the case of Germany's energy transition, by modeling the competitive relationship between RETs and nuclear energy. In [6], the substitution between traditional sources and renewables was modeled for four countries (USA, Europe, China, and India), by also accounting for the effect of external shocks. In [9] the case of Australia was analyzed, by focusing on the relationship, either competitive or collaborative, between coal, gas, and renewables.

The study proposed in this paper aims at providing further insight within this branch of literature.

4 Model

Consistent with the aforementioned literature, this paper employs a general model for a diachronic duopolistic competition proposed in [12]. This model, termed the *unbalanced competition and regime change diachronic model*, UCRCRCD for brevity, is a bivariate generalization of the Bass model, whose purpose is to represent diffusion processes in a competitive environment with two competitors (see [12, 13, 21, 22], and [29]). The UCRCRCD model describes a diffusion process evolving through two sequential phases: an initial phase where only one agent (“player”) occupies the market, thus giving rise to a monopolistic setting, and a second stage where, following the entrance of a concurrent player, true competition occurs. Given these different phases, the market potential is assumed to take different levels: m_a , the market potential in the monopolistic phase, and m_c , the market potential in the competition phase. Letting $z(t)$ denote the overall cumulative number of adoptions at time t , in the competition phase, it holds $z(t) = z_1(t) + z_2(t)$ where $z_1(t)$ and $z_2(t)$ denote adoptions from the first player, the one occupying the market in the monopolistic phase, and from his competitor, respectively. During the competition phase, the residual market $m - z(t)$ is assumed to be shared. The second player enters the market at time $t = c_2$ with $c_2 > 0$.

The model is described through the following system of two differential equations where the time derivatives $z'_1(t)$ and $z'_2(t)$ represent instantaneous adoptions of the first and of the second competitor, respectively, and I_A is the indicator function of time interval A :

$$\begin{aligned} z'_1(t) &= m \left\{ \left[p_{1a} + q_{1a} \frac{z(t)}{m} \right] (1 - I_{t>c_2}) \right. \\ &\quad \left. + \left[p_{1c} + (q_{1c} + \delta) \frac{z_1(t)}{m} + q_{1c} \frac{z_2(t)}{m} \right] I_{t>c_2} \right\} \left[1 - \frac{z(t)}{m} \right], \\ z'_2(t) &= m \left[p_2 + (q_2 - \gamma) \frac{z_1(t)}{m} + q_2 \frac{z_2(t)}{m} \right] \left[1 - \frac{z(t)}{m} \right] I_{t>c_2}, \\ m &= m_a (1 - I_{t>c_2}) + m_c I_{t>c_2} \\ z(t) &= z_1(t) + z_2(t) I_{t>c_2}. \end{aligned} \quad (1)$$

In the monopolistic phase spanning over the time period $t \leq c_2$, the trajectory of the first player, $z'_1(t)$, is described according to a standard Bass model with parameters p_{1a} , q_{1a} , and m_a . Following Bass’ terminology, these parameters represent the market’s innovation rate (p_{1a}), reflecting the hazard of spontaneous adoption under the pressure of the existing communication system; the imitation rate (q_{1a}), reflecting the strength of adoptions due to social contacts (*word-of-mouth*) between agents; and the market potential (m_a), respectively.

In the competition phase, occurring when $t > c_2$, competitors influence each other. This requires a number of additional parameters. The first market player is

now described through (i) the innovation coefficient under competition, p_{1c} ; (ii) the *within* imitation coefficient $q_{1c} + \delta$, describing internal growth dynamics; and (iii) the *cross*-imitation one, q_{1c} , which multiplies z_2/m and measures the influence of the second market player on the first.

The second player has a symmetric structure: (i) the innovation coefficient p_2 ; (ii) the *within* imitation coefficient q_2 , modulating internal growth through the ratio z_2/m ; and (iii) the *cross*-imitation coefficient $q_2 - \gamma$, which tunes the effect of the first player on the second. When parameters δ and γ are identical, i.e., the restriction $\delta = \gamma$ holds, the model has a reduced form, called *standard UCRCD* [12], implying a symmetric behavior between the two competitors. The standard UCRCD model has a closed-form analytical solution.

Typically, internal growth parameters $q_{1c} + \delta$ and q_2 have a positive sign, and their magnitude provides a measure of the intensity of growth. Instead, the cross-imitation parameters may take either a negative or a positive sign: a negative sign implies a competition effect, that is, the competitor has a negative effect on the absolute rate of change of the given player, while a positive one describes a collaborative dynamics.

4.1 Estimation and Model Selection

The statistical implementation of the UCRCD models is based on nonlinear least squares, NLS, [30]. The structure of a nonlinear regression model is as follows:

$$w(t) = \eta(\beta, t) + \varepsilon(t), \quad (2)$$

where $w(t)$ is the observed response, $\eta(\beta, t)$ is the deterministic component depending on parameter vector β and time t , and $\varepsilon(t)$ is a residual term, generally independent and identically distributed (i.i.d.). As for the deterministic component, the literature has considered either the cumulative adoption function $z(t)$ or the instantaneous adoption function $z'(t)$.

Model goodness-of-fit may be evaluated through the determination R^2 index. Moreover, the choice between the unrestricted UCRCD model, with $\delta \neq \gamma$, U , and the standard UCRCD model, with $\delta = \gamma$, S , may be evaluated through a squared multiple partial correlation coefficient \tilde{R}^2 (lying in the interval $[0; 1]$):

$$\tilde{R}^2 = (R_U^2 - R_S^2)/(1 - R_S^2), \quad (3)$$

The \tilde{R}^2 coefficient has a monotone relationship with the F -ratio, i.e.,

$$F = [\tilde{R}^2(n - v)]/[(1 - \tilde{R}^2)k], \quad (4)$$

where n is the number of observations, v the number of parameters of the extended model U , and k the incremental number of parameters from S to U .

5 Application

In order to statistically test the existence of a dynamic relationship between renewables and the main energy source for electricity production in each country, both an unrestricted ($\delta \neq \gamma$) and a standard ($\delta = \gamma$) UCRCD models were fitted. After conducting a model comparison by means of the statistical tools described in Sect. 4.1, the standard model ($\delta = \gamma$) was selected in all cases, except for the USA where the unrestricted model ($\delta \neq \gamma$) had a better performance. A graphical inspection of the model fitting in Fig. 2 allows to appreciate the adequacy of the UCRCD model in reproducing the competing diffusion dynamics of RETs against the top source in all countries considered. Specifically, the increasing trend in renewables observed in all countries was efficiently described, and just in the case of Italy, the model slightly overestimated the behavior of the series. Pairwise, the model reproduction of the key features of the top source temporal trend was also highly satisfactory. Some lack-of-fit arose only when the patterns in the data revealed more complicated than what the flexibility of the model itself can afford, e.g., the fall-rise-fall phases observed in many countries due to the 2008 subprime crisis, the subsequent slow recovery, and the subsequent, possibly definitive, decline in recent years. Some lack-of-fit also occurred, though in some cases only, at the switch between the “monopolistic” and the competition phase.

Table 1 displays parameter estimates of the full UCRCD for the 12 countries. As a general insight, it may be observed that almost all parameters are significant, providing evidence of meaningful relationships in all the cases analyzed. The only exception is represented by parameter p_{2c} , which always proved nonsignificant, confirming findings in previous studies about the lack of meaningful external supports to the market for renewables [4, 11]. Since the focus of the analysis was to characterize the dynamic interplay between renewables and the underlying “top sources” for electricity production, parameter estimates referred to the competition phase, $t > c_2$, are analyzed with more detail and displayed in Table 2. As for the top source, both innovation, p_{1c} , and within imitation coefficient, $q_{1c} + \delta$, are significant and strictly positive. The significant and strongly negative value assumed by the cross-imitation parameter q_{1c} suggests the existence of a strongly competitive pressure exerted by RETs toward the top source.

Further interesting insights on the dynamic relationship come from the coefficients concerning renewables. The nonsignificance of the innovation coefficient p_{2c} and the significant, high value of parameter q_{2c} may be interpreted as an indication of an essentially logistic process meaning that the early stages of RETs life cycle were primarily driven by internal growth forces, namely, imitation and collective learning, in the absence of a sustained support by the public and the media system. Further policy efforts, aiming to foster the continued adoption of renewables by overcoming existing barriers, should keep this into account. The cross-imitation parameter, $q_2 - \gamma$, though significant, appears very small in all countries, especially if compared with q_{1c} : this detected regularity would suggest a regime change characterized by an independent path behavior of renewables, being marginally

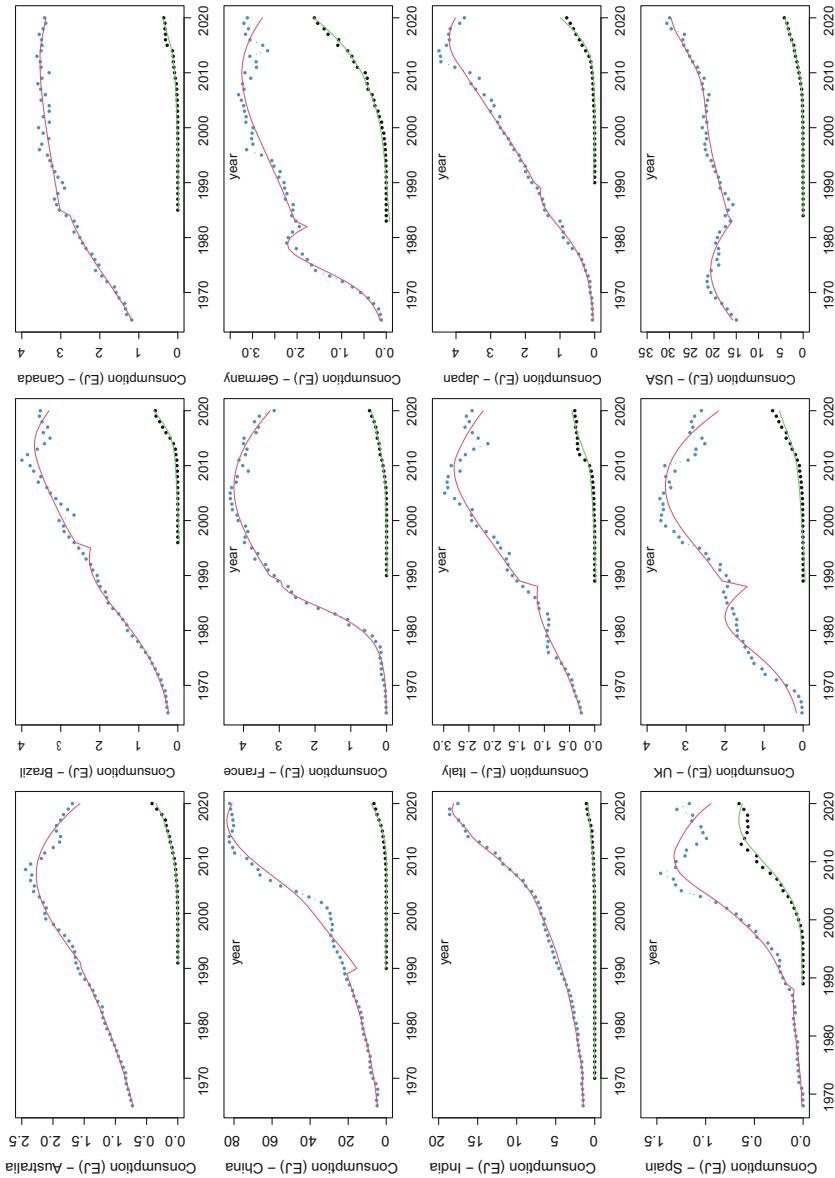


Fig. 2 Dynamic relationship between consumption of the top energy source (blue dots, observed; red curve, predicted) and of renewables (black dots, observed; green curve, predicted) in the countries considered, 1965–2020, as resulting from the fit of the UCRC model

