Universidad de Huelva

Departamento de Economía

Prediction and interpretation of events through variable relevance analysis in machine learning models

Memoria para optar al grado de doctor
presentada por:

Pedro Cadahia Delgado

Fecha de lectura: 18 de noviembre de 2022

Bajo la dirección de los doctores:

Antonio Aníbal Golpe Moya

Manuel Emilio Gegúndez Arias

Huelva, 2022

UNIVERSITY OF HUELVA

DOCTORAL THESIS

Prediction and interpretation of events through variable relevance analysis in machine learning models

Author:
Pedro Cadahía Delgado

Supervisor:
Dr. Antonio Golpe, Dr.
Manuel Emilio Gegundez

A thesis submitted in fulfillment of the requirements for the degree of Philosophy in the Doctoral Programme: Economics, Business, Finance and Computing Science.

in the

Economics Department

July 18, 2022

Declaration of Authorship

I, Pedro Cadahía Delgado, declare that this thesis titled, "Prediction and interpretation of events through variable relevance analysis in machine learning models" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"We can build a much brighter future where humans are relieved of menial work using AI capabilities."

Andrew Ng

UNIVERSITY OF HUELVA

Abstract

Faculty of Business
Economics Department

Doctoral Programme: Economics, Business, Finance and Computing Science.

Prediction and interpretation of events through variable relevance analysis in machine learning models

by Pedro Cadahía Delgado

Regression and Classification techniques has been studied in Machine learning in order to solve several tasks. In particular, the creation of models based both on predicting and interpretability is one of the central themes of this work. The development of this literature has been possible, due to the necessity of both predicting and describing the reality learnt by statistical models. In this regard, Most representative decision-tree ensemble methods has been used not only for predict events but also to examine the variable importance in order to understand the elements that inform and make it relevant to understand diverse phenomena. However, the literature and the empirical applications are still scarce. For this reason, this thesis tries to empirically analyze these models and to develop new models that allow progress in the understanding of the relationships and relevance between variables in the field of econometrics. Chapter 2 proposes a new accurate model for US economic recessions giving as an output of the work the most important treasury term spreads and rules for US economic recession detection, finding the most relevant term spread found is 3-month–6-month, which is proposed to be monitoring by economic authorities. Chapter 3 proposes a new accurate model in order to study the relevance of price, GDP and affordability as a mechanism for controlling the demand for cigarettes, finding that although the demand functions estimated so far are useful to make predictions about the behavior of cigarette demand, the government must consider that price is a good tool to control tobacco consumption from a certain point of affordability. Besides, Chapter 4 analyze if the EPSs established in Spanish provinces were fulfilled and the anomalies observed in provinces where sales exceed expected values are measured, finding that the provinces in which sales below reasonable values are observed (as detected by the EPSs) present a clear geographical pattern. Furthermore, the values provided by the EPSs in Spain, as indicated in the previous literature, are slightly oversized. Finally, there are regions bordering other countries or with a high tourist influence in which the observed sales are higher than the expected values. Finally, there are regions bordering other countries or with a strong tourist influence in which the observed sales are higher than the expected values. Finally, the chapter presents a new methodology in the field of time series forecasting, creating a disruptive model capable of standing out from other classical models in short-term forecasting. Moreover, this model, based on the prediction of future events based on past data, is able to perform in a non-parametric way more flexible than other more classical linear models.

UNIVERSITY OF HUELVA

Abstract

Faculty of Business
Economics Department

Doctoral Programme: Economics, Business, Finance and Computing Science.

Prediction and interpretation of events through variable relevance analysis in machine learning models

by Pedro Cadahía Delgado

Las técnicas de regresión y clasificación se han estudiado en el aprendizaje automático para resolver diversas tareas. En particular, la creación de modelos basados tanto en la predicción como en la interpretabilidad es uno de los temas centrales de este trabajo. El desarrollo de esta literatura ha sido posible, debido a la necesidad tanto de predecir como de describir la realidad aprendida por los modelos estadísticos. En este sentido, la mayoría de los métodos de conjuntos de árboles de decisión representativos se han utilizado no sólo para predecir eventos, sino también para examinar la importancia de las variables con el fin de comprender los elementos que informan y hacen que sean relevantes para entender diversos fenómenos. Sin embargo, la literatura y las aplicaciones empíricas son todavía escasas. Por ello, esta tesis pretende analizar empíricamente estos modelos y desarrollar nuevos modelos que permitan avanzar en la comprensión de las relaciones y relevancia entre variables en el campo de la econometría. En el capítulo 2 se propone un nuevo modelo preciso para las recesiones económicas en EE.UU. dando como resultado del trabajo los diferenciales de plazos del tesoro más importantes y las reglas para la detección de recesiones económicas en EE.UU., encontrando que el diferencial de plazos más relevante es el de 3 meses a 6 meses, el cual se propone para ser monitoreado por las autoridades económicas. En el capítulo 3 se propone un nuevo modelo preciso para estudiar la relevancia del precio, el PIB y la asequibilidad como mecanismo de control de la demanda de cigarrillos, encontrando que aunque las funciones de demanda estimadas hasta el momento son útiles para hacer predicciones sobre el comportamiento de la demanda de cigarrillos, el gobierno debe considerar que el precio es una buena herramienta para controlar el consumo de tabaco a partir de un determinado punto de asequibilidad. Seguidamente, en el capítulo 4 se analiza si se han cumplido las EPS establecidas en las provincias españolas y se miden las anomalías observadas en las provincias donde las ventas superan los valores esperados, encontrando que las provincias en las que se observan ventas por debajo de los valores razonables (detectados por las EPS) presentan un claro patrón geográfico. Además, los valores proporcionados por las EPS en España, como se indica en la literatura anterior, están ligeramente sobredimensionados. Por último, existen regiones limítrofes con otros países o con gran influencia turística en las que las ventas observadas son superiores a los valores esperados. Por último, en el capítulo 5 se presenta una nueva metodología en el campo de predicción de series temporales, creando un modelo disruptivo capaz de destacar sobre otros modelos clásicos en predicciones a corto plazo. Además este modelo basado en la predicción de sucesos futuros basados en los datos del pasado, es capaz de realizar de una manera no paramétrica más flexible que otros modelos lineales más clásicos.

Acknowledgements

The purpose of this dissertation submitted to the University of Huelva is to shed light on the relationship between innovation and business problem solving in predictive modeling, as well as to unravel the effects of the variables included in such modeling.

This main topic emerged after finishing my Master's thesis at the University of Huelva and International University of Andalusia. Due to my great interest in continuing my learning, collaborating and contributing to the scientific world was my great motivation. Subsequently, I decided to start the present written research under the supervision of Jose Manuel Bravo, Antonio Golpe and Manuel Emilio Gegúndez who introduced me to the different topics exposed in this work, showing me the possibility of innovating and solving problems. My sincere gratitude to them for their help, patience, effort and knowledge.

Likewise, I would like express my gratitude to the Department of Economics of University of Huelva and the Master's faculty staff and colleagues, in particular, Antonio Golpe, Juan Manuel Martín and Jose Manuel Bravo and Manuel Emilio Gegúndez. I could not quantify how much I have learned from all of you. I am also indebted to Jose Carlos Vides for his valuable contributions on a chapter of this dissertation.

Finally, I would like to express my sincere thanks to my friends and my family for their support during this journey, specially to my mother María Dolores Delgado Burgueño, my father Pedro Cadahía Fernandez, my godmother Esther Cadahía Fernandez and great uncle Domingo Cadahía Cicuendez whose constant effort and dedication have shaped me as person.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

The use of Statistics and Mathematics play an essential vital role in researches, it importance comes from the use in data collection to the analysis, interpretation, explanation and presentation. These fields helps researchers in research for proper characterization, summarization, presentation and interpretation of the result of research. As a result, Machine learning(ML) is a field related with statistics and mathematics share common goals by using different techniques (Davidian and Louis, 2012; Dodge and Commenges, 2006).

ML is a branch of artificial intelligence that began to gain importance in the 1980s. It is a type of AI that no longer relies on rules and a programmer, but rather the computer can set its own rules and learn by itself. Machine learning is composed by algorithms. An algorithm is nothing more than a series of ordered steps taken to perform a task.

The goal of machine learning is to create a model that allows us to solve a given task. The model is then trained using a large amount of data. The model learns from this data and is able to make predictions. Depending on the task to be performed, it will be more appropriate to work with one algorithm or another.

The crux of the matter is to clearly define the objective. To solve our problem, then, we will consider what type of task we will have to undertake.

These task can be classified depending on their target variable availability, this is, Supervised learning algorithms base their learning on a previously labeled training data set. By labeling we mean that for each occurrence of the training data set we know the value of its target attribute. This will allow the algorithm to be able to "learn" a function capable of predicting the target attribute for a new dataset.

- Classification problems,when the outcome to be predicted is a categorical attribute, such as spam detection.
- Regression problems, when the outcome to be predicted is a numerical attribute, such as finding out how much a customer will use a certain service (determining a value).

Unsupervised methods are algorithms that base their training process on a dataset without previously defined labels or classes. That is, a priori no objective or class value, either categorical or numerical, is known. Unsupervised learning is dedicated to grouping tasks, also called clustering or segmentation, where its goal is to find similar groups in the dataset.

- Clustering problems, such as recommending a book to a user based on his previous purchases (recommender system)

If we consider the classic customer retention problem, we see that we can approach it from different approaches. We want to segment customers, yes, but what

strategy is the most appropriate? Is it better to treat it as a classification, clustering or even regression problem? It is essential to be clear at all times about the objectives sought by the company when using these techniques, in order to be able to ask the right questions of the data. And, of course, always work with quality data.

Nowadays, The science of ML is used in most industrial operations and companies to plan, take right actions and improve the processes' efficiency. ML refers to the methods and algorithms used to enable computers and systems to learn. The overall goal of ML approaches is to gradually improve the efficiency of computational tasks in terms of accuracy, automation and prediction. ML has been revolutionizing sectors such as medicine, healthcare, manufacturing and banking, among others. As a result, it has already become an important part of contemporary industry (Mottaqi, Mohammadipanah, and Sajedi, 2021).

ML is a field that has a collection of methods that computers can use to predict or improve predictions from data. No matter which task, i.e.: estimation of house prices, product recommendations, street sign detection, credit default prediction and fraud detection, whenever there is data, ML techniques can be helpful to solve in some degree this task. For instance, for predicting the price of a house, the computer may learn patterns from the past sales of houses. The thesis focuses on supervised machine learning, which comprises all the prediction tasks where we have a data set for which we already know the output of interest (e.g., past house prices) and we aim to learn how to predict the output for the incoming data. The goal of supervised learning is to learn a prediction model that links data characteristics (e.g., house size, location, floor type, etc.) to an outcome (e.g., house price). If the output is categorical, the task is called classification, and if it is numerical, it is called regression. The algorithm learns a model by estimating parameters (weights) or by learning structures (trees). The optimal solution is obtained by a loss function that is minimized, as an example, the difference between the estimated home price and the predicted price.

As a result of the advance of the development of this techniques, nowadays in several tasks, humans have been surpassed by machines, such as playing chess or predicting the weather (Granter, Beck, and Papke Jr, 2017; Weyn, Durran, and Caruana, 2019). In addition, another advantage of automating tasks with machines is performance, in terms of speed, reproducibility and scalability. Human learning can take years and is very expensive. In addition, the use of ML can have the drawback that sometimes understanding the solution is complex or has been impossible in some models. For this reason, ML have been used to identify from novel viruses and guess the nature of the virus through the world to the e-commerce business, between other examples (Chatfield and Collins, 1981; Mottaqi, Mohammadipanah, and Sajedi, 2021).

ML techniques have the benefit of generalization compared to traditional prediction models. These techniques can be classified into two main groups: multivariate and univariate. Multivariate models refer to statistical models with more than one dependent variable, these models are usually more accurate as they incorporate more information (Wei, Lu, and Song, 2015).

However, when variables are added to a model, it is not only the prediction that is relevant, but also the interpretation of the model to comprehend "reality". Supervised machine learning models have remarkable predictive capabilities. But how can the model be trusted, and will it work in implementation? What else can it tell you about the world?

As interpretability is a concept, there is no mathematical definition for this. There

are some non-mathematical definitions proposed for interpretability, i.e.: "Interpretability is the degree to which a human can understand the cause of a decision" (Miller, 2019) and "Interpretability is the degree to which a human can consistently predict the model's result" (Kim, Khanna, and Koyejo, 2016). Interpretability of a machine learning model helps human to comprehend why certain decisions or predictions have been made by a model. In this work, the terms interpretable and explainable are interchangeable. However, there are some works (Miller, 2019) that sets the terms interpretability/explainability and explanation as different terms, meaning "explanation" for explanations of individual predictions. Moreover, When a ML model performs well, why do not we just trust the model and ignore why it made a certain decision? "The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." (Doshi-Velez and Kim, 2017).

There are many aspects about why interpretability is a key point into the ML framework:

- **Building trust**: When safety-critical decisions have to be made, e.g. in medical applications, it is important to provide explanations so that the domain expert involved can understand how the model came to its decisions and thus can decide whether to trust the model or not.
- **Failure Analysis**: Other applications, such as autonomous driving might not involve an expert when deployed. But if something goes wrong, interpretable methods can help to retrospectively inspect where bad decisions have been made and understand how to improve the system.
- **Discovery**: Imagine you have an algorithm that can accurately detect early-stage cancer, and on top of that, also prescribe the optimal treatment. Being able to use this as a black-box is great already, but it would be even better if experts could inspect why the algorithm does so well and subsequently gain insights into the mechanisms of cancer and the efficacy of treatments.
- **Verification**: When training ML models, it is often hard to tell how robust the model is (even if the test error is great) and why it does well in some cases but not in others. Especially heinous are so-called spurious correlations: features that correlate with the class you want to predict in the training data, but are not the true underlying reason of why this class is correct. There are great examples of spurious correlations and also cases documented in academia.
- **Model Improvement**: If your model does not well, and you don't know why, it is sometimes possible to take a look at explanations of the model decision and identify whether the problem lies in the data or the model structure.

In predictive modeling, usually there is a trade-off between interpretability and accuracy when building a model. The necessity of interpretability arises from an incompleteness in problem formalization (Doshi-Velez and Kim, 2017), which means that for specific problems or tasks it is not enough to get the prediction (the what). The model must also explain how it came to the prediction (the why), because a correct prediction only partially solves the original problem (Doshi-Velez and Kim, 2017; Miller, 2019). In this regard, an experiment was performed with the aim to measure that explanations are used to manage social interactions (Heider and Simmel, 1944). By creating a shared meaning of something, the explainer influences the actions, emotions and beliefs of the recipient of the explanation.

In many applications with the use of multivariate models, variable importance is something to be taken into account. It is important to note that attention that variable importance analysis (VIA) techniques have been developed independently in many disciplines, a unified review has been performed by several authors (Wei, Lu, and Song, 2015). Accordingly, researchers as well as professionals involved in ML programming may struggle with the complex issues of selecting a relatively small set of significant inputs from a large number of candidate variables, fixing the large set of non-influential inputs to their nominal values while not impacting the prediction accuracy or the uncertainty of the model output, and identifying how a decrease in the uncertainty of each input will impact the uncertainty of the output variable (Saltelli et al., 2008). In general terms, for this work, the most appropriate definition of variable relevance is: "A quantitative indicator that quantifies the change of model output value w.r.t. the change or permutation of one or a set of input variables" (Wei, Lu, and Song, 2015). Independently from the approach, the integration of machines and algorithms in real life requires interpretability to increase social acceptance.

As previously mentioned, artificial intelligence is making headway in companies, although its implementation is still very limited. It brings benefits in all its processes, thanks in part to its interrelation with new technological innovations, such as augmented reality or the Internet of Things.

Artificial intelligence (AI) can be understood as the ability of machines to perform tasks that usually require the involvement of human intelligence. In this sense, AI has helped to reduce the working day.

Although its presence is still very incipient, artificial intelligence is making its way into all business processes. The consulting firm Gartner highlights in one of its latest reports that by 2025 (Rimol and Costello, 2021) the use of AI will be widespread and will lead the technological investment of companies. Its importance in companies and its hegemonic role in the short and medium term is clear.

As is evident, the development of this technology is much more complex than designing a series of protocols for the movement of a machine in a typical production line. Hence, the applications for different industries are also very specific. This slows down its implementation, because it is not as simple as acquiring a new robot or a standard vehicle.

Because of the importance of their development, the possibilities of use in each sector are the subject of attention from the managers of large corporations. However, there is still no clear strategy for its implementation in most companies.

Gradually, we can glimpse what the company of the future will be like in the medium term and how new intelligent technologies will affect the performance of its employees. Designing better strategies, standing out from the competition or getting to know customers and products better are just some of the benefits. The impact on employees' lives and on the reconfiguration of the labor market will not be negligible.

1.1 Contribution of this thesis

This thesis contributes to the mentioned literature through empirical applications, modification of existing models, and analysis and development of new models.

First, in terms of empirical applications, this work has allowed these models to be used in fields of regional analysis to find relevant results in different areas of study such as the field of economic recessions, cigarette consumption.

Second, by applying these techniques and, this work analyzes the meaning of the results found in these models and their relationship with other results of analyses considered key in recession research using treasury terms spreads. Furthermore, it demonstrates and proposes the control of a certain treasury term spreads overlooked in literature and research, that were selected as relevant by a novel model and its application to analyze the usefulness of the results founds.

Third, this work has shown the usefulness of some of the model proposed in a different way from the original point of view. Specifically, using the traditional common factors way of modelling, it has allowed us to understand the cigarette consumption, understanding relevant factors (price, income and affordability) which importance value changes over the time. Additionally, in the same field the important contribution of the first method as alternative for the Empty Pack Surveys, that aids in the detection of anomalies by avoiding the collection of field samples of tobacco packs, using a ML model.

Forth, in the time series field, algorithms are used to study the causal relationship between several variables that change over time and influence each other. From a probabilistic point of view a time series is a succession of random variables indexed by increasing parameter over time. Representing business data as time series often helps companies to visualize business activity. In turn, time series are usually used to predict the future behavior of the measured variable. Time series forecasting is relevant regardless of the field of application, in general, companies generate an efficient management of resources according to the expectations generated by the model or prediction. Efficient forecasting will have a positive impact on making strategic decisions prematurely and therefore being better prepared to make decisions. Every forecast has a scope associated with it, which can be short, medium or long term. In general terms, the following summary table is presented, although the scope varies according to the industry. Besides, this work also offers a contribution in the research of new methods in time series prediction, within the non-parametric methods a new predictor has been offered to be able to perform accurate predictions in short term compared to other existing methodologies. The relative measurement of the error in percentage helps to compare against other models what margin of improvement it is offering against other models, in this way the economic impact on the companies can be measured.

In summary, through an analysis of the most recent models of tree-ensemble methods and some model interpretability techniques of these models, this thesis contributes to the existing literature trying to advance in the knowledge and development of models that take into account the nested structure of the data that occur between different geographic areas, different time periods and variable importance for understanding information that levers some phenomena. In addition, a new methodology in time series forecasting is performed with good results.