Table 1 Parameter estimates of the UCRCRD model for the 12 countries selected

Country	Parameter	Estimate	s.e.	Lower c.i.	Upper c.i.	p-value	R^2
Australia	m_c	98.5	3.24	92.2	104.9	<0.0001	0.999980
	p_{1c}	0.015	0.0004	0.014	0.016	<0.0001	
	p_{2c}	-0.0001	0.0003	-0.0007	0.0006	0.861	
	q_{1c}	-0.3606	0.055	-0.4687	-0.2524	<0.0001	
	q_{2c}	0.4204	0.054	0.3152	0.5255	<0.0001	
	δ	0.4199	0.0547	0.3127	0.5272	<0.0001	
Brazil	m_c	225.8	42.1	143.2	308.4	<0.0001	0.999941
	p_{1c}	0.011	0.0019	0.008	0.015	<0.0001	
	p_{2c}	-0.0002	0.0003	-0.0008	0.0004	0.561	
	q_{1c}	-0.253	0.063	-0.378	-0.129	<0.0001	
	q_{2c}	0.298	0.061	0.178	0.418	<0.0001	
	δ	0.296	0.063	0.173	0.418	<0.0001	
Canada	m_c	580.2	206.5	175.5	984.9	0.0065	0.999992
	p_{1c}	0.0052	0.0018	0.0017	0.0087	0.0053	
	p_{2c}	-0.0001	0.0001	-0.0002	0.0001	0.447	
	q_{1c}	-0.139	0.026	-0.189	-0.088	<0.0001	
	q_{2c}	0.154	0.025	0.105	0.202	<0.0001	
	δ	0.153	0.025	0.103	0.202	<0.0001	
China	m_c	3445.5	270.2	2916.0	3975.0	<0.0001	0.999712
	p_{1c}	0.0039	0.0004	0.0032	0.0046	<0.0001	
	p_{2c}	0.0001	0.0004	-0.0006	0.0008	0.782	
	q_{1c}	-0.456	0.188	-0.825	-0.086	0.0189	
	q_{2c}	0.551	0.185	0.188	0.914	0.0043	
	δ	0.552	0.187	0.185	0.919	0.0046	
France	m_c	208.5	6.2	196.3	220.7	<0.0001	0.999987
	p_{1c}	0.0152	0.0004	0.0145	0.0160	<0.0001	
	p_{2c}	-0.0001	0.0002	-0.0005	0.0003	0.541	
	q_{1c}	-0.262	0.049	-0.357	-0.166	<0.0001	
	q_{2c}	0.309	0.048	0.216	0.403	<0.0001	
	δ	0.308	0.048	0.214	0.403	<0.0001	
Germany	m_c	467.6	122.3	227.9	707.3	0.0003	0.999905
	p_{1c}	0.0043	0.0010	0.0022	0.0063	<0.0001	
	p_{2c}	-0.0002	0.0001	-0.0004	0.0001	0.208	
	q_{1c}	-0.0935	0.009	-0.112	-0.075	<0.0001	
	q_{2c}	0.127	0.011	0.106	0.147	<0.0001	
	δ	0.122	0.110	0.101	0.144	<0.0001	

(continued)

Table 1 continued

Country	Parameter	Estimate	s.e.	Lower c.i.	Upper c.i.	p-value	R^2
India	m_c	690.9	60.2	573.0	808.9	<0.0001	0.999948
	p_{1c}	0.0096	0.0007	0.0081	0.0110	<0.0001	
	p_{2c}	0.0001	0.0003	-0.0004	0.0006	0.576	
	q_{1c}	-0.281	0.163	-0.600	0.039	0.0951	
	q_{2c}	0.377	0.160	0.064	0.069	0.0249	
	δ	0.380	0.163	0.059	0.700	0.0271	
Italy	m_c	133.1	7.7	118.1	148.2	<0.0001	0.999874
	p_{1c}	0.0106	0.0005	0.0095	0.0116	<0.0001	
	p_{2c}	-0.0002	0.0004	-0.0010	0.0006	0.610	
	q_{1c}	-0.189	0.053	-0.293	-0.085	0.0007	
	q_{2c}	0.254	0.051	0.155	0.354	<0.0001	
	δ	0.252	0.052	0.149	0.355	<0.0001	
Japan	m_c	427.0	107.5	216.2	637.8	0.0002	0.999929
	p_{1c}	0.0039	0.0009	0.0021	0.0056	<0.0001	
	p_{2c}	0.0001	0.0001	-0.0001	0.0003	0.440	
	q_{1c}	-0.273	0.033	-0.337	-0.208	<0.0001	
	q_{2c}	0.320	0.032	0.257	0.383	<0.0001	
	δ	0.322	0.033	0.257	0.3871	<0.0001	
Spain	m_c	50.9	2.0	47.1	54.8	<0.0001	0.999358
	p_{1c}	0.0029	0.0007	0.0017	0.0042	<0.0001	
	p_{2c}	0.0002	0.0007	-0.0012	0.0015	0.810	
	q_{1c}	-0.089	0.026	-0.141	-0.037	0.0014	
	q_{2c}	0.229	0.025	0.179	0.278	<0.0001	
	δ	0.224	0.032	0.162	0.286	<0.0001	
UK	m_c	143.8	6.6	130.9	156.6	<0.0001	0.999674
	p_{1c}	0.0014	0.0007	0.0124	0.0150	<0.0001	
	p_{2c}	-0.0004	0.0007	-0.0017	0.0009	0.550	
	q_{1c}	-0.291	0.083	-0.454	-0.128	0.0009	
	q_{2c}	0.363	0.081	0.205	0.521	<0.0001	
	δ	0.360	0.083	0.198	0.522	<0.0001	
USA	m_c	1340.3	102.3	1139.8	1540.7	<0.0001	0.999976
	p_{1c}	0.0123	0.0009	0.0105	0.0140	<0.0001	
	p_{2c}	0.000004	0.0002	-0.0003	0.0004	0.981	
	q_{1c}	1.266	0.295	0.687	1.845	<0.0001	
	q_{2c}	0.392	0.062	0.269	0.514	<0.0001	
	δ	-1.232	0.294	-1.807	-0.656	<0.0001	
	γ	0.392	0.063	0.268	0.515	<0.0001	

Table 2 Parameter estimates of the UCRCD model for the dynamic relationship between the consumption of the top energy source (in brackets) and that of renewables for the 12 countries considered. The reported parameter estimates refer to the competition phase only ($t > c_2$)

Country	m_c	p_{1c}	$(q_{1c} + \delta)$	q_{1c}	q_2	$(q_2 - \gamma)$	δ	γ
Australia (R vs C)	99	0.015	0.06	-0.36	0.42	0.0004	0.42	
Brazil (R vs H)	226	0.011	0.04	-0.25	0.30	0.0022	0.30	
Canada (R vs H)	580	0.005	0.01	-0.14	0.15	0.0010	0.15	
China (R vs C)	3445	0.004	0.10	-0.46	0.55	-0.0012	0.52	
France (R vs N)	209	0.015	0.05	-0.26	0.31	0.0011	0.31	
Germany (R vs G)	468	0.004	0.03	-0.09	0.13	0.0042	0.12	
India (R vs C)	691	0.010	0.10	-0.28	0.38	-0.0029	0.38	
Italy (R vs G)	133	0.011	0.06	-0.19	0.25	0.0023	0.25	
Japan (R vs G)	427	0.004	0.05	-0.27	0.32	-0.0017	0.32	
Spain (R vs G)	51	0.003	0.14	-0.09	0.23	0.0043	0.22	
UK (R vs G)	144	0.014	0.07	-0.29	0.36	0.0029	0.36	
USA (R vs G)	1340	0.012	0.03	1.27	0.39	-0.0002	-1.23	0.39

influenced by the dominant source, as already pointed out in previous studies based on a duopolistic model [9, 10].

To sum up, some evident regularities emerge from the analysis, highlighting a highly competitive pressure of renewables on the top source, while the top source seems to aid the integration of renewables, although with a very small quantitative effect.

Notable exceptions to this general pattern are China, India, and Japan, where both q_{1c} and $q_2 - \gamma$ are negative, indicating the presence of a pure competition dynamic relationship between the two energy sources.

Last but not least, a unique relationship characterizes the case of the USA, where q_{1c} is extremely high and positive, while $q_2 - \gamma$ is negative, though negligible, suggesting a competitive effect of the top source toward the growth of renewables. This exceptional behavior is arguably connected to the renewed strength of natural gas in the US energy market, where the growth of renewables appears to “collaborate” to establish a new energy regime of electricity production, namely, from a previous one essentially based on coal and nuclear to a new one centered on natural gas, where renewables appear to play –at least in this stage of RETs lifecycle– a supporting role to the gas dominance.

6 Discussion

This paper performed a multicountry analysis, which aimed at identifying significant dynamic relationships between the diffusion process of renewables and that of the top source for electricity production currently prevailing in each country considered. Since the transformation of electricity systems through a progressive

expansion and integration of renewables is seen as a necessary step for the current ecological and energy transitions, grounding the growth of renewables against that of the top source in the electricity mix has appeared a natural choice, in order to provide a credible representation of current trends.

From this viewpoint, the present findings appear especially interesting since the same model has been applied to different countries, giving rise to a number of evident regularities in diffusion patterns: a clear competitive effect exerted by renewables that seems to follow a robust and somehow independent growth path within the electricity market.

A special attention should be devoted to the last observation available in the data, referring to year 2020, for which a decline in electricity consumption from the top source has been observed in almost all countries with respect to year 2019. The only exception is represented by China, for which a slight increase in coal consumption has been observed in 2020. However, these variations do not indicate a negative shock in consumption, as confirmed by the good UCRCD fit in this part of the data: indeed electricity continued to be a primary good, also, or perhaps especially, during the pandemic crisis and the related lockdown periods during 2020, so that a strong decrease in consumption was not a plausible outcome.

This work has a number of limitations. First, in order to follow a parsimonious approach, the study was based on some key simplifying hypotheses, purposely overlooking sources of heterogeneity among countries, such as the availability of multiple energy sources, socioeconomic conditions, national policies, international agreements, etc. In this sense, some lack-of-fit was observed especially in the first phase, i.e., the one prior to the start of the market for renewables, that we termed the “monopolistic” phase owing to the two-dimensional nature of the adopted UCRCD modeling framework. Clearly, representing the initial phase prior to renewables onset as a “monopolistic” one driven by a simple Bass-like trend is a strong hypothesis entailing a simplification. Indeed, it ignores that in the energy market of most countries considered, there already was (in the pre-renewables epoch) a competition among several energy sources (e.g., in the US case among natural gas, coal, and nuclear). Therefore, the trajectory of the top energy source in the “monopolistic” phase was generated by a dynamic process far more complicated than the simple Bass model adopted here. So we unavoidably expected that –at least in some cases– the fit to the data would be sub-optimal. Future research may aim at improving the fit also for this phase, by relaxing the assumption of a Bass-like behavior, allowing for the introduction of a more suitable function.