1.2 Chapter overview

This work includes the contributions mentioned above in 3 different chapters. Two chapters are devoted to the literature of feature relevance from tee-based ensemble models, while the last chapter is applied to the literature of anomaly detection.

Chapter 2 review most representative decision-tree ensemble methods in order to examine the variable importance of Treasury term spreads to predict US economic recessions with a balance of generating rules for US economic recession detection. A strategy is proposed for training the classifiers with Treasury term spreads data

and compare the results in order to select the best model for interpretability. We also discuss the use of SHapley Additive exPlanations (SHAP) framework to understand US recession forecasts by analyzing feature importance. Consistently with the existing literature we find the most relevant Treasury term spreads for predicting US economic recession and a methodology for detecting relevant rules for economic recession detection. In this case, the most relevant term spread found is 3-month–6-month, which is proposed to be monitoring by economic authorities. Finally, the methodology detected rules with high lift on predicting economic recession that can be used by these entities for this propose. This latter result stands in contrast to a growing body of literature demonstrating that machine learning methods are useful for interpretation comparing many alternative algorithms and we discuss the interpretation for our result and propose further research lines aligned with this work.

Chapter 3 review the literature, where it is commonly accepted that the best mechanism to control smoking is by increasing tobacco prices via taxes. However, there are some studies that indicate that the decrease in tobacco consumption when prices rise is because consumer income is not capable of counteracting said rise. In other words, they associate the decrease in tobacco consumption with a lower affordability of the products and not with the simple fact that prices rise.

The empirical analysis was developed using a panel of data from the Spanish provinces covering 2002 to 2018. By using Machine Learning assembly models, the importance of price, GDP and affordability as a mechanism for controlling the demand for cigarettes is estimated.

The importance of affordability to control tobacco consumption in Spain has grown over time. Furthermore, until 2010, income has generally better explained the demand for cigarettes in the Spanish provinces. However, as of 2010, price is the explanatory variable of the demand function that best explains the behavior of the demand for cigarettes. In these circumstances, the separate estimates of price and income elasticity that have been carried out in Spain so far must be interpreted considering that as of 2010, price is more important than income in explaining the demand for cigarettes. Although the demand functions estimated so far are useful to make predictions about the behavior of cigarette demand, the government must consider that price is a good tool to control tobacco consumption from a certain point of affordability.

Chapter 4 is based on publication (Cadahia et al., 2021), It review the literature that questions the veracity of the studies commissioned by the transnational tobacco companies (TTC) to measure the illicit tobacco trade. Furthermore, there are studies that indicate that the Empty Pack Surveys (EPS) ordered by the TTCs are oversized. The novelty of this study is that, in addition to detecting the anomalies analyzed in the EPSs, there are provinces in which cigarette sales are higher than reasonable values, something that the TTCs ignore.

This study analyzed simultaneously, firstly, if the EPSs established in each of the 47 Spanish provinces were fulfilled. Second, anomalies observed in provinces where sales exceed expected values are measured. To achieve the objective of the paper, provincial data on cigarette sales, price and GDP per capita are used. These data are modeled with machine learning techniques widely used to detect anomalies in other areas.

The results reveal that the provinces in which sales below reasonable values are observed (as detected by the EPSs) present a clear geographical pattern. Furthermore, the values provided by the EPSs in Spain, as indicated in the previous literature, are slightly oversized. Finally, there are regions bordering other countries

or with a high tourist influence in which the observed sales are higher than the expected values. These results are important because they show that cigarette sales in Spain are conditioned by the effect of tourism and by the price differential with border countries. Along these lines, cooperation between countries in tobacco control policies can have better effects than policies developed based on information from a single country. The lack of control over the transactions of tourists and inhabitants of border countries can cause important anomalies that distort the vision that governments have on tobacco consumption based on official data.

Chapter 5 is based on publication (Cadahía and Caro, 2021), in this work revisits the prediction problem in time series framework using a new non-parametric approach. In this approach, the prediction is obtained from a weighted sum of past observed data. These weights are computed by solving a constrained linear optimization problem that minimizes an outer bound of the prediction error. The novelty in our approach consists in considering both deterministic and stochastic assumptions in order to obtain the upper bound of the prediction error. A tuning hyperparameter is used to balance these deterministic-stochastic assumptions in order to improve the predictor's performance. We include a benchmark example illustrating that the proposed predictor can obtain suitable results in a prediction scheme and can be an suitable alternative to existing classical non-parametric methods. Additionally, it is shown how this model can outperform the preexisting ones in a short term forecast.

1.3 Publications

As a result of this dissertation, the following works have been developed:

- **Chapter 2**: Cadahía, P., Golpe, A. A., Vides, J. C.(2021). The yield curve as a recession leading indicator. An application for Gradient boosting and Random Forest. **International Journal of Interactive Multimedia and Artificial Intelligence**.
- **Chapter 3**: Cadahía, P., A., Golpe, A. A., Martín, J. M., Asensio, E. (2021). The importance of price, income and affordability in the demand for cigarettes: A Machine Learning approach for Spanish provinces.**Journal,pages,doi url**.
- **Chapter 4**: Cadahía, P., A., Golpe, A. A., Martín, J. M., Asensio, E. (2021). Measuring anomalies in cigarette sales by using official data from Spanish provinces: Are there only the anomalies detected by the Empty Pack Surveys (EPS) used by Transnational Tobacco Companies (TTCs)?. **Tobacco Induced Diseases 19,1–12,https://doi.org/10.18332/tid/143321**.
- **Chapter 5**: Cadahía, P., Caro, J. M.(2021). Short-term prediction of Time Series based on bounding technique. **arXiv**.

Chapter 2

Chapter 2. The yield curve as a recession leading indicator. An application for Gradient boosting and Random Forest.

2.1 Introduction

Since the decade of the 80's, economic crises have been more recurrent and deeper. In this respect, researchers and practitioners have tried to understand, model and even, predict a recession in different ways. One popular forecasting tool suggested in the literature and followed by economists is the analysis of the slope of the yield curve or the term spread, i.e., the difference between long-term and short-term interest rates (Estrella, 2005a).

According to this idea, in a competitive financial environment, the term structure should respond to international market forces, considered as key for assessing the impact of monetary policy and more importantly, to expression the behavior of the economy. Indeed, if a monetary policy is effective, changes in short-term policy interest rates should impact on long-term ones (Holmes, Otero, and Panagiotidis, 2015). In this sense, the need to forecast and prevent economic recessions has become of great importance to policy makers, practitioners and researchers. In this respect, the use of economic and financial variables as predictive information containers joint to the application of several econometric methods and machine learning models have focused in detect a better accuracy in prediction the possible turning points of business cycle and, in a deeper way, economic recessions (Liu and Moench, 2016). This literature review has tried to shed some light to the more important and highlighted works of the topic.

As previously mentioned, the term structure holds implications in macroeconomics or finance and in the shape of the yield curve, see Shiller and McCulloch, 1990 for a survey. According to this, an upward sloping yield curve suggests that future short-term rates are expected to rise. Contrariwise, a descending sloping yield curve may mean that future short-term rates are expected to drop. The slope of the yield curve –the difference between longer maturity of interest rates and the shorter maturity– gives an important source of information of the real economy evolution (Estrella and Hardouvelis, 1991). Accordingly, it is found that a positive curve slope is associated with future increases in real economic activity when using macroeconomic variables, possessing a significant predictive power or its economic implications in the monetary policy (Weber and Wolters, 2012; Weber and Wolters, 2013). To understand the backgrounds of the term structure, we briefly treat the Expectations

Hypothesis of Term Structure (EHTS). This hypothesis illustrates the relationship between short and long-term interest rates and represents the most influential theory explaining the term structure relations. In fact, this hypothesis establishes that long-term interest rates are defined by an average of the contemporary and expected short-term interest rate (Campbell, 1995). Therefore, this relationship between both types of interest rates indicates that their spread holds meaningful information on future changes in short-term rates and is an important function in the potential effectiveness of monetary policy (Bernanke, Blinder, et al., 1992; Vides, Iglesias, and Golpe, 2018) or reflecting economic agents' anticipations of future events such as recessions (Vetzal, 1994). Regarding this, the inversion of the yield curve is viewed as a consistent predictor of recessions and future economic activity, providing an important reason to explain the flattening or inversion of the yield curve: a monetary lightening (Estrella and Trubin, 2006). A tightening monetary policy would be considered to a rise in short-term interest rates, focusing to a reduction of the inflation. The consequence of the monetary lightening is that the economy may slowdown.

Consequently, shorter-term interest rates are considered as indicators of demand for credit and future inflation so, longer-term interest rates would tend to decrease and flatten the yield curve, being this an example of the relation between the yield curve behavior and recessions. Definitely, the steepness of the yield curve would help us to predict and determine a future recession (Estrella and Mishkin, 1996).