A further clear limitation of the proposed analysis is the absence of an out-of-sample forecasting exercise. However, especially in light of the likely structural changes that will be induced by the enduring pandemic and especially by the national post-pandemic recovery plans, we preferred not to provide forecasts for the next years and to definitely bound our analysis on the characteristics of the competition on energy markets until the year 2020. A future perspective of research will focus on clarifying and possibly modeling the still uncertain role of the pandemic and related recovery plans in the growth of renewables, in order to propose realistic future scenarios.

Acknowledgments This research has been partially funded by the grant BIRD188753/18 of the University of Padua, Italy. We warmly thank two anonymous reviewers whose valuable comments allowed us sharply improve the Discussion of this work. Usual disclaimers apply.

References

1. Bass, F.M.: A new product growth for model consumer durables. *Manag. Sci.* **15**(5), 215–227 (1969)
2. Bass, F.M., Krishnan, T.V., Jain, D.C.: Why the BM fits without decision variables. *Market. Sci.* **13**(3), 203–223 (1994)
3. BP Statistical Review of World Energy 2021. Available at www.bp.com
4. Bunea, A.M., Della Posta, P., Guidolin, M., Manfredi, P.: What do adoption patterns of solar panels observed so far tell about governments' incentive? Insights from diffusion models. *Technol. Forecast. Soc. Chang.* **160**, 120240 (2020)
5. Dalla Valle, A., Furlan, C.: Forecasting accuracy of wind power technology diffusion models across countries. *Int. J. Forecast.* **27**(2), 592–601 (2011)
6. Furlan, C., Mortarino, C.: Forecasting the impact of renewable energies in competition with non-renewable sources. *Renew. Sust. Energ. Rev.* **81**, 1879–1886 (2018)
7. Geels, F.W., Sovacool, B.K., Schwanen, T., Sorrell, S.: Sociotechnical transitions for deep decarbonization. *Science* **357**(6357), 1242–1244 (2017)
8. Grubler, A.: Energy transitions research: insights and cautionary tales. *Energ. Policy* **50**, 8–16 (2012)
9. Guidolin, M., Alpcan, T.: Transition to sustainable energy generation in Australia: interplay between coal, gas and renewables. *Renew. Energ.* **139**, 359–367 (2019)
10. Guidolin M., Guseo R.: The German energy transition: modelling competition and substitution between nuclear power and Renewable Energy Technologies. *Renew. Sust. Energ. Rev.* **60**, 1498–1504 (2016)
11. Guidolin, M., Mortarino, C.: Cross-country diffusion of photovoltaic systems: modelling choices and forecasts for national adoption patterns. *Technol. Forecast. Soc. Chang.* **77**(2), 279–296 (2010)
12. Guseo R., Mortarino C.: Within-brand and cross-brand word of mouth for sequential multi-innovation diffusions. *IMA J. Manag. Math.* **25**(3), 287–311 (2014)
13. Guseo, R., Mortarino, C.: Modelling competition between two pharmaceutical drugs using innovation diffusion models. *Ann. Appl. Stat.* **9**(4), 2073–2089 (2015)
14. Guseo, R., Dalla Valle, A., Guidolin, M.: World Oil Depletion Models: price effects compared with strategic or technological interventions. *Technol. Forecast. Soc. Change* **74**(4), 452–469 (2007)
15. Heffron, R.J., Körner, M.F., Schöpf, M., Wagner, J., Weibelzahl, M.: The role of flexibility in the light of the COVID-19 pandemic and beyond: Contributing to a sustainable and resilient energy future in Europe. *Renew. Sust. Energ. Rev.* **140**, 110743 (2021)
16. Henry, M.S., Bazilian, M.D., Markuson, C.: Just transitions: histories and futures in a post-COVID world. *Energ. Res. Soc. Sci.* **68**, 101668 (2020)
17. Hoang, A.T., Nižetić, S., Olcer, A.I., Ong, H.C., Chen, W.H., Chong, C.T., Nguyen, X.P.: Impacts of COVID-19 pandemic on the global energy system and the shift progress to renewable energy: opportunities, challenges, and policy implications. *Energy Policy* **154**, 112322 (2021)
18. IEA: Renewable Energy Market Update 2021, IEA, Paris (2021). <https://www.iea.org/reports/renewable-energy-market-update-2021>
19. IEA: Electricity Market Report – July 2021, IEA, Paris (2021). <https://www.iea.org/reports/electricity-market-report-july-2021>

20. Jiang, P., Van Fan, Y., Klemeš, J.J.: Impacts of COVID-19 on energy demand and consumption: challenges, lessons and emerging opportunities. *Appl. Energ.* **116**441 (2021)
21. Krishnan T.V., Bass F.M., Kumar V.: Impact of a late entrant on the diffusion of a new product/service. *J. Market. Res.* **37**, 269–278 (2000)
22. Laciana, C.E., Gual, G., Kalmus, D., Oteiza-Aguirre, N., Rovere, S.L.: Diffusion of two brands in competition: cross-brand effect. *Physica A* **413**, 104–115 (2014)
23. Marchetti, C.: Society as a learning system: discovery, invention, and innovation cycles revisited. *Technol. Forecast. Soc. Change* **18**(4), 267–282 (1980)
24. Marchetti, C., Nakicenovic, N.: The dynamics of energy systems and the logistic substitution model. Research Report RR-79-13. Laxenburg: International Institute for Applied Systems Analysis (1979)
25. Negro, S.O., Alkemade, F., Hekkert, M.P.: Why does renewable energy diffuse so slowly? A review of innovation system problems. *Renew. Sustain. Energ. Rev.* **16**(6), 3836–3846 (2012)
26. Otto, I.M., Donges, J.F., Cremades, R., Bhowmik, A., Hewitt, R.J., Lucht, W., Schellnhuber, H.J.: Social tipping dynamics for stabilizing Earth's climate by 2050. *Proc. Natl. Acad. Sci.* **117**(5), 2354–2365 (2020)
27. Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., ... & Ziel, F. Forecasting: theory and practice. *Int. J. Forecast.* **38**(3), 705–871 (2022)
28. Rao, K.U., Kishore, V.V.N.: A review of technology diffusion models with special reference to renewable energy technologies. *Renew. Sustain. Energ. Rev.* **14**(3), 1070–1078 (2010)
29. Savin, S., Terwiesch, C.: Optimal product launch times in a duopoly: balancing life-cycle revenues with product cost. *Oper. Res.* **53**(1), 26–47 (2005)
30. Seber, G.A.F., Wild, C.J.: Nonlinear Regression. Wiley, New York (1989)

The State and Perspectives of Employment in the Water Transport System of the Republic of Croatia



Drago Pupavac Ljudevit Krpan and Robert Maršanić

Abstract The main aim of this paper is to investigate the state of employment and employment trends in the water transport system of the European Union and the Republic of Croatia. The purpose of this paper is to find answers to the question of how to turn negative employment trends in the Croatian water transport system into positive ones. To answer this question, several scientific methods were applied, in particular descriptive statistics and correlation and regression analysis. The increase in goods transport and the growth of the gross domestic product have been recognized as major factors in increasing employment in the water transport system. The main findings of this paper can be helpful to transport managers at all levels for human resource planning in the water transport system.

Keywords Water transport · Maritime transport · Inland waterway transport · Employment

1 Introduction

Water transport covers the transport of goods and persons by ships that travel on the sea or on inland waterways. The Republic of Croatia is a maritime country [1] and Croatia's water transport system is part of both the European and the global transport system [2]. Since there is no strongly oriented maritime economy and the connection between sea and river ports is rather poor, this precious resource remains insufficiently used in Croatia [3]. This is especially true with regard to Croatia's

D. Pupavac ()

Polytechnic of Rijeka, Rijeka, Croatia

e-mail: drago.pupavac@veleri.hr

L. Krpan

Primorsko-goranska County & University North, Rijeka, Croatia

R. Maršanić

Road Administration Primorje and Gorski Kotar County & University North, Rijeka, Croatia

inland waterway transport [4]. The largest transport volume in inland waterway transport, amounting to 5.48 million tonnes (7.7% of the total land transport), was recorded in 1980, whereas today it is almost nine times smaller and amounts to only 632 thousand tons [5]. In the pre-transition period in 1988, water transport employed 15.5 thousand people whereas today it employs only 3529 people or 4.39 times less [6]. Reasons for this should be looked for in decreased economic activity, insecure industrial production [7], decreased employment in transport activity as a whole [8], insufficient integration between Croatian sea ports and river ports, loss of large shipping companies [9], closing of steelworks and the Oil Refinery in Sisak, non-maintenance of inland waterways, and the lack of qualified employees (river ship operators). The negative economic trends that emerged in 2009 had an adverse effect on employment in the whole Croatian transport sector [10]. About 57,000 transport system jobs were lost, of which water transport has lost 13,000 in the past three decades [11]. The water transport system is a part of Croatia's transport system, which has suffered the adverse effects of numerous crises over the past three decades: the transition crisis (1991–2000), the global financial crisis, the fallout of which was felt in Croatia from 2009 to 2014, and the crisis caused by the COVID-19 pandemic (2020–2021). Accordingly, the purpose of this paper is to find an answer to the question of how to turn negative employment trends in the water transport system into positive ones.

2 Theoretical Framework

The most important types of water transport are shown in Fig. 1.

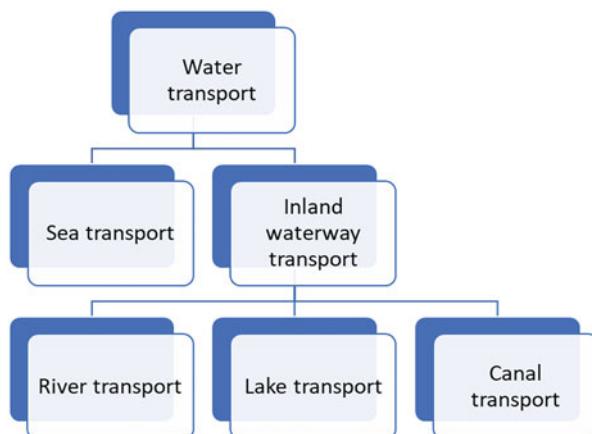


Fig. 1 Types of water transport

1. *Sea transport.* Sea transport is also called sea shipping. The main features of sea transport are that it is undertaken at sea, on a natural and free-of-charge waterway, in various types of ships and vessels, and that it requires artificially built starting and finishing points—seaports. The world merchant fleet is made up of some 100,000 commercial vessels, about one third of which are controlled by EU countries. Maritime industries are an important source of employment and income for the European economy [12]. The total capacity of Croatian maritime shipping has declined in the last three decades and at the end of 2016 it amounted to 127 ships—86 passenger ships and 41 cargo ships [5].
2. *Inland waterway transport* is, together with road and rail transport, one of the three main land transport modes. Goods are transported by ships via inland waterways, such as: (a) rivers, (b) lakes, and (c) canals. A brief description of each type of inland waterway transport is given below:
 - (a) River transport and traffic. It is carried out on navigable rivers, on a natural and free-of-charge waterway, in various types of vessels: ships, cargo barges, small vessels, barges, push boats, and tug boats, and it requires artificially built starting and finishing points—docks. The European inland waterway network represents the most significant and most developed regional market of river transport. The basic feature of river waterways worldwide is their under-utilization. River transport of the developed countries of the European Union accounts for 25% of their total transport, resulting in multiple savings in costs for the economy.
 - (b) Lake transport and traffic. It is maintained on navigable lakes, on a natural and free-of-charge waterway, in various types of vessels, and similar to sea and river transport and traffic, it requires artificially built starting and finishing points—docks.
 - (c) Canal transport and traffic. It has all relevant features of sea, river, and lake transport and traffic, however with a significant difference—it is carried out on artificial canals. Canal transport has particular significance in the international transport of goods and passengers. The basic feature of canal transport is reflected in its ability to fundamentally change the transport importance of individual parts of the world and the transportation routes of goods by removing geographical barriers to transport. By connecting different seas, canal transport has contributed to the increase of international trade exchange almost as much as technical progress in transport. On a global level, the Suez Canal and the Panama Canal are of particular importance whereas the Kiel Canal and the Corinthian Canal are of regional significance.

Water transport as a whole, together with its individual (sub)types (sea, river, lake, and canal transport) has numerous technical, technological, organizational, economic, and legal idiosyncrasies that all active stakeholders of this system should follow, know, and put into practice, since only by doing so can they materially affect the safety, speed, and rationality of the very complex process of transport services production.

3 Descriptive Analysis of Employment in Water Transport in the EU

3.1 Descriptive Analysis of Employment in Sea Transport in the EU

Maritime countries, among which the EU-27 plays a special role, are the primary drivers of the development of international exchange and trade around the world. Bringing together production and consumption across the far corners of the world, sea transport has always been a factor in the integration of countries and the world market as a whole [13]. The steady growth of international seaborne trade, measured in the number of tonnes carried, has increased 4.25 times [14] in the past 50 years, implying the continuous growth of total ship capacities and the number of persons employed in sea transport. The number of seafarers in the world is estimated at 1.65 million [15], with the EU countries accounting for about 230,000 [16], and Croatia for over 18,000. Due to continuous improvements to the technological structure of sea-going ships, the number of people employed in sea transport has not grown proportionally with the increase in ship capacities. Hence, it comes as no surprise that there is a positive and weak correlation ($r = 0.23$; $p < 0.5$) between the number of sea-going ships and the number of employees in Croatia's sea transport system [6] and that the global demand for officers is greater than the global supply, while the global supply of ratings exceeds the global demand (see Fig. 2).

The graphs in Fig. 2a, b demonstrate what occurs when, given the economic reality on the global seafarer labor market, the rate of pay is set at the W_1 level for ratings and the W_2 level for officers. Namely, in economic reality, wages do not adjust to the equilibrium wage rate but rather are inelastic and slow to respond to economic changes. When wages fail to adjust to the equilibrium rate, a disbalance

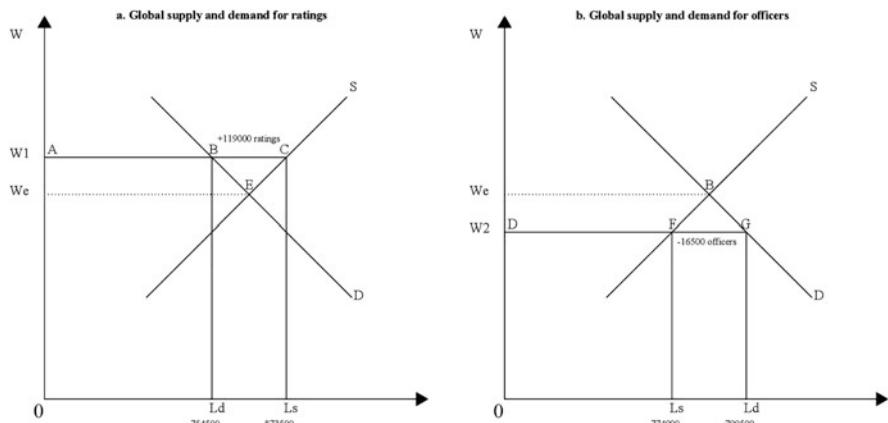


Fig. 2 Global supply and demand for officers and ratings in sea transport

may occur between the number of workers looking for a job and the number of job vacancies. In the first case (Fig. 2a), the number of ratings looking for work (Ls or AC) is larger than the demand for ratings according to the number of job vacancies (Ld or AB). When Ld (or CE) ratings find a job, Ls-Ld or BC become involuntarily unemployed ratings (119,000), that is, ratings who wish to work for current wages but cannot find a job. In the second case (Fig. 2b), the situation differs essentially, because wages are set below the We level. In this case, there is a shortage of workers and shipping companies are not able to fill all the vacant officer positions. This occurs because the number of workers seeking work (Ls or DF) is smaller than the number of vacant officer positions (Ld or DG). As Ld (or DF) officers become employed, Ld-Ls (or FG) represents the shortage of workers (16,500) that shipping companies want to employ at current wages but cannot find.

The number of persons employed in sea transport in the EU-28 in 2017 varied considerably, from only 200 employed in Slovenia to over 48,000 employed in Italy. The average number of people employed in sea transport per member country (see Table 1) was 6.37 (SD = 10.30).

According to the number of sea transport workers, the Republic of Croatia, with 3429 workers, is below the EU-28 average. However, having 18,658 seafarers [6], Croatia is substantially above the average of EU countries ($M = 10,491$; $SD = 1807.73$). Eastern Member States such as Bulgaria (33,269), Romania (24,343), Poland (22,669), and Croatia (18,658) employ more than 35% of the total number of seafarers.

Table 1 Descriptive statistics on employment in sea transport in EU-28, 2017 (000)

Mean	6.37
Standard Error	1.94
Median	1.3
Standard Deviation	10.3
Sample Variance	106.16
Kurtosis	10.35
Skewness	2.86
Range	48.9
Minimum	0
Maximum	48.9
Sum	178.3
Count	28
Confidence Level (95.0%)	3.99

Source: Prepared by the authors according to [17]

3.2 Descriptive Analysis of Employment in the EU Inland Waterway Transport

Inland waterway transport is very developed in the EU-27, in particular in the Netherlands, Germany, and France. The network of the inland waterways of EU countries has more than 37,000 km [18] of waterways.

The Rhine–Danube network, with a length of 14,360 km, represents the main international inland waterway network. The most important basins are: (a) the Rhine basin—around 80% of the overall inland waterway freight transport is carried on this river and (b) the Danube basin—around 9% of the overall inland waterway transports is carried out on the Danube and the Rhine–Main–Danube canal [19]. The Danube basin has the potential to guarantee river navigation between the North Sea and the Black Sea.

Croatia possesses significant natural and geographical potential for the development of river transport. The determining factors of this potential are the rivers Danube and Sava, and a part of the waterway of the river Drava (to 22 rkm).

Croatia has 534.7 km of waterways, of which 287.4 km or 53.75% complies with the requirements of international waterway norms [20]. The inland waterways are located in the northern part of the country. The construction of the Danube–Sava canal would result in the creation of a singular national navigable network, which would become an integral part of the singular navigable waterway network of the EU-27. This would contribute considerably to increasing the volume of traffic on Croatian waterways.

The EU-28 inland water transport system in 2017 employed 44,300 workers or 0.42% of the total number of employees in the EU-28 transport system, or 27.4% of the total number of employees in the EU-28 water transport system. Five European Union countries do not keep records of the number of employees in inland water transport while five European Union countries have only 100 employees in inland water transport. Croatia is one of them. Below is a brief overview of employment in the EU-28 inland water transport traffic based on the method of descriptive statistics (Table 2).

Based on data from Table 2, it is evident that average employment per member country is 1580 people ($SD = 3.07$). The largest number of people 13,400 are employed by the inland water transport system of the Netherlands. According to the number of inland water transport workers, Croatia is below the EU-28 average.

4 Data and Research Methodology

In order to make a model for evaluating total employment in waterway transport in Croatia (TEWT), it is necessary to first make two partial models (a separate model for evaluating employment in sea transport and another one for inland waterway transport), and then combine these two to obtain one integral model.

Table 2 Descriptive statistics on employment in inland water transport in EU-28, 2017 (000)

Mean	1.58
Standard Error	0.58
Median	0.6
Standard Deviation	3.07
Sample Variance	9.42
Kurtosis	9.86
Skewness	3.15
Range	13.14
Minimum	0
Maximum	13.4
Sum	44.3
Count	28
Confidence Level (95.0%)	1.19

Source: Prepared by the authors according to [17]

The first model to be presented starts from the premise that the number of persons employed in sea transport depends on (1) the number of passenger ships—NPS, (2) the number of cargo ships—NCS, (3) passengers carried—PC, (4) passenger miles—PM, (5) goods carried—GC, (6) tonne-miles—TM, and (7) GDP.

The model can be presented in the following way:

$$\text{NES} = b_0 + b_1 \text{NPS} + b_2 \text{NCS} + b_3 \text{PC} + b_4 \text{GC} + b_5 \text{PM} + b_6 \text{TM} + b_7 \text{GDP} \quad (1)$$

b_i —($i = 0, 1, 2, 3, 4, 5, 6, 7$) = model parameters.

The second model to be presented starts from the premise that the number of persons employed in inland waterway transport depends on the (1) transport of goods in national transport, (2) transport of goods in international transport, and (3) gross domestic product.

Its linear form would be as follows:

$$\text{NER} = b_0 + b_1 \text{NT} + b_2 \text{IT} + b_3 \text{GDP} \quad (2)$$

b_i —($i = 0, 1, 2, 3$) = model parameters.

This approach to inland waterway transport was chosen because the Croatian Bureau of Statistics does not keep records of the total transport volume of passengers in Croatian river ports, and data on river fleet per years are not available.

By combining the two models, an integrated model for estimating the total number of persons employed in Croatia's waterway transport system is obtained.

$$\text{TEWT} = \text{NES} + \text{NER} \quad (3)$$

Table 3 Employment in sea transport, passenger ships, cargo ships, passengers carried, passenger miles, goods carried, tonne-miles and GDP, 2005–2016

Year	NE	PM in mln	NPS	TM in mln	NCS	PC in 000	GC in 000t	GDP constant price in HRK
2005	4255	233	86	68,069	69	11,440	29,975	292,859.83
2006	4203	245	86	73,971	69	12,079	31,423	306,739.8
2007	4290	265	91	74,230	67	12,723	32,420	323,522.76
2008	4154	265	88	77,199	68	12,861	30,768	331,155.41
2009	3862	263	88	74,160	64	12,550	31,371	308,305.68
2010	3870	266	85	87,878	68	12,506	31,948	301,214.65
2011	3830	315	80	83,929	67	12,926	30,348	301,214.65
2012	4018	325	91	67,861	64	12,474	25,636	295,190.36
2013	3397	331	85	68,727	46	12,770	24,744	292,238.45
2014	3281	335	84	58,158	45	13,029	20,335	290,777.26
2015	3427	337	84	65,995	43	13,082	21,376	295,430.00
2016	3429	352	86	61,071	41	13,525	20,951	303,997.47

Source: Prepared by the authors according to [21]

Table 4 Employment trends, transport of goods on inland waterways, and GDP, 2005–2014

Year	NER	National (000t)	International (000t)	GDP constant price in HRK
2005	552	195	1251	292,859.83
2006	671	189	1320	306,739.8
2007	714	163	1305	323,522.76
2008	759	141	739	331,155.41
2009	695	127	406	308,305.68
2010	134	145	370	301,214.65
2011	133	91	411	301,214.65
2012	128	50	596	295,190.36
2013	104	42	535	292,238.45
2014	100	102	441	290,777.26

Source: Prepared by the authors according to [21]

Accordingly, the available relevant data for sea transport (see Table 3) and inland waterway transport (see Table 4) were collected from secondary sources.