The literature of this topic has tried to demonstrate the role of the term structure or the yield curve as a good forecasting tool for recessions (Poole, Rasche, Thornton, et al., 2002). It should be noted the influential papers (Estrella and Hardouvelis, 1991; Estrella and Hardouvelis, 1990). In these works, they evidenced that the yield curve might be employed to predict real growth in consumption, investment, or aggregate GNP, and more importantly, they demonstrated the relation with NBER-dated recessions. For its part, Dueker et al., 1997 suggests that among different variables used in his work, the term spread is the major predictor of recessions at horizons beyond three months. In this respect, many previous papers have treated the topic by relating the GDP growth with the yield curve slope, see Laurent, 1988; Harvey, 1989; Stock and Watson, 1989; Chen, 1991; Harvey, 1993; Dotsey, 1998; Hamilton and Kim, 2002; Estrella, 2005b; Ang, Piazzesi, and Wei, 2006, among others or Wheelock, Wohar, et al., 2009 for a deep survey of the topic. Another important work argues the convenience of applying models which use the yield curve to predict recessions (Estrella, Rodrigues, and Schich, 2003). In other influential paper in the literature, the term spread is also found as an useful manner in predicting recession even for professional forecaster, a suggestion and combination of the term spread with stock returns in order to measure the accuracy of the latter to predict recessions is performed (Rudebusch and Williams, 2009; Nyberg, 2010). His results were positive and the term spread was found as a useful predictor of recessions for German and US economies. A similar work compared the strength of the yield curve in forecasting recessions (Rudebusch and Williams, 2009; Lahiri, Monokroussos, and Zhao, 2013) with the data used in Rudebusch and Williams, 2009, evidencing the power of the former and suggesting the suitability of using this indicator. For its part, also it is treated the capability of predicting recessions of the term structure and highlighted the power of this indicator over other leading indicators and its strength decreased as a predictor after the financial crisis due to the volatility of macroeconomic variables but unfortunately, its predictive power the last decade decreased (Chinn and Kucko, 2015). Furthermore, Liu and Moench, 2016 in line with the previous literature, find that the ability of the term structure to predict recessions is stronger over the twelve-month horizon when using a similar probit model than

Estrella and Hardouvelis, 1991 or Estrella and Mishkin, 1996 used. Additionally, Evgenidis, Tsagkanos, and Siriopoulos, 2017 further evidenced the potential of the yield curve in forecasting future situations of the US economy over horizons ranging from one quarter to two years. Besides, Gebka and Wohar, 2018 recognized that the yield curve contains information on future GDP growth and that its predictability varies with time, forecast horizons, and quantiles of distribution of future growth nonetheless, an important empirical contribution of their work is that it seems more efficient to predict future expansionary phases, which are more common than recessions, for which the latter appear to perform better. Finally, although Evgenidis, Papadamou, and Siriopoulos, 2020 find that developments in the stock market diminish the efficacy of the yield curve in forecasting future economic activity, they show the fitness of this indicator for predicting economic activity in many most important world economies, such as the US, Canada and Europe and, more importantly, when periods of financial stress are analyzed.

From another empirical perspective, it emerges in the literature the use of techniques based on machine learning algorithms. In this sense, Ng, 2017 claims the suitability of machine learning techniques on central banking or monetary policy issues as it has been applied in other topics of real-life. In this sense, Berge, 2015 demonstrated the yield curve as robust and consistent predictor of economic activity when US business cycle turning points are checked by using four different methods, i.e., equally weighted forecasts, Bayesian Model Averaging (BMA), and linear and nonlinear machine learning boosting algorithms. An important paper in the literature by Gogas et al., 2015 compares different Support Vector Machine (SVM hereafter) and logit models when using the yield curve as a leading indicator, being "the first empirical investigation on the relation between the yield curve and an economy's real output, using an SVM classifier". The model created is useful for policy makers in order to forecast future recessions. In order to reaffirm this latter study, the yield curve is a useful tool for assessing future economic activity, achieving a 100% forecasting accuracy for recessions Gogas, Papadimitriou, and Chrysanthidou, 2015. For its part, Döpke, Fritsche, and Pierdzioch, 2017 demonstrated that the predictive power of boosted regression trees is considerably better than standard probit models. Their findings show that short rates and the yield curve are crucial leading indicators for recession forecast during the 1974-2014 period. Finally, Vrontos, Galakis, and Vrontos, 2021 employ several machine learning methods such as Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net, Discriminant Analysis classifiers, Bayesian classifiers, and classification and regression trees (CART), in line to the existing literature and reveal the ability of the yield curve to act as an early warning system to predict recessions in the United States is reconfirmed. Specifically, the yield curve keeps on a consistent and reliable predictor of recession over the 12-month forecast horizon and Bluwstein et al., 2020 also apply a battery of machine learning methods: decision trees, random forests, extremely randomized trees, support vector machines (SVM), and artificial neural networks, finding that almost all the machine learning models appropriately predict the global financial crisis of 2007-2008 and, additionally, they indicate that the flatter or more inverted the yield curve is, the higher the chance of a crisis, exposing the tendency of chasing performance or increased risk taking that can often be seen before financial crises.

To the best of our knowledge, our approach, i.e., Gradient Boosting and Random Forest Machine Learning methods allows us to reach a better accuracy than in those previous paper in the topic. These Machine Learning algorithms let us to identify the more relevant variables associated to the main variable, that is something that has not been done before in the literature and additionally, we extend time horizon, i.e.,

we update data comparing to previous studies. Indeed, our results indicate that, amongst the term spreads analyzed, our algorithm let us to signal and choose the most influential variables for predicting economic recessions. In this case, highlighting some of the most important term spreads as 3-month–6-month, 2-year–5-year and 5-year–10-year. Furthermore, with respect to these variables, the lift metric is computed in order to detect intervals with higher probability of accounting for a recession, applied into the rules description methods. Results suggests that most important term spread is 3-month–6-month comparing with the term spreads mentioned in literature. Results give some considerations for monetary authorities policymakers and practitioners, such as the monitorization of this term spread above mentioned as a tool for evidencing economic recessions.

The rest of the paper is as follows. Section 2.2 presents the data and methodology used in the paper. Later, section 2.3 show and discuss the results, the concluding remarks are in section 2.4.

2.2 Data and methodology

A supervised method is proposed in order to predict economic crisis cycles and also capable to identify the key factors that levers this phenomenon. Assessing variable importance is an important task, this is reflected in many studies fields, besides there are several approaches that addresses this question (Wei, Lu, and Song, 2015; Yun et al., 2016; Yang et al., 2015; Vladislavleva et al., 2013a).

A decision-tree ensemble classification method is proposed for interpretability rather than only prediction of economic recessions from the different term spread as independent variables. In this way, the variable importance is computed to measure which variables are the most relevant to predict economic crisis cycles. More interpretation of the model is performed by analyzing the dependencies with the most correlated variables and the feature value dependency regarding the target variable in order to get a wider understanding of this phenomenon. Finally, it is proposed a rule extraction process that could be useful for interpreting and detection of economic recession.

2.2.1 Data description

For our empirical analysis, we employ a monthly sample of Treasury Constant interest rates at 9 different maturities over the period January 1969 to November 2020 (amounting 601 observations for each interest rate series). The data corresponds to the 3-month, 6-month, 1-year, 2-year, 3-year, 5-year, 7-year, 10-year and 20-year constant maturity rates.

The data is collected from the Federal Reserve Economic Data (FRED) collected by the Economic Research Division of the Federal Reserve Bank of St. Louis. Since 1-month Treasury Constant maturity rate is only accessible since January 2001, we have picked these maturities considering the availability of consistent interest rate data with the period studied. We reveal 3-month, 6-month and 1-year as short-run; including the latter variable 1-year as short-term because it offers more robustness in our assessment. Conversely, we contemplate the rest of the maturity rates as long term. Table 2.1 shows descriptive statistics related with each interest rate in different maturities. In terms of volatility.

TABLE 2.1: Descriptive statistics for the data.

	M3	M6	Y1	Y2	Y3	Y5	Y7	Y10	Y20
Mean	4.57	4.69	5.08	5.18	5.54	5.84	6.07	6.23	6.31
Median	4.86	4.95	5.27	5.03	5.77	5.97	6.17	6.20	6.01
Min	0.01	0.04	0.10	0.13	0.16	0.27	0.56	0.62	1.06
Max	16.30	15.52	16.72	16.46	16.22	15.93	15.65	15.32	15.13
S.D.	3.41	3.40	3.64	3.78	3.51	3.53	3.23	3.11	3.05

[a] Data from January of 1969 to November of 2020.
[b] M and Y refers to month and year respectively.

From 9 interest rates, 36 spread variables are obtained, the calculation being a subtraction of two elements, this follows a combination without repetition $C(n,r)$, being n and r the set and subset size respectively. As shown in Table 2.1, the interest rates shows similar statistical properties. Nevertheless, the short term interest rates 3-month and 6-month presents lower mean and median and higher standard deviation, on the contrary long term interest rates shows the opposite higher mean and median and lower standard deviation. Henceforth for representing term spread at figures and tables, due to saving space, an abbreviation is used, being M and Y for month and year interest rates respectively, i.e. M3-Y10 for 3-month–10-year term spread.

FIGURE 2.1: Original data interest rates(A) and Computed Term spreads(B).

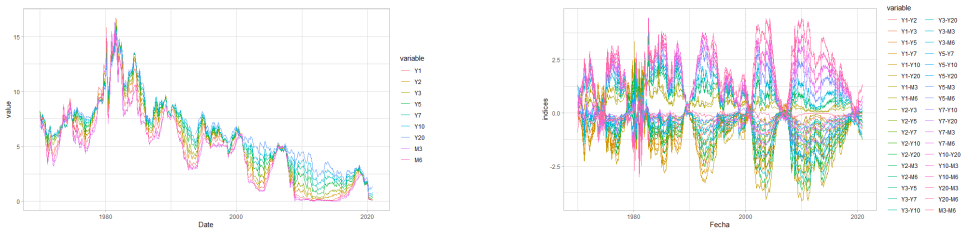
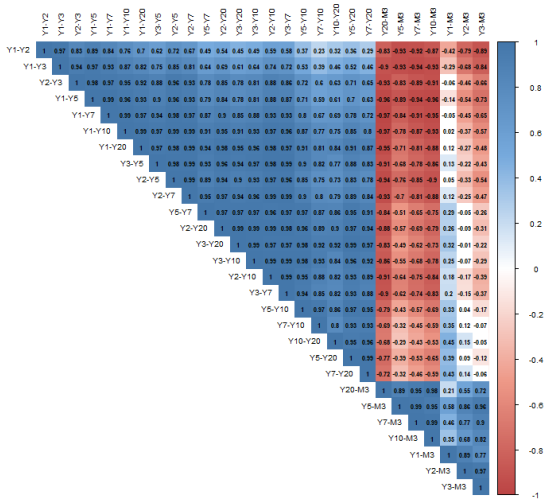


Figure 2.1 presents a plot analysis of the time series traced for all maturities. In this picture the interest rates are plotted where the general trend is decreasing(A). Besides, in the secondary plot(B) shows the computed Term spread for all combinations of interest rates, it is stated that there are some expansion stages with a behavior of divergence and flattening stage where the term spreads are inverted with a behavior of convergence which could be an early indicator of economic recession.