Data on employment in inland waterway transport (NER) and goods carried in national and international transport are taken from the Croatian Bureau of Statistics, while data regarding the GDP in constant prices are the result of the author's calculations (see Table 4).

5 Research Result and Discussion

Based on the data collected in Tables 3 and 4, correlation analysis was performed first for sea transport (see Table 5) and then for river transport (see Table 6).

The conducted correlation analysis confirms there is a strong and positive correlation between the number of people employed in sea transport and the number of cargo ships ($r = 0.91$; $p < 0.05$) and goods carried ($r = 0.84$; $p < 0.05$). The negative correlation between the number of employees in maritime transport and indicators of labor in passenger traffic can be explained by the under-utilization of the existing capacities, especially off season. This claim is substantiated by the findings of Pupavac, Plazibat, Krcum [22] in their research, which identified a statistically positive and strong correlation between the number of tourist arrivals and sea passenger demand in Croatia ($r = 0.81$; $p < 0.05$). It is the main reason why we focused on the employment effects in the sea transport related to the transport of goods. The second reason is that calculations conducted to determine the value of parameters of the function in the form (2) yielded no conclusive and logical regression models.

To estimate the future number of employees in the sea transport system, a simpler conclusive model has been developed.

$$\text{NES} = 2008.510 + 0.066\text{GC} \quad (R = 0.84; F(1, 10) = 24.758; p < 0.01) \quad (4)$$

The results of correlation analysis for inland waterway transport are shown in Table 6.

Table 6 shows there is a strong positive interdependence between the number of people employed in inland waterway transport and GDP ($r = 0.75$; $p < 0.05$) and a positive but moderate interdependence between the number of people employed in inland waterway transport and national ($r = 0.71$; $p < 0.05$) and international ($r = 0.64$; $p < 0.05$) goods transport.

Regression analysis between the number of employees in inland waterway transport and gross domestic product has resulted in the following model of linear regression:

$$\text{NER} = -4605.97 + 0.02 \text{GDP} \quad (R = 0.75; F(1, 8) = 10.29; p < 0.01) \quad (5)$$

Accordingly, total employment in water transport in Croatia can be expressed as the sum of employment in sea transport and employment in inland waterway transport, or

$$\text{TEWT} = -2597.46 + 0.066\text{GC} + 0.02 \text{GDP}. \quad (6)$$

Table 5 Correlation analysis for sea transport

Correlations (sea transport) Marked correlations are significant at $p < .05$ (000 N = 12 (Casewise deletion of missing data))										
	Means	Std. Dev.	PM	NPS	TM	NCS	PC	GC	GDP	NE
PM	294.3	41.79	1.00	-0.29	-0.51	-0.83	0.78	-0.86	-0.43	-0.84
NPS	86.2	3.07	-0.29	1.00	-0.13	0.27	-0.19	0.23	0.45	0.50
TM	71,770.7	8658.86	-0.51	-0.13	1.00	0.72	-0.22	0.81	0.39	0.50
NCS	59.3	11.62	-0.83	0.27	0.72	1.00	-0.67	0.92	0.44	0.91
PC	12,663.8	529.66	0.78	-0.19	-0.22	-0.67	1.00	-0.56	0.13	-0.67
GC	27,607.9	4689.86	-0.86	0.23	0.81	0.92	-0.56	1.00	0.55	0.84
GDP	303,553.9	12,598.9	-0.43	0.45	0.39	0.44	0.13	0.55	1.00	0.55
NE	3834.7	367.56	-0.84	0.50	0.50	0.91	-0.67	0.84	0.55	1.00

Table 6 Correlation analysis for inland waterway transport
 Correlations (rivers) Marked correlations are significant at $p < .05000$ $N = 10$ (Casewise deletion of missing data)

	Means	Std.Dev.	NER	N	1	GDP
NER	399.0	299.01	1.000000	0.714016	0.643454	0.750118
N	124.5	53.02	0.714016	1.000000	0.695854	0.393733
I	737.4	397.99	0.643454	0.695854	1.000000	0.308091
GDP	304, 321.9	13, 638.05	0.750118	0.393733	0.308091	1.000000

Table 7 Estimate of total employment in water transport by 2027 in Croatia

Year	Goods carried in 000t	GDP constant price in HRK	Employment in sea transport	Employment in inland waterway transport	Total employment in water transport
2016	22,220.60355	304,059.4927	3478.317	394.68	3872.997
2017	22,887.22165	308,620.3851	3522.411	469.69	3992.101
2018	23,573.8383	313,249.6909	3567.828	545.83	4113.658
2019	24,281.05345	317,948.4363	3614.608	623.11	4237.718
2020	25,009.48505	322,717.6628	3662.791	701.54	4364.331
2021	25,759.76961	327,558.4277	3712.419	781.15	4493.569
2022	26,532.56269	332,471.8042	3763.536	861.96	4625.496
2023	27,328.53957	337,458.8812	3816.187	943.98	4760.167
2024	28,148.39576	342,520.7644	3870.417	1027.23	4897.647
2025	28,992.84763	347,658.5759	3926.274	1111.73	5038.004
2026	29,862.63306	352,873.4545	3983.807	1197.49	5181.297
2027	30,758.51205	358,166.5564	4043.066	1284.55	5327.616

Assuming that goods transport in sea transport were to increase by 3% and GDP by 1.5% per year, employment in water transport in Croatia would range as follows:

The implications of COVID-19 crisis are not included in this projection because we have no sufficient data. Also, we believe that negative effects of Covid-19 crisis for GDP moving and total goods transport will be short-term. The assumption that goods transport would increase by 3% in sea transport is based on the fact that in the next decade Croatia should reach the average level of tonnes of goods carried (30,361.24) as in the period from 1996 to 2016. The assumption that the GDP would increase by 1.5% per year seems realistic for the next decade although it is not high enough. Given the above and based on the data in Table 7 it can be concluded that employment in water transport in Croatia by 2027 is expected to increase by slightly more than 1798 jobs, specifically by 614 jobs in sea transport and by as many as 1184 jobs in river transport. Although the assumption about employment increase in sea transport seems realistic, the assumption about employment increase in river transport can be argued as unsustainable. Nevertheless, the obtained information points to the enormous potential of river transport when new workplaces are created in Croatia. To exploit that potential, however, it will be necessary to re-industrialize Croatia, create a single inland waterway network in Croatia, and develop economic cooperation with countries in the region, primarily with Bosnia and Herzegovina, and Serbia.

6 Conclusion

Wide-ranging, frequent, and unpredictable economic, technological, and political changes have marked and aggravated business conditions in the global, European, and Croatian waterway transport markets. As waterway transport is vulnerable to impacts from the global market, its long-term development must be aligned with international business conditions. Employment in Croatia's waterway transport has shrunk 4.37 times relative to employment in the pre-transition period. The reasons behind this decline should be sought in the deindustrialization of the Croatian economy, the lack of singular navigable waterway networks, insufficient integration of Croatian sea ports and river ports, the loss of large shipping companies, and the failure to fully tap into the potential of Croatia's geographic and traffic position. Croatia is below the EU average considering the number of water transport workers. Sea transport prevails in employment in water transport in Croatia, which is understandable considering that Croatia is a maritime country. Due to a series of objective and subjective factors, employment in river transport has become almost irrelevant. The increase of goods transport in sea transport and the growth of the gross domestic product, i.e., an increase in industrial production in river transport, have been recognized as major factors in increasing employment in water transport. Assuming modest economic growth at an average annual rate of only 1.5% and an average increase of tonnes of goods carried of 3%, employment in Croatia's waterway transport system could grow by about 1800 jobs by 2027. The potential for increasing employment in waterway transport in the short run is especially high in river transport. The main preconditions to achieving the above projections are economic growth, the integration of sea ports and river ports, greater international exchange of goods, and cooperation with countries in the region. The rapid development of the maritime economy would spur economic growth, as a special feature of the maritime economy is the fact that its multiplier effect on the development of the land economy is much greater than the multiplier effect that the land economy has on the maritime economy.

References

1. Stražićić, N.: Croatia – a coastal and maritime country. *GeoJournal*. **38**(4), 445–453. <http://www.jstor.org/stable/41146865> (1996) Accessed 3 June 2020
2. Božičević, J., Steiner, S., Smrečki, B.: Evaluation of the Croatian Transport System, ISEP 2008, Ljubljana. https://bib.irb.hr/datoteka/368329.ISEP08_bozicevic.pdf. Accessed 8 May 2019
3. Nikšić, M., Blašković Zavada, J., Golubić, J.: Valorizacija prometnog položaja Republike Hrvatske (in English: Evaluation of transport position of the Republic of Croatia). In: Steiner, S., Koroman, V., Josip, B., Blašković Zavada, J., Nikšić, M., Brnjac, N. (eds.) HAZU-Znanstveno vijeće za promet (2014)
4. Krčum, M., Plazibat, V., Jelić Mrčelić, G.: Integration sea and river ports – the challenge of the Croatian transport system for the 21st century. *NAŠE MORE/Our See*. **62**(4), 247–255 (2015) <https://doi.org/10.17818/NM/2015/4.2>