As a result of the combinatory, the term spread variables show several strong correlations. Correlation coefficient is used to verify collinearity, it is argued that collinearity is certain at the 0.9 level of a correlation coefficient or higher (Dohoo et al., 1997). A correlation analysis is shown between variables at Figure 2.1 where the correlation plot shows the coefficients:

FIGURE 2.2: Pearson correlation between term spread variables



Pearson's correlation results at Figure 2.2 shows high correlated features. Aligned with literature, results shows that there is a consistent negative relationship in the difference between long-term and short-term interest rates and consequently in the term spreads (Estrella, 2005a). This is taken into account in order to interpret the results regarding with the features importance exposed in the results.

Literature mainly has been focused on continuous variables whose values, for instance, growth rates in GNP, GDP, industrial production, consumption, investment among others (Estrella, 2005a). In this work only interest rates are used as predictors as the main purpose of this work is not to offer the better predictive model results of literature but understanding the relationships, importance and rules that regarding interest rates with economic recession.

2.2.1.1 Variable Target Lift

In machine learning, Lift is a metric used to assess the performance of a targeting model at predicting or classifying cases as having an enhanced response with respect to the population as a whole. This metric is pretty straightforward to understand, a targeting model is performing good if the response within the target is much better than the average for the population as a whole. In other words, Lift is simply the ratio of these values: target response divided by average response (Tufféry, 2011).

It is defined as:

$$\text{Lift } = \frac{P(A \cap B)}{P(A)P(B)} \tag{2.1}$$

These indicators which are shown at Table 2.2 are useful in the exploratory data analysis stage in order to understand at each variable's decile which range of values of the response variable have more impact on positive target. This can be used as an early exploratory rules for detecting economic recession, this is a complementary information as the decile split does not guarantee the optimal value range for a variable for maximizing the lift, on the contrary, the computed lift for tree base rules ranges may give a better separation as it is a supervised method, for this reason it helps initially to understand this economic processes.

TABLE 2.2: Lift for crisis per Deciles for the most relevant features.

Decile	M3-M6	Y3-M3	Y5-Y10	Y2-Y5	Y2-M6	Y3-Y7
1	1.46	1.09	0.46	0.16	1.52	0.33
2	0.74	1.60	1.14	0.91	0.62	0.87
3	1.20	0.75	0.90	1.40	0.00	1.11
4	0.51	0.53	0.64	1.67	0.91	0.96
5	0.85	1.42	1.63	1.55	1.71	1.26
6	0.77	1.29	0.56	0.78	1.71	0.62
7	0.34	1.42	0.31	0.62	0.62	1.09
8	0.41	0.66	0.71	1.26	0.30	0.33
9	1.88	0.54	0.62	0.30	0.78	1.42
10	1.79	0.75	2.95	1.34	1.83	2.04

[a] Term spread abbreviations contains M and Y for monthly term and yearly term interest rates respectively.

For the sake of simplicity, at Table 2.2 the target lift is computed only for the most important variables as shown in section 2.3. From this table some initial patterns there can be found. Generally, almost for every term spread at high deciles there is a high lift in economic recession with the exception of 3-year–3-month. On the contrary, for the 3-year-3–month and 2-year–6-month term spreads shows high lift either for low and mid deciles, this is an initial indicator due to the nature of higher probability of recession in those deciles, for specific range values the decile's interval table can be found at appendix.

2.2.2 Methodology

The main purpose of this work is not only to offer a model for predicting economic recessions but also to offer a methodology of a good enough model that is able to explain variable importance, dependencies and economic recession detection rules. Decision-tree ensemble methods are supervised learning methods for modeling the relationship between the dependent variable y with the characteristic vector x. Besides, these techniques are a common choice on the actual machine learning research scenario, it has a wide range of applications for regression, classification and other tasks (Ferreira and Figueiredo, 2012; Sun and Pfahringer, 2011a).

The two main decision-tree ensemble methods in bagging and boosting for classification scenario are applied in this work for estimating the economic crisis cycles. The advantage of this methods is that often provides predictive accuracy that cannot be beat, it can optimize on different loss functions and provides several hyperparameter tuning options that make the function fit flexible, generally no data preprocessing required and often works great with categorical and numerical values. To train the models, a training and test data split is performed, where the training set consist on all available variables for all observations from January of 1969 to December of 1999 and the test set comprises from January of 2000 to January of 2020, with the correspondent binary supervised target of economic crisis cycle. In other words, the models should learn which features are relevant in order to predict from an time interval selected for another more recent time interval which should be relevant not only for predicting the economic crisis cycles but also for interpretability of the actual situation.

2.2.2.1 Random Forest Classifier

Random Forest (RF) was proposed by Ho, 1995 as an ensemble method for regression based on individual decision trees, the original classification approach based on Stochastic Discrimination was proposed by Kleinberg, 1990; Kleinberg, 2000.

In this way, Ranger is a fast implementation of RF (Breiman, 2001) or recursive partitioning, particularly suited for high dimensional data. The R implementation Ranger was used to adjust a RF model respectively the considered optimal settings (Wright and Ziegler, 2015).

Which makes Random forest powerful is that builds several weak decision trees in parallel, resulting computationally cheap process, by combining the trees to form a single, strong learner by averaging or taking the majority vote results often to be accurate learning algorithms.

The pseudocode is illustrated at algorithm scheme 1. The algorithm works as follows: for each tree in the forest, a bootstrap sample is selected from S where $S^{(i)}$ is the *ith* bootstrap. Then it is trained a decision-tree as follows: at each node of the tree, instead of examining all possible feature-splits, a random features subsect selection is made $f \subseteq F$, where F is the set of features.

The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F. By narrowing the set of features, it drastically speeds up the learning of a tree.

Algorithm 1: Random Forest Algorithm

Precondition : A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$ being **F** the features and **B** the number of trees in forest.

Function `RandomForest(`S, F`)`**:**
  $\mathbf{H} \longleftarrow \emptyset$
  for $i \in 1, \ldots, B$ **do**
    $\mathbf{S}^{(i)} \longleftarrow$ **A** boostrap sample from **S**
    $\mathbf{h}_i \longleftarrow$ `RandomizedTreeLearns(`$S^{(i)}$, F`)`
    $\mathbf{H} \longleftarrow \mathbf{H} \cup \{h_i\}$
  end
  return H
Function `RandomizedTreeLearns(`S, F`)`**:**
  foreach *node* **do**
    $f \longleftarrow$ small subset of **F**
    Split on best feature f
  end
  return learned tree

RF algorithm is a bagging technique for building an ensemble of decision trees, this technique is known to reduce the variance of the algorithm. Traditionally bagging with decision trees, the constituent decision trees may end up to be highly correlated because the same features will tend to be used repeatedly to split the bootstrap samples. At the same time by restricting each split-test to a small, random sample of features, it is decreased the correlation between trees in the ensemble and improve the performance of the algorithm.

2.2.2.2 Gradient Boosting Machine

The gradient boosting machines (GBM) proposed by Friedman, 2001 is a robust machine learning algorithm due to its flexibility and efficiency in performing regression

tasks Friedman, 2001. The main difference among boosting and traditional machine learning techniques is that optimization is held out in the function space. In other words, the function estimate $\hat{f}$ is parametrized in the additive functional form:

$$\hat{f}(x) = \hat{f}^M(x) = \sum_{i=0}^{M} \hat{f}_i(x) \tag{2.2}$$

In this notation, M is the number of iterations, $\hat{f}_0$ is the initial guess and $\left\{\hat{f}_i\right\}_{i=1}^{M}$ are the function increments, also known as "boosts".

To ensure that the functional approach is achievable in practical terms, a comparable approach to parameterization of the family of functions can be implemented. It is introduced to the reader the parameterized *base-learner* functions $h(x, \theta)$ to differentiate it the overall ensemble functions estimates $\hat{f}(x)$. Different families of basic learners can be chosen, such as decision trees and loss functions.

The *greedy stagewise* approach of function incrementing with the *base-learners* can be formulated.

For the function estimate at the $t - th$ iteration, the optimization function is:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h\left(x, \theta_t\right) \tag{2.3}$$

$$\left(\rho_t, \theta_t\right) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} \psi\left(y_i, \hat{f}_{t-1}\right) + \rho h\left(x, \theta_t\right) \tag{2.4}$$

The optimal *step-size* ρ, should specified at each iteration.

The gradient boosting algorithm proposed by Friedman (Friedman, 2001), can be summed up with the following pseudocode at algorithm 2.

Algorithm 2: Friedman's GBM Algorithm

Precondition :

Input: data$(x, y)_{i=1}^{N}$

- Number of iterations M
- Choice of loss-function $\psi(y, f)$
- Choice of the base-learner model $h(x, \theta)$

Initialize $\hat{f}_0$ with a constant
for $t = 1$ *to* M **do**
  Compute the negative gradient $g_t(x)$
  Fit a new base-learner function $h(x, \theta_t)$
  Find the best gradient descent step size ρ_t:
  $\rho_t = \arg\min_{\rho,\theta} \sum_{i=1}^{N} \psi\left(y_i, \hat{f}_{t-1}\right) + \rho h\left(x, \theta_t\right)$
  Update the function estimate $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h\left(x, \theta_t\right)$
end

The theory and formulation of GBM is available in reference (Friedman, 2001), which interested readers of a deeper explanation can refer to for better understanding of this method.

In this work the so called Extreme Gradient Boosting Training(XGB), proposed by (Chen and Guestrin, 2016) which is a version of GBM was applied as a boosting method for classification with the R library xgboost.

2.2.2.3 Classifier Evaluation

For training the model a data partition was performed, as explained in the aforementioned sections, the predictive accuracy of the models was measured by splitting the data into training and test sets.

The training set comprehends from 1970 to 1999 with 360 instances and a binary target variable with 16% of positives(5 crisis cycles). The test set, comprehends from 2000 to 2020 which are 251 instances with a 14% of positives in the binary target(3 crisis cycles).

As a classification task the error assessment was performed by using the predicted class for the selected models and computing some accuracy metrics from the confusion matrix.

Let $\{P, N\}$ the positive a negative instance class and let $\{\tilde{P}, \tilde{N}\}$ be the predictions produced by a classifier. Let $P(P|I)$ be the posterior probability that an instance I is positive

TABLE 2.3: Classification Metrics for classification model assessment.

Metric	Formula
$Recall(TPR)$	$\mathrm{P}(\tilde{P} \mid \mathrm{P}) \approx \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
$Specificity(TNR)$	$\mathrm{P}(\tilde{N} \mid \mathrm{N}) \approx \frac{\mathrm{TN}}{\mathrm{TN+FP}}$
$Precision(PPV)$	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

[a] TP, P, TN and N refers to true positives, total positives, true negatives and total negatives respectively.

There is no unique metric for assessing a classification task, depending on the characteristics to be evaluated, we consider precision as the most suitable metric for this purpose as considers the positives correctly classified within the observations correctly classified.

2.2.2.4 Model Interpretation

The interpretability of an statistic model helps to understand why certain decisions or predictions have been made, for this reason, measuring variable importance is an important task in many applications. In this sense, this is the era of making machine learning explainable, several authors have conducted an extensive review of methods (Otte, 2013; Wei, Lu, and Song, 2015).

The most common variable importance based has been tested by several researchers using both simulated and real data, this metric tend to be biased in many scenarios (Wei, Lu, and Song, 2015; Strobl et al., 2007a; Strobl et al., 2008). As studied at subsection 2.2.1, there is presence of mutually correlated and/or collinearity, Gini variable importance is expected to be biased (Strobl et al., 2007a; Strobl et al., 2008).

Nevertheless, there is also other classification for interpretability, it could be either local or global, in other words, it is explaining an individual prediction or the entire model behavior (Lundberg et al., 2020).

2.2.2.4.1 SHAP Variable importance

SHapley Additive exPlanations(SHAP), is a model additive explanation approach in which each prediction is explained by the contribution of the features of the

dataset to the model's output (Lipovetsky and Conklin, 2001; Lundberg and Lee, 2017). SHAP comes from game theory field, that is, the solution for the problem of computing the contribution to a model's prediction of every subset of features given a dataset with m features.

A model retraining is required on all feature subsets $S \subseteq F$, where F are all the available features. A value of importance it is assigned to every variable that accounts for the impact on the model's prediction of incorporating that feature. A model $f_{S\cup\{i\}}$ is trained with that feature present and another model f_S is trained with the feature withheld in order to compute this effect. Then, both models predictions are compared on the current input $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)$, where x_S are the values of the input variables in the set S. Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S \subseteq F\backslash\{i\}$. The feature attributions are the computed Shapley values.

They are a weighted average of all possible subsets of S in F:

$$\phi_i = \sum_{S\in F\backslash\{i\}} \frac{|S|!(M-|S|-1)!}{M!}\left[f_{S\cup\{\{\}}\left(x_{S\cup\{\}}\right) - f_S\left(x_S\right)\right] \tag{2.5}$$

SHAP value is the only possible locally accurate and consistent feature contribution values (Lipovetsky and Conklin, 2001; Lundberg and Lee, 2017), they can provide high quality explanation both local and global.

Calculating the importance of the features based on SHAP contributions, the mean of each feature is retrieved for each SHAP matrix. Then, the resulting vectors are summed.

2.2.2.4.2 SHAP Dependence Plots

For every feature and data instance, a point is plotted with the feature value on the x-axis and the corresponding Shapley value on the y-axis, this is the SHAP feature dependence plot.

Mathematically, the plot contains the following points:

$$\left\{\left(x_j^{(i)}, \phi_j^i\right)\right\}_{i=1}^{n} \tag{2.6}$$

SHAP dependence plots are an alternative to partial dependence plots and accumulated local effects. While other methods show average effects, SHAP dependence also shows the variance on the y-axis.

2.2.2.4.3 Rules Extraction

Tree ensembles such as random forests and boosted trees are accurate but difficult to understand. In this work, the interpretable trees framework (*inTrees*) is used in order to extract, measure, prune, select, and summarize rules from a tree ensemble, and calculates frequent variable interactions (Deng, 2019).

Tree ensemble methods consists of multiple decision trees (Breiman, 2001; Friedman, 2001). A rule can be extracted by means of a decision tree's root node to a leaf node. This rule summarization process explained at algorithm 3, is relevant in order to understand and filter the rules for phenomenon interpretability.

Given a rule $\{\mathrm{C} \Rightarrow \mathrm{T}\}$, where C is the condition's rule, being a conjunction of variable-value pairs aggregated from the path from the root node to the current

node, C_{node} denote the variable-value pair used to split the current node, $leaf_{Node}$ denote the flag whether the current node is a lead node, $pred_{node}$ denote the prediction at a leaf node, and T for rule's output.

Algorithm 3: rulesExtract Algorithm

Input: $ruleSet \leftarrow null, node \leftarrow rootNode, C \leftarrow null$
Output: ruleSet

```
Function ruleExtract(ruleSet, node, C):
    if Leaf_Node = true then
        currentRule ⟵ {C ⟶ pred_node}
        ruleSet ⟵ {ruleSet ⟶ currentRule}
        return ruleSet
    end
    for child_i = every child of node do
        C ← C ∧ C_node
        ruleSet ← ruleExtract(ruleSet, child, C)
    end
    return ruleSet
end
```

The method ruleExtract explained at psecudocode Algorithm 3 shows the method used to extract rules from a decision tree. As tree ensembles are multiples decision trees, the final rules are a combination of rules extracted each decision tree in the tree ensemble.

In the following work it is applied the inTrees framework to the data set. For the winning classifier, the ruleExtract method is applied, as a result several rules are extracted and a post processing rules step is performed, this post-processing it is comprised of de-duping rules and rules metrics computation for rules quality. The rule's metrics are: length which is the number of conditions within a rule, support which is the percentage frequency of observations that fulfill the rule, the rule's error for classification tasks which is the number correct classified instances within a rule condition and the target lift (see subsection 2.2.1.1) for every rule as the number proportion of positive targets in the rule condition compared with the variable range.

2.3 Results and Discussion

In this work, a methodology is proposed for understanding the economic recession phenomenon and extracting rules as an early economic recession detection method with a balance of getting a model with a suitable accuracy for prediction which is the main scope of interpretable models in machine learning. This methodology begins with the benchmarking of proposed models in order to get the feature importance for the winning model (see epigraph 2.2.2.4.1). From this step, the main variables that levers the economic recession are detected, by understanding the dependencies with the most correlated variables and the feature value interaction regarding the target variable in order to get a wider understanding of this phenomenon (see epigraph 2.2.2.4.2). To conclude, a rule extraction process is performed for proposing rules useful for early detection of economic recession (see epigraph 2.2.2.4.3).

As first step, two tree-based classification models are fitted to the data, as a result Table 2.4 shows the results for the proposed accuracy metrics for the fitted models.

When assessing the predictive accuracy, the yield curve performs quite well. Additional information can improve its predictive performance Estrella and Mishkin, 1998. Thus, the main purpose of this work is by means of term spreads as unique independent variables to build a model for interpretability with a balance on predictive accuracy.

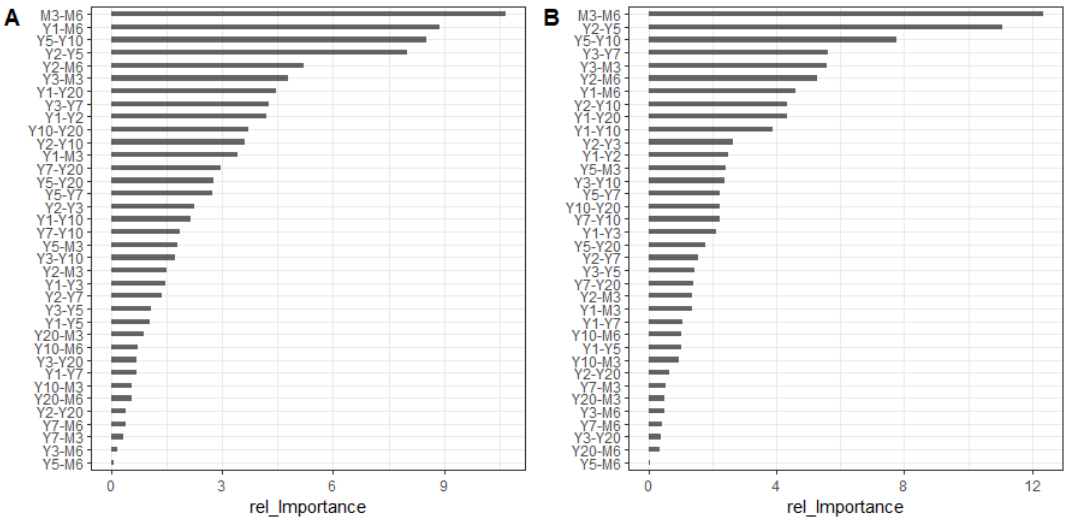
TABLE 2.4: Classification metrics results.