5. Croatian Bureau of Statistics: Statistical Yearbook of the Republic of Croatia (2017)
6. Pupavac, D., Maršanić, R., Krpan, L.J., Drašković, M.: Analysis and evaluation of employment in the maritime transport system of the Republic of Croatia. Montenegrin J. Econ. **15**(2), 181–193 (2019)
7. Mihaljević, D.: The deindustrialisation process of the Croatian economy. Kurswechsel. **3**, 63–73 (2013)
8. Pupavac, D., Baković, I.: Zaposlenost u prometnom sustavu za 21. stoljeće – skica jedne vizije (in English: Employment in the Transport System in the 21st century – Draft of a Vision). In: Suvremeni promet – Modern Traffic, vol. 37, no. 1–2. Hrvatsko znanstveno društvo za promet, Zagreb (2017)
9. Žuvić, M.: “Jugolinija”: The Myth and the Truth. Trans. Maritime Sci. **05**(01), 69–81 (2016)
10. Pupavac, D., Drašković, M.: Employment in transportation: looking backward and looking forward. In: Barković, D., Runzheimer, B. (eds.) Interdisciplinary Management Research XI. Faculty of Economics Osijek; Hochschule Pforzheim University, Opatija (2015)
11. Malić, A., Badanjak, D., Rajsman, M.: Employment dynamics in the Croatian traffic system. In: Zanne, M., Fabjan, D., Jenček, P. (eds.) ICTS 2005 Transportation Logistics in Science and Practice. Faculty of Maritime Studies and Transport Portorož, Portorož (2005)
12. European Commission: Study on EU Seafarers’ Employment. <https://ec.europa.eu/transport/sites/transport/files/modes/maritime/studies/doc/2011-05-20-seafarers-employment.pdf> (2011). Accessed 22 Feb 2018
13. Rodrigue, J.P.: The Geography of Transport Systems. Routledge, New York (2006)
14. https://unctad.org/system/files/official-document/rmt2020_en.pdf
15. <https://themaritimepost.com/2020/05/21/shipping-fact-global-supply-and-demand-of-seafarers/>. Accessed 20 July 2021
16. https://www.etf-europe.org/our_work/maritime-transport/
17. European Commission: Directorate-General for Mobility and Transport. EU transport in figures: statistical pocketbook 2020. <https://data.europa.eu/doi/10.2832/491038>
18. Platz, T., Klatt, G.: The role of inland waterway transport in changing logistics environment. In: Wiegmans, B., Konings, R. (eds.) Inland Waterway Transport: Challenges and prospects. Routledge, New York (2017)
19. European Court of Auditors: Inland Waterway Transport in Europe: No Significant Improvements in Modal Share and Navigability Conditions Since 2001. Special Report. https://www.eca.europa.eu/Lists/ECA/Documents/SR15_01/SR15_01_EN.pdf (2015)
20. www.mppi.hr. Accessed 15 Mar 2019
21. Croatian Bureau of Statistics: Statistical Yearbook of the Republic of Croatia, different years
22. Pupavac, D., Plazibat, V., Krčum, M.: Modelling transport demands in maritime passenger traffic. NAŠE MORE/Our See. **62**(1), 8–12 (2015) <https://doi.org/10.17818/NM.1.2.2015>

Reversed STIRPAT Modeling: The Role of CO₂ Emissions, Population, and Technology for a Growing Affluence



Johannes Lohwasser , Axel Schaffer, and Tom Brökel

Abstract The presented paper analyzes the relationship between economic growth, demographic development, and CO₂ emissions for 30 industrialized countries using time-series data from 1982–2014 in the well-known IPAT/STIRPAT setting. In contrast to the general assumption of IPAT/STIRPAT modeling, which in most cases proposes a one-way causality running from the anthropogenic factors to the environment, applied Granger-causality tests indicate a reversed causal relationship. Therefore, the paper suggests to add a new perspective to the IPAT/STIRPAT approach by setting up a stochastic model that explains impacts on economic growth (affluence) by regression on population, carbon emissions (as a proxy for energy use or ecosystem services), and technology. The results confirm that GDP per capita growth rates of highly industrialized economies are significantly driven by the development of CO₂ emissions, population, and energy intensity. Coefficients remain robust with or without integrating structural and energy variables and for the short- and long-run perspective.

Keywords STIRPAT · IPAT · Energy · Carbon emissions · GDP per capita · Population

1 Introduction

Despite broad consensus that economic production has substantially altered the global environment, empirical findings on the causal relationship between economic growth and environmental impacts are (at least in some parts) inconclusive. While some authors identify a monocausal relationship running from economic growth to

J. Lohwasser (✉) · A. Schaffer
Bundeswehr University Munich, Neubiberg, Germany
e-mail: johannes.lohwasser@unibw.de

T. Brökel
UIS Business School Stavanger, Stavanger, Norway

the production of anthropogenic greenhouse gases, others find strong evidence for a reversed causality running from environmental emissions to economic growth. Yet others observe bidirectional relationships or no causal link at all. In conclusion, the rich portfolio of empirical studies reveal no universal direction of causality, findings rather depend on the considered time periods as well as countries' stage of development and sectoral structure [1, 2].

Against this background the presented paper seeks to analyze the relationship between economic growth, demographic development, and CO₂ emissions for 30 industrial countries in the well-known STIRPAT (STochastic Impacts by Regression on Population, Affluence and Technology) setting. However, in contrast to the general assumption of STIRPAT modeling, which proposes a one-way causality running from the anthropogenic factors to the environment, applied Granger-causality tests indicate a reversed causal relationship for the sample at hand. Thus, in contrast to existing applications of the STIRPAT model, this paper uses, to our best knowledge for the first time, the STIRPAT framework to estimate environmental impacts on economic growth. This means CO₂ emissions can be, for industrial countries and the time period between 1982 and 2014, considered a driver of economic growth rather than vice versa. Based on these results we suggest to complement the STIRPAT model family by a reversed version that explains stochastic impacts on affluence (rather than on the environment) by regression on population, technology, and environmental impacts or inputs.

The remainder of the paper is organized as follows. Section 2 discusses the general issue of causality and offers a new perspective on the IPAT and STIRPAT modeling. Section 3 continues with methodological remarks followed by the empirical application of the revised model for 30 advanced economies and the discussion of results in Sects. 4 and 5, respectively. Finally, the paper closes with concluding remarks and some brief policy implications in Sect. 6.

2 Perspectives of Causality in the STIRPAT Model Approach

One way to analyze the relationship of anthropogenic factors and the environment is the so-called *IPAT* approach, which presumes that environmental impacts (*I*) are the multiplicative product of population (*P*), affluence (*A*), and technology (*T*) [3]:

$$I = P \cdot A \cdot T. \quad (1)$$

Notably the formula proposes a functional relation between anthropogenic factors and the environment but does not tell us much about the causality of this relationship (e.g., [4]). As a mathematical identity, the equation can be solved for

any variable, e.g., for technology T , defined as environmental impact per unit output (e.g., CO₂ emissions per unit of GDP; [5, 6]) or affluence A (Eq. 2):

$$A = \frac{I}{P \cdot T} \quad (2)$$

Accordingly affluence (typically operationalized as GDP per capita) rises with environmental impacts or inputs (operationalized by CO₂ emissions) and technical progress T (if defined as *decreasing* fossil fuel consumption per unit of GDP).¹ At the same time it decreases with an increasing population P . Or, the other way around, a shrinking population pushes GDP per capita.

While clarity and simplicity certainly add to the popularity of the *IPAT* approach, the pure identity undermines hypothesis testing and causal interpretation (e.g., [4]). This is why [7] suggests to transfer the IPAT equation into the so-called STIRPAT model that explains *Stochastic Impacts on the environment by Regression on Population, Affluence and Technology* and provides the framework for empirical analysis (Eq. 3):

$$I_{i,t} = c_t \cdot P_{i,t}^{\alpha} \cdot A_{i,t}^{\beta} \cdot T_{i,t}^{\gamma} \cdot e_{i,t}, \quad (3)$$

where $I_{i,t}$ is the environmental impact of country i at time t , $P_{i,t}$ is the population, $A_{i,t}$ is the affluence, $T_{i,t}$ is the technology, c_t is the constant, and $e_{i,t}$ is the residual error term. α , β , and γ are the environmental outcome elasticities with respect to population, affluence, or technology, respectively. In order to address the skewness and non-stationarity of variables, STIRPAT models commonly take logs and use first-differences (Eq. 4) (e.g., [8, 9]):

$$\Delta \ln I_{i,t} = \Delta \ln C_t + \alpha \cdot \Delta \ln P_{i,t} + \beta \cdot \Delta \ln A_{i,t} + \gamma \cdot \Delta \ln T_{i,t} + \Delta \ln e_{i,t}. \quad (4)$$

where $\Delta \ln I_{i,t}$ is the change of log CO₂ emissions in country i from time $t-1$ to t , $\Delta \ln P_{i,t}$ is the change of log population, $\Delta \ln A_{i,t}$ is the change of GDP per capita, $\Delta \ln T_{i,t}$ is the change of log technology, $\Delta \ln c_t$ is the change of the log constant, and $\Delta \ln e_{i,t}$ is the change of the log error term.

In contrast to the simple IPAT identity, the very thought of setting up the main STIRPAT equation already implies the assumption of causality. Considering affluence, population, and technology as *key driving forces, contributing factors, predictive or explanatory variables* that *explain, determine, or lead to* environmental impacts further strengthens the underlying assumption of causality (e.g., [4, 9, 10]). After all, it is probably fair to say that the large majority of STIRPAT models assume a one-way causal impact through affluence (typically GDP per capita), population,

¹ For a better traceability the environmental/energetic input is still denoted as I .

and technological progress (measured as environmental impact per output) on the environment (typically CO₂ emissions).

However, this monocausal perspective is not undisputable. *First*, many studies analyzing the relationship between economic growth and the environment propose a bidirectional causality running from economic growth to the environment but also—e.g., through the provision of ecosystem services—from the environment to economic growth (e.g., [11]). This can easily be shown for CO₂ emissions, which are frequently used to illustrate and measure regulative services of the terrestrial ecosystem. Increasing carbon concentration, possibly exceeding the ecosystem's regulative capacity, is not only the result of industrial production but might as well affect production factors and outputs and hamper economic growth (and affluence) in the long-run.

Second, the estimates of CO₂ emissions, probably the most common indicator for measuring environmental impacts within the STIRPAT analysis, generally derive from (fossil) energy consumption. This means they are not only a proxy for environmental impacts but equally reflect the use of (cheap) fossil fuels, which, until now, clearly dominates global energy use.

Empirical findings in this field cannot answer the question of causality unequivocally. Following the “conservation hypothesis,” mainstream economics literature seems to focus on how a growing economy affects energy consumption rather than the other way around. Significant results indicating a causality in this direction can be found for developing and advanced countries (e.g., [12, 13]).

In comparison with this and deeply rooted in Georgescu-Roegen's (e.g., [14]) work on the physical basis of economic production, biophysical economists argue that any production process relies on material and energy inputs (flows), which are transformed by use of human labor, physical capital, and Ricardian land (funds) into production outputs. Thus, the availability of (cheap) energy can be considered a key prerequisite for economic growth and the constitution of the “age of affluence” [15, p. 155]. This so-called growth hypothesis played a minor role in economics for a long time, but gained in importance when some economists could convincingly explain the economic recession in the aftermath of the major oil crises by the declining availability of cheap fossil fuels (e.g., [16]). Since then many empirical studies in this field confirm the idea that energy use drives economic growth (e.g., [17, 18]).

3 Methodological Remarks

Overall, both ways of causality are plausible and there is good reason to assume a bidirectional relationship over a longer-term perspective, as the economy is passing through different stages of development. We therefore propose to check for the direction of causality before setting up the final (STIRPAT) model. One way to do so is the application of the Granger-causality test, which provides valuable insights about the forecasting quality of one variable on another by the help of its past values.

For example, a vector autoregression (VAR) model with two variables y and x allows to test whether, after controlling for past values of y , past values of x help to forecast y [19]. Formally, x Granger-causes y if

$$E(y_t | I_{t-1}) \neq E(y_t | J_{t-1}), \quad (5)$$

where I_{t-1} contains past information on y and x , and J_{t-1} contains only information on past y . Thus, Granger-causality does not mean causality per se and does not imply a contemporaneous causality between variables but rather a variable's feasibility of predicting the other variable according to its past development.