Model	Class	Precision	Recall	Specificity
RF	0	0.88	0.96	0.25
	1	0.52	0.25	0.96
XGB	0	0.96	1.00	0.80
	1	1.00	0.80	1.00

Despite of adding only variables about interest rate nature, suitable classification metrics are obtained by means of term spread variables for predicting economic recession. XGB model has better classification metrics results, for the positive target class, the precision show us how no false positive are obtained, for this reason specificity has also the maximum value. However, recall has a high value but not the maximum showing that there despite of there is a balanced classification of negative and positive labels, the false negatives are present. After fitting and selecting the winning model, the model interpretation of the model for understanding the phenomenon as the most important part of this work comes with the feature importance as the first relevant output in order to interpret which variables are the main predictors for economic recessions. The variable importance is obtained by computing the mean of absolute SHAP value for all instances for every feature at the training and test set, as a result Table A.3 which is at the appendix A is plotted at Figure 2.3 for better understanding. At Figure 2.3, the features are sorted by variable importance in descending order from top to bottom for the most relevant to the less relevant respectively. Besides, by only considering the presence of variables Figure 2.3 shows similar results at the most important variables both in train and test sets, however, as the test set has the more recent data, it is expected to be more representative for future values and may be more accurate in order to extrapolate this information for a near future, due to this, the main analysis is focused in the test set analysis.

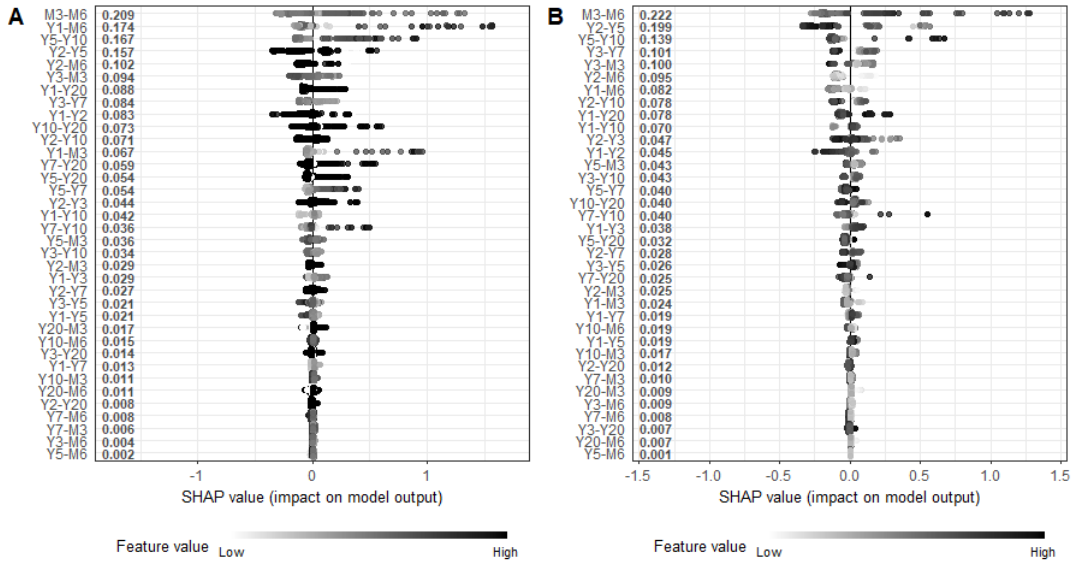
When forecasting economic recession, in previous studies, best results are obtained by taking the difference between two interest rates whose maturities are far apart. Estrella and Mishkin, 1998 suggested that the 3-month–10-year term spread provides a reasonable combination of accuracy and validity in a long term in order to predict economic recessions. However, most term spreads are highly correlated and provide similar information about the economy, so the particular choices regarding the maturity amount mainly to fine tuning process.

FIGURE 2.3: Training(A) and Test(B) SHAP values for the variables



At Figure 2.3 is shown the model's variable importance, results suggests that the most important term spreads are 3-month–6-month, 2-year–5-year, 5-year–10-year, 3-year–7-year, 3-year–3-month and 2-year–6-month. Despite of this work has more recent data than previous studies, literature suggests as rule of thumb that the difference between 10-year and 3-month Treasury rates becomes negative in early the recessions providing a reasonable accuracy and time prevalence (Estrella and Mishkin, 1998). Despite of not having this term spread as the more relevant, most term spreads are highly correlated and provide similar information about the behavior of the economy, so the particular choices with regard to maturity amount mainly to fine tuning and not to reversal of results (Estrella and Mishkin, 1998). The cautionary is that a reference point that works for one spread may not work for other spreads. As an example, the 2-year to 10-year term spread may reverse in advance of the 3-month to 10-year term spread, which it tends to be higher (Estrella, 2005a). In this line, some of the most important variables like 5-year - 10-year term spread which is aligned with the literature statements as could invert earlier than 10-year–3-month term spread.

FIGURE 2.4: Training(A) and Test(B) SHAP contribution values results



SHAP contribution values are plotted for training and test sets at Figure 2.4. This method gives an estimation of an individual sample, due to they are local explainers.

Nonetheless, this can lead for different results as training and test set have different instances, in this case there are small differences between both results. Besides, this plot retrieve additional information about the feature value analysis and the position of the instances on the plot. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction from right to left respectively, vertical location shows the variable importance. Color gradient shows whether that variable is high (dark) or low (light) for that observation.

As argued before, the analysis is focused on test set results, SHAP contribution values analysis could be complementary to decile target lift results at Table 2.2 as it is a preliminary analysis which has not the best splitting method for finding a range with the maximum split. SHAP contribution analysis shows that 3-year–6-month and 5-year–10-year term spreads has higher lift for higher values, which are the 9-10 deciles. The aforementioned term spreads shows this relationship information at the SHAP contribution plot at Figure 2.4, the dark gradient color for instances are at the right side of the plot and the light ones at the left which indicates that high values are associated to positive predictions of economic recession. On the contrary, an opposite behavior is shown on 2-year–5-year, 3-year–7-year, 3-year–3-month and 2-year–6-year spreads, which is somehow aligned with the decile target lift values of Table 2.2, the lower values the higher lift, in other words, higher probability of economic recession. As SHAP contribution plot shows a local interpretability and the decile target lift is not an optimized method for splitting ranges for maximizing lift, these complementary results also may present different nuances at both results due to are different perspective analysis.

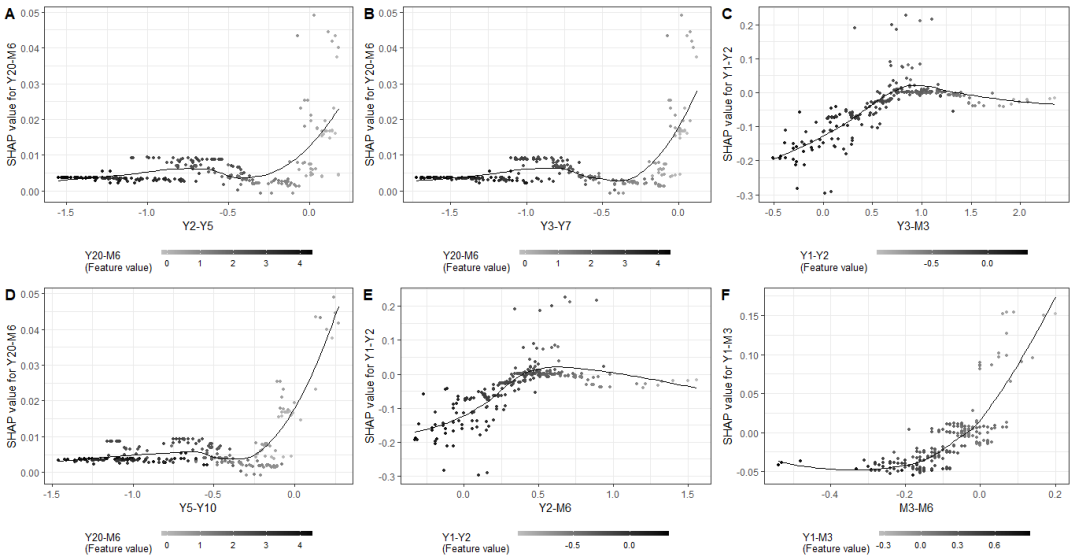
Once detected the main features with more impact on economic recession prediction, the dependence of variables with more important variables on the target variable are studied. Dependence plots has been explained at epigraph 2.2.2.4.2, more information can be found at Lipovetsky and Conklin, 2001; Lundberg and Lee, 2017. In essence, this plot shows feature values of the most important variables on

the x-axis and SHAP values of the most correlated variable on the y-axis, additionally a gradient color to the points by the feature value of the designated variable is added.

For selecting the most correlated variable, with the results from the pairwise Pearson's correlation performed at subsection 2.2.1, by sorting the correlation coefficient, the most correlated feature is selected by most important variable giving as a result Table A.1 at the appendix A. Results suggests that the most correlated variables for the most important ones are in the same time term, for long term time spreads most correlated are long term ones. The relevance of this information is to complement the previous findings with the dependencies of other variables in order to know the dependence and relationship between the most important variables and the most correlated to them, this helps completes the overview of the processes that affect to the economic recessions.