In order to test for Granger-causality most empirical studies either apply time series and cointegration analysis (e.g., [20, 21]) or use VAR models (e.g. [22, 23]). For the paper at hand we follow the VAR approach and estimate a panel vector autoregression (PVAR) model by the cross-sectional series of variables. The general PVAR structure is given by:

$$y_{i,t} = c_i + A y_{i,t-1} + e_{i,t}, \quad (6)$$

where $y_{i,t} = (I_{i,t}, Y_{i,t})'$. $I_{i,t}$ is CO₂ emissions (or population or energy intensity) and $Y_{i,t}$ is GDP per capita of country i at time t . c_i is a country-specific intercept term, A is the coefficient matrix, and $e_{i,t}$ is the residual term. In a next step Eq. (6) is transformed by taking logs and applying first-differences (Eq. 7):

$$\Delta \ln y_{i,t} = A \cdot \Delta \ln y_{i,t-1} + \Delta \ln e_{i,t}. \quad (7)$$

Equation (7) is estimated by the generalized method of moments (GMM) while applying lagged values as instruments. The PVARs include first-order lags according to the Moment Model Selection Criterion (MMSC) and Akaike Information Criterion (AIC).

In case the empirical analysis reveals a monocausal relationship running from anthropogenic factors to the environment, the conventional STIRPAT model (Eq. 4) should be applied. If, however, findings indicate a reverse causality, we suggest to add a new perspective to the STIRPAT approach. By analogy with the transformation from IPAT to STIRPAT [7], a stochastic model could then be based on Eq. (2) and explain stochastic impacts on economic growth (affluence) by regression on population, carbon emissions (as a proxy for energy use or ecosystem services), and technology:

$$A_{i,t} = c_t \cdot P_{i,t}^\alpha \cdot I_{i,t}^\delta \cdot T_{i,t}^\gamma \cdot e_{i,t}, \quad (8)$$

where $A_{i,t}$ is the affluence of country i at time t , $P_{i,t}$ is the population, $I_{i,t}$ is the environmental input (e.g., energy use, availability of energy or ecosystem service measured by CO₂ emissions), $T_{i,t}$ is technology and $e_{i,t}$ is the residual error

term.² α , δ and γ are the economic outcome elasticities with respect to population, environmental input or technology, respectively.

After taking logs and applying first-differences Eq. (8) yields:

$$\Delta \ln A_{i,t} = \Delta \ln C_t + \alpha \cdot \Delta \ln P_{i,t} + \beta \cdot \Delta \ln I_{i,t} + \gamma \cdot \Delta \ln T_{i,t} + \Delta \ln e_{i,t}. \quad (9)$$

where $\Delta \ln A_{i,t}$ is the change of log GDP per capita in country i from time $t-1$ to t . $\Delta \ln P_{i,t}$ is the change of log population, $\Delta \ln I_{i,t}$ is the change of CO₂ emissions, $\Delta \ln T_{i,t}$ is the change of log technology, $\Delta \ln c_t$ is the change of the log constant, and $\Delta \ln e_{i,t}$ is the change of the log error term.

4 Empirical Application

4.1 Granger-Causality, Non-stationarity, and Cointegration

For the empirical part, a balanced yearly cross-country panel dataset of 30 advanced countries from 1982 to 2014 is used. The classification of advanced economies is according to IMF [24]. CO₂ emissions are measured in kilotons and the data stem from Oak Ridge National Laboratory [25]. The variables GDP (in millions US\$2011), population (in millions) and technology [defined as the energy intensity level of primary energy (in MJ per US\$2011)],³ are taken from the Penn World Tables version 9.0 [27] and the World Bank data base [28], respectively.

Following Eqs. (5)–(7), the Granger-causality between CO₂ emissions (environment) and GDP per capita (affluence) is estimated and tested in the first step (results for the underlying PVAR estimations are available upon request). Findings for the logarithmized and first-differenced variables confirm the “growth hypothesis” (with a causality running from CO₂ emissions to GDP per capita). Equally population Granger-causes GDP per capita but not vice versa. With regard to technology, however, Granger-causality only runs from GDP per capita to energy intensity (Table 1). In general, the findings of the Granger-causality test support the idea to consider environmental impacts or inputs a main driving factor for affluence in industrially mature economies rather than vice versa (Eq. 8). As the Hadri Lagrange Multiplier (LM) test, the Im-Pesaran-Shin (IPS) test and the Levin-Lin-Chu (LLC) test suggest that niveau parameters (order of differences: 0) are not stationary but first-differences variables are (results are available upon request), we setup the modified STIRPAT model according to (Eq. 9).

² In contrast to Eq. (2), P and T are not expressed inversely. This does not affect the estimation results.

³ Generally, the STIRPAT studies treat technology differently. This paper uses energy intensity in order to stay close to existing STIRPAT literature [26]. Often, technology is approximated and assumed to be partly captured of the error term.

Table 1 Granger-causality Wald test (Chi2-statistic) based on PVARs (Eq. 7)

GDP per capita → CO ₂ -emission	CO ₂ -emission → GDP per capita	GDP per capita → population	Population → GDP per capita	GDP per capita → Energy intensity	Energy intensity → GDP per capita
1.20	19.30***	0.01	8.82***	9.67***	1.67

***p < 0.01.; H₀: Variable does not Granger-cause the other variable

In addition, the variables are tested for panel cointegration. Cointegration can be interpreted as evidence of a long-run equilibrium relationship between variables (e.g., [29]). In case of cointegration, the evaluation of short-run dynamics between variables by using a first-differences regression should be complemented by the evaluation of long-run dynamics by using error correction models. In order to check for cointegration, the Kao and the Pedroni tests are applied. Most test statistics reject the null hypothesis assuming no cointegration (see appendix, Table A.1). Thus, there is evidence for a long-run cointegrating relationship among economic impacts, carbon emissions, population and structural variables (see next section).

Consequently, both short-run and long-run impacts on economic growth are estimated. In order to evaluate the short-run dynamics, a standard random-effects (RE) estimator is used (estimation of Eq. 9). In order to evaluate long-run dynamics, the fully modified ordinary least squares (FMOLS) estimator is applied. In addition to FMOLS, dynamics ordinary least squares (DOLS) and canonical cointegration regression (CCR) estimators are applied. Results confirm the findings of FOMLS qualitatively. Further, the pooled mean group estimator is used. This approach allows for estimation of short- and long-run dimensions within one error correction model. Results confirm the validation of RE OLS first-differenced results (results are available upon request).

4.2 Reversed STIRPAT

Coefficients are estimated for three slightly different model variations (Table 2). In the first basic setup affluence (GDP per capita) is explained by CO₂ emissions, population, and energy intensity [Table 2, column (1)]. Not surprisingly, and in line with the results of the Granger-causality tests, CO₂ emissions positively and significantly affect GDP per capita. In fact, GDP per capita growth rises by 0.3% when CO₂ emissions growth rises by 1%. In contrast, impacts of population growth have a negative impact on affluence. Further, an increase in energy intensity has a negative and significant effect on GDP per capita. This means that an improvement of energy intensity (i.e., a *decrease* of energy intensity measured in MJ per \$) positively relates to GDP per capita.

In the second model setup, the basic model is augmented by structural variables. In accordance with the most STIRPAT models, we control for the share of urban

Table 2 Determinants of GDP per capita (RE Model)

	(1)	(2)	(3)
Ln GDP p.c.			
Ln CO ₂	0.30***(0.05)	0.28***(0.06)	0.26***(0.04)
Ln Population	-0.56***(0.13)	-0.65***(0.12)	-0.73*(0.42)
Ln Energy Intensity	-0.43***(0.08)	-0.40***(0.08)	-0.39***(0.07)
Ln Urban		0.05(0.44)	-0.33(0.44)
Ln Globalization		-0.18*(0.10)	-0.11(0.09)
Ln Expectancy		0.55***(0.21)	0.60(0.77)
Ln Nuclear			0.04***(0.01)
Ln Renewable			-0.01(0.01)
Constant	0.01*(0.01)	0.02**(0.01)	0.01(0.01)
R ² (<i>within</i>)	0.42	0.42	0.56
R ² (<i>between</i>)	0.58	0.63	0.19
R ² (<i>overall</i>)	0.43	0.43	0.54
Observations	685	685	336
Countries	30	30	15

Robust standard errors clustered at country level in parentheses; Year fixed-effects are included; all variables first-differenced

***p < 0.01, **p < 0.05, *p < 0.1

population (% of total population; [28]) and thus for the effects of an increasing urbanization on economic growth. It can be assumed to have a positive impact on affluence due to agglomeration effects [30]. Furthermore, the impacts of globalization (Globalization Index; [31]) on economic growth are investigated. At least in the long-run, globalization is assumed to have positive effects on economic growth due to various scale and spill-over effects [32]. Finally, we test for possible effects of life expectancy (at birth in years; [28]) on economic growth. Life expectancy is assumed to play a crucial role regarding the so-called quantity-quality trade-off. In this context, educational attainment rises if life expectancy increases. This process affects economic growth (e.g., [33]).

With regard to the size and sign of the coefficients, impacts of the key variables (CO₂ emissions, population, and technology) remain almost unchanged compared to the basic model [Table 2, column (2)]. Further, results show that globalization has a negative impact on affluence in the short-run. In contrast, life expectancy positively and significantly drives GDP per capita. At the same time, we find no significant impact of urbanization.

Assuming that CO₂ emissions reflect the utilization of terrestrial regulation services and fossil energy inputs, affluence might further be affected by the use of other (less carbon intensive) energy sources. For this reason, the third model setup additionally accounts for the share of renewable energy consumption (% of total; [28]) and electricity production from nuclear sources (% of total; [28]). The findings on short-run impacts suggest that the use of (comparatively cheap) nuclear energy positively and significantly relates to affluence [Table 2, column (3)]. With regard to renewable energy, no significant impacts can be observed in the short-run.

Results are confirmed for most variables in the long-run. Interestingly, globalization and renewable energy consumption turn significantly positive (see Appendix, Table A.2). Generally, results are hardly affected qualitatively, if niveau parameters and size effects are taken into consideration (short- versus long-run model).

5 Discussion of Results

The results indicate that GDP per capita growth rates are significantly driven by the development of CO₂ emissions, population, and energy intensity. Coefficients remain rather robust with or without integrating structural and energy variables as well as for the short- and long-run perspective.

In conclusion, the empirical results confirm the “growth hypothesis,” which considers (cheap) energy inputs a key driver of affluence. The positive impact of (comparatively cheap) nuclear energy further supports this hypothesis. In contrast, increasing shares of renewable energy have, in the short-run, no particular effect on welfare. However, results show that renewable energy consumption drives affluence in the long-run. Reasons are, for example, a slow accompanying infrastructure or market accessibility needed for renewable energy sources.

Similar to conventional STIRPAT results, the findings should not be interpreted in a general way but with respect to the underlying country group [10]. This means, the results of this paper particularly hold for advanced economies but not necessarily for other countries. However, it is the governments of the advanced economies that have a particular responsibility to decarbonize their economies and to implement the intended energy turnaround toward renewable energy. This will not necessarily boost the welfare, but as long as prices are reasonably low, switching to renewables will not hamper the economic development either—renewables are more or less growth neutral in the short-run.

Further, findings indicate that population growth has a negative impact on economic growth. This is in line with unified growth theory, according to which positive impacts of a shrinking population on the economy are still visible for advanced industries, even long time after the demographic transition (i.e., process from high to low mortality and fertility rates) has taken place [34].

The findings are less conclusive on the role of technology. On the one hand, there is clear evidence that technical progress (in the form of decreasing energy intensity) relates significantly and positively to GDP per capita. On the other hand, causality analysis indicates that GDP per capita Granger-causes energy intensity rather than vice versa.

Increases in the share of urban population cannot be identified as a significant factor. This does not mean that the degree of urbanization is irrelevant for affluence. Rather advanced countries show generally very high levels of urbanization for the whole period of observation, so further increases might be less important in this case or even hamper economic growth [35].

In contrast, there is evidence that globalization affects economic growth negatively in the short-run and positively in the long-run. The various channels of globalization take time to gain momentum regarding clear positive effects on economic growth. For example, an increasing knowledge acquisition due to globalization cannot immediately translate into research improvements and thus economic growth [36].

Finally, life expectancy significantly and positively affects affluence in the short- and long-run. Existing literature points out that life expectancy increases economic growth due to effects on the age structure or improvements on educational attainment and labor productivity [33].

6 Concluding Remarks

The STIRPAT approach is commonly used to estimate anthropogenic impacts (growing affluence, increasing population, and technological change) on the environment. Today, much of the debate in a continuously developing *STIRPAT* literature is on the choice of the control variables and the relative contribution of an increasing population, economic growth, and technological change to the production of greenhouse gases and other environmental impacts. We largely stay clear of this discussion. Instead, our main interest lies in the causal relationship of the key variables.

The presented paper proposes an alternative extension of the IPAT identity for analysis. Similar to the STIRPAT studies a directional relationship between variables is presumed. However, in contrast to STIRPAT literature and based on a Granger-causality test it seems plausible, at least for the sample at hand, to activate the IPAT identity toward affluence and to estimate stochastic impacts on economic growth (affluence) by regression on population, carbon emissions (as a proxy for energy use or alternatively ecosystem services), and technology. The significant and robust regression results (short- and long-run estimations) with respect to the main variables (CO_2 emissions, population, and energy intensity) in all model variants demonstrate the reasonableness of applying this setup in addition and complementary to the traditional STIRPAT model.

In addition, the findings confirm the ongoing high dependence of advanced economies on the availability and consumption of cheap energy. Breaking the fossil path dependency and decarbonizing the economy, which in light of climate change is without alternatives, could in case of rising energy prices be accompanied with comparatively small growth rates of affluence (if measured as GDP per capita) in advanced economies in the near future. Policies should enhance the use of renewable energy and further support the substitution of non-renewable with renewable energy sources. So, a framework could be created that is able to foster economic growth during the energy transition.

Without doubt, IPAT and particularly STIRPAT modeling has evolved to a powerful tool for illustrating and estimating anthropogenic impacts on the environment.

However, this approach could be extended and also used to identify the relevance of environmental inputs on affluence. Following this line of thought further research might analyze other country groups (e.g., emerging economies) or earlier stages of development of industrialized countries. Furthermore it might be interesting to use other, arguably more inclusive measures of environmental impacts, such as ecological footprints or ecosystem services, rather than fossil energy inputs.

Appendix

Table A.1 Results of the Kao- and Pedroni Cointegration Tests

Kao-test		Pedroni-test	
<i>H</i> ₀ : No cointegration		<i>H</i> ₀ : No cointegration	
<i>GDP per capita, CO</i> ₂ -Emission, Population, Energy Intensity, Urban, Globalization, Life Expectancy (all variables logged)			
Modified Dickey-Fuller t	1.27 (0.10)	Modified Phillips-Perron t	4.78*** (0.00)
Dickey-Fuller t	1.30* (0.09)	Phillips-Perron t	-2.25** (0.01)
Augmented Dickey-Fuller t	1.09 (0.14)	Augmented Dickey-Fuller t	-2.87*** (0.00)

p-value in parentheses; Kao-test assumes a constant cointegration vector; Pedroni-test assumes panel-specific AR parameters; Cross-sectional averages are subtracted. More results regarding cointegration between variables are available upon request

***p<0.01, **p<0.05, *p<0.1

Table A.2 Determinants of GDP per Capita for the Long-run (FMOLS)

Ln GDP p.c.	(1)	(2)	(3)
Ln CO ₂	0.37*** (0.07)	0.44*** (0.06)	0.55*** (0.06)
Ln Population	-0.34*** (0.07)	-0.42*** (0.06)	-0.51*** (0.07)
Ln Energy Intensity	-0.13 (0.10)	-0.20** (0.09)	-0.42*** (0.07)
Ln Urban		-0.09 (0.18)	-0.17 (0.21)
Ln Globalization		1.68*** (0.23)	1.17*** (0.25)
Ln Expectancy		6.18*** (1.58)	3.75*** (1.13)
Ln Nuclear			0.04** (0.02)
Ln Renewable			0.08*** (0.02)
Constant	6.83*** (0.58)	-27.14*** (7.08)	-14.99*** (6.09)
R ²	0.53	0.30	0.06
Observations	551	551	263
Countries	30	30	15

Standard errors in parentheses; Year fixed-effects are included

***p<0.01, **p<0.05, *p<0.1

References

1. Costantini, V., Martini, C.: The causality between energy consumption and economic growth: a multi-sectoral analysis using non-stationary cointegrated panel data. *Energy Econ.* **32**(3), 591–603 (2010)
2. Ozturk, I.: A literature survey on energy–growth nexus. *Energy Policy* **38**(1), 340–349 (2010)
3. Ehrlich, P.R., Holdren, J.P.: Impact of population growth. *Science* **171**(3977), 1212–1217 (1971)
4. York, R., Rosa, E.A., Dietz, T.: STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. *Ecol. Econ.* **46**(3), 351–365 (2003)
5. Raskin, P.: Methods for estimating the population contribution to environmental change. *Ecol. Econ.* **15**(3), 225–233 (1996)
6. Ehrlich, P., Holdren, J.: Impact of population growth. *Popul. Resour. Environ.* **3**, 365–377 (1972)
7. Dietz, T., Rosa, E.A.: Effects of population and affluence on CO₂ emissions. *Proc. Natl. Acad. Sci.* **94**(1), 175–179 (1997)
8. Lohwasser, J., Schaffer, A., Brieden, A.: The role of demographic and economic drivers on the environment in traditional and standardized STIRPAT analysis. *Ecol. Econ.* **178** (2020)
9. Liddle, B.: Impact of population, age structure, and urbanization on carbon emissions/energy consumption: evidence from macro-level, cross-country analyses. *Popul. Environ.* **35**(3), 286–304 (2014)
10. Singh, M.K., Mukherjee, D.: Drivers of greenhouse gas emissions in the United States: revisiting STIRPAT model. *Environ. Dev. Sustain.* **21**(6), 3015–3031 (2019)
11. Guo, X.R., Cheng, S.Y., Chen, D.S., Zhou, Y., Wang, H.Y.: Estimation of economic costs of particulate air pollution from road transport in China. *Atmos. Environ.* **44**(28), 3369–3377 (2010)
12. Akinlolu, A.E.: Energy consumption and economic growth: evidence from 11 SubSaharan African countries. *Energy Econ.* **30**(5), 2391–2400 (2008)
13. Bowden, N., Payne, J.E.: The causal relationship between US energy consumption and real output: a disaggregated analysis. *J. Policy Model.* **31**(2), 180–188 (2009)
14. Georgescu-Roegen, N.: Feasible recipes versus viable technologies. *Atl. Econ. J.* **12**, 21–31 (1984)
15. Hall, C.A., Klitgaard, K.: Energy and the wealth of nations: an introduction to biophysical economics. Springer (2018)
16. Cleveland, C.J., Costanza, R., Hall, C.A., Kaufmann, R.: Energy and the US economy: a biophysical perspective. *Science* **225**(4665), 890–897 (1984)
17. Apergis, N., Payne, J.E.: Renewable energy consumption and economic growth: evidence from a panel of OECD countries. *Energy Policy* **38**(1), 650–655 (2010)
18. Lee, C.C., Chang, C.P., Chen, P.F.: Energy-income causality in OECD countries revisited: the key role of capital stock. *Energy Econ.* **30**(5), 2359–2373 (2008)
19. Wooldridge, J.: Introductory econometrics: a modern approach (with economic applications online, econometrics data sets with solutions manual web site printed access card). MIT press (2015)
20. Lee, C.C., Chang, C.P.: Energy Consumption and GDP revisited: a panel analysis of developed and developing. *Energy Econ.* **29**, 1206–1223 (2007)
21. Stern, D.I.: A multivariate cointegration analysis of the role of energy in the US macroeconomy. *Energy Econ.* **22**(2), 267–283 (2000)
22. Stern, D.I.: Energy and economic growth in the USA: a multivariate approach. *Energy Econ.* **15**(2), 137–150 (1993)
23. Hamilton, J.D.: Oil and the macroeconomy since World War II. *J. Polit. Econ.* **91**(2), 228–248 (1983)
24. IMF: World Economic Outlook, October 2015. International Monetary Fund (2016)

25. Boden, T., Marland, G., Andres, R.: Global, Regional, and National Fossil-Fuel CO₂ Emissions in Trends. Carbon Dioxide Information Analysis Centre (CDIAC), UK (2015)
26. Vélez-Henao, J.A., Vivanco, D.F., Hernández-Riveros, J.A.: Technological change and the rebound effect in the STIRPAT model: a critical view. *Energy Policy*. **129**, 1372–1381 (2019)
27. Feenstra, R.C., Inklaar, R., Timmer, M.P.: The next generation of the penn world table. *Am. Econ. Rev.* **105**(10), 3150–3182 (2015)
28. The World Bank: Population ages 15–64 (% of total), urban population (% of total), energy intensity level of primary energy (in MJ per US\$ 2011), renewable energy consumption (% of total) and electricity production from nuclear sources (% of total). Data retrieved from World Bank Open Data, <http://data.worldbank.org> (2018)
29. Liddle, B.: Consumption-driven environmental impact and age structure change in OECD countries: a cointegration-STIRPAT analysis. *Demogr. Res.* **24**, 749–770 (2011)
30. Turok, I., McGranahan, G.: Urbanization and economic growth: the arguments and evidence for Africa and Asia. *Environ. Urban.* **25**(2), 465–482 (2013)
31. Gygli, S., Haelg, F., Potrafke, N., Sturm, J.E.: The KOF globalisation index—revisited. *Rev. Int. Organ.* **14**(3), 543–574 (2019)
32. Chang, C.P., Lee, C.C.: Globalization and economic growth: a political economy analysis for OECD countries. *Glob. Econ. Rev.* **39**(2), 151–173 (2010)
33. Cervellati, M., Sunde, U.: Life expectancy and economic growth: the role of the demographic transition. *J. Econ. Growth.* **16**(2), 99–133 (2011)
34. Reher, D.S.: The demographic transition revisited as a global process. *Popul. Space Place.* **10**(1), 19–41 (2004)
35. Nguyen, H.M.: The relationship between urbanization and economic growth: an empirical study on ASEAN countries. *Int. J. Soc. Econ.* (2018)
36. Grossman, G.M., Helpman, E.: Globalization and growth. *Am. Econ. Rev.* **105**(5), 100–104 (2015)