The dependence plot for the most important variables are shown at Figure 2.5 At x-axis, the horizontal location is the actual value from the most correlated variable, and at y-axis, the vertical location shows what having that value did to the prediction. Additionally, the relationship between both information is shown with a loess regression line. For positive slopes, this trends says that the more variable value the higher the model's prediction is for the most correlated variable, with negative slopes it is the opposite.

FIGURE 2.5: SHAP dependence plot for most important variables and their most correlated features



As a result, two kinds of relationships are found: one with positive trend at Figure 2.5 subplots A, B, D and F with a positive slope, having the highest correlation with 20-year–6month and 1-year–3-months term spread respectively. Besides, a the positive trend with an asymptotic behavior at Figure 2.5 subplots C and E is found with a correlation of 1-year–2-year term spread. In addition, the color gradient shows the y-axis feature value from light to dark when variables value is low to high respectively. Generally speaking, the bigger value of the most correlated variables the smaller SHAP value of this variable is. At this point, decile target lift, feature contribution, feature importance and feature dependence are presented, this information let understanding as early indicators which initial range variable values

have more probability to have economic recessions and which variable are the most relevant for economic recession process respectively.

To finalize, at epigraph 2.2.2.4.3 is proposed a methodology for identifying rules for economic recession detection. As a result of rules extraction and initial post-processing, 359 rules are extracted followed by rules metrics, due to saving space, the table is not presented at appendix this can be asked in a document enclosed to this work.

The extracted rules from the winning model can be filtered in several ways, as an initial exploratory study, this work propose a frequency maximization and Lift Maximization criterion for discovering interesting rules. Frequency maximization criterion is when rules are sorted by support at descending order, the first rules are those ones which are the most frequent. Frequency maximization criterion does not sort results by lift, error or length metric for the rules.

TABLE 2.5: Top 5 XGB Max support Rules.

Rule	Error	Length	Support	Lift
$M3 - M6 \leq 0.19$	0.14	1	0.95	1.02
$Y5 - Y10 \leq 0.35$	0.12	1	0.95	0.84
$Y2 - Y5 \leq 0.15$	0.13	1	0.90	0.95
$Y3 - M3 > 0.26$	0.12	1	0.85	0.90
$Y3 - Y7 \leq -0.1$	0.09	1	0.74	0.59

[a] Source is in an enclosed document, rules are obtained by lift and selecting by the presence of top variables from SHAP results.
[b] M and Y are referred for monthly term and yearly respectively.

Table 2.5 shows some rules for Frequency maximization criterion, results show a maximum Support for a rule of 0.95% of observations that satisfy the condition. By analyzing lift criterion, these rules shows values nearly to 1 which is equivalent to say that these rules could guarantee that there is no a special probability of finding an economic recession compared with other data range however, a rule with values near to 0 could show a high probability of not finding an economic recession. As previously explained, XGB is a tree-ensemble model, by means of assembling simple trees making a complex non-linear model, in this way the rules extraction may provide rules with a low level of complexity. Due to this sorting method, the most important rules present low Length, low Lift and error rate, qualifying this as simplistic and inaccurate rules.

By sorting rules by lift at descending order, the first rules are those ones which impacts more on the economic recession detection. Nevertheless, these rules could affect for a little observations, but as recession is a rare event, support for recession identification should be a small percentage.

TABLE 2.6: Top 5 XGB Max lift and support Rules.

Rule	Error	Length	Support	Lift
$Y2 - M6 \leq -0.145 \& Y20 - M3 > 0.79$	0	2	0.01	7.17
$Y2 - Y3 \leq -0.12 \& Y5 - Y10 \leq 0.04 \& M3 - M6 > 0.01$	0	3	0.03	6.32
$Y1 - Y2 > -0.585 \& Y2 - M6 > 1.02$	0	2	0.04	6.32
$Y5 - Y10 > 0.12 \& Y5 - Y20 \leq 0.43 \& Y20 - M6 > -0.66$	0	3	0.04	6.32
$Y3 - M3 > 0.45 \& Y5 - Y20 > 0.22 \& Y5 - M3 \leq 1.32$	0	3	0.04	6.32

[a] Source is in an enclosed document, rules are obtained by lift and selecting by the presence of top variables from SHAP results.
[b] M and Y are referred for monthly term and yearly respectively.

Table 2.6 shows some rules for lift maximization criterion, results show a maximum Lift for a rule of 7.17 times more probability of economic recession for the observations that satisfy the condition comparing the overall observations. Nonetheless, as economic recession is a rare event, these rules usually have a low support due to the nature of economic recession which is a rare event. By this sorting criterion more complex rules are found, with a low error rate and high probability of economic recession, therefore, the more interesting rules may be found. The interpretation of these rules is pretty straightforward, a condition value is presented for every term spread involved in the rule, when this condition is satisfied a support which is the percentage of observation that satisfies this rule is computed with the respective lift.

For the first rule, 2-year–6-month and 20-year–3-month are involved, this also indicates that there is an interaction in the rule between these variables regarding the economic recession detection. Besides, the 20-year–3-month term spread is also an important term spread indicator as may invert earlier than the 3-month–10-year term spread stated as relevant in previous studies (Estrella and Mishkin, 1998).

Regarding the threshold values interpretation, the values are compared with the min, mean and max values for all the historic data for every term spread(see Table A.2 at appendix A) in order to interpret the threshold value as a small, average or big value as those thresholds are closer to any of this feature descriptive statistics, in the case a value is close to two statistics the priority for the average is given. As a result, the first threshold number is labeled as small value and the second as average value, in this way, the qualitative interpretation of this rule will be formulated as follows: "When the 2-year–6-month term spread is lower or equal a small value and 20-year–3-month term spread is greater than the average value there is over 7 times more probability of economic recession than the probability of economic recession for the complementary conditions". Besides, historically this rule fulfilled the economic recessions accounted for 2008 year.

For the second rule, 2-year–3-year, 5-year–10-year and 3-month–6-month are involved which mainly describes that there is an interaction between these variables regarding the economic recession detection. "When the Y2–Y3 and Y5–Y10 term spread is lower or equal of the average value of this term spread and greater than the average value of M3–M6 term spread there is over 6 times more probability of economic recession than the probability of economic recession for the complementary conditions". Besides, this conditions where fulfilled in the economic recessions accounted at 1990, 1991, 2001 and 2008 year.

The other rules from Table 2.6 can be described in the similar way to the previous explained rules, this rules fulfill the conditions of the economic recession accounted at 1980, 1981, 1982, 1974 and 1970 years. This technique allow us having in a small

summary table, a set of rules for detecting economic recession, with a proper data updating and model retraining, this rules can be used in real life and act consequently with economic policies among other uses.

To conclude and sum up the findings, Table 2.7 shows the main results with exception of dependencies analysis results.

TABLE 2.7: Summary Table of empirical results

Rule	Error	Length	Support	Lift	
Variables	Most Correlated	Decile lift	SHAP(+)	Rules Support	Rules Lift
M3-M6	Y1-M3	Low-High	High	✓	✓
Y2-Y5	Y20-M6	Mid-High	Low-Mid	✓	✗
Y5-Y10	Y20-M6	Mid-High	High	✓	✓
Y3-Y7	Y20-M6	Mid-High	Low-Mid	✓	✗
Y3-M3	Y1-Y2	Low-Mid	Low-Mid	✓	✓
Y2-M6	Y1-Y2	Low-High	Low	✗	✓

As a result, main variables on predicting economic recession are detected and the variable dependence with respect the most correlated is studied, the SHAP value for positive economic recession is taken into account with the preliminary information of Decile Target Lift, besides, some of the top rules are presented containing the most important variables and fulfilling the ideas mentioned in this work.

2.4 Conclusion

Regarding the term structure, long-term rates could explain changes in future short-term rates. Understanding the term structure and hence, the yield curve, our goal is to create an interpretable forecasting model that is able enough accurately inform us about future recessions, which could be a useful tool for practitioners, researchers, governments and central banks. for three main groups, the public sector, the private sector which are the households, banks and investors, and the Federal Reserve. From investors point of view, this information could be useful to take right decisions for investing considering different strategies regarding this information, as the expanding economic activity is correlated with the stock market expansion (Cameron, 1978). Federal Reserve by using the term spread to know in advance a possible economic recession could modify the interest rates in order to trying reduce the effect of this phenomenon.

Relevant term spreads are found, 3-month - 6-month, 2-year–5-year, 5-year–10-year, 3-year–7-year, 3-year–3-month and 2-year–6-month. Furthermore, with respect to these variables, the lift metric is computed in order to detect initial intervals with higher probability of accounting for a recession which are complementary to the SHAP contribution values analysis, applied into the rules description methods implementing the necessary policy mix they can dampen the effects of the recession, minimize its duration, or steer the economy away from it all together. As the model provides some false negative alarms, we expect that implementing fiscal and monetary policy in this manner may put some inflationary pressure to the economy.

Finally, the methodology propose a novelty application in this topic by extracting rules for economic recession understanding and detection. With this technique several descriptive conditions allow the user not only to understand this phenomenon but also to have indicators with the goal of detecting in order to minimize the magnitude of the effect of the recession. It is important to note that the yield curve's

predictive power is statistical evidence and that, despite its accuracy, it is impossible to assure the future results. Thus, as the market evolves, we encourage validating and updating these rules with reasonable frequency. The literature suggests that the best predictor of economic recessions for the USA is the 3-month-10-year term spread. Nevertheless, we found that the 3-month-6-month spread is the most relevant for detecting recessions, being included in the main recession detection rules. Therefore, monitoring this spread can be a useful tool for recession identification and could also be a valid indicator for market expectations. In this context, it is found that the best rule, associates this short-term 3-month-6-month predictor with the long-term term spreads, such as 5-year-10-year and 2-year-3-year, illustrating the rule as "When the Y2-Y3 and Y5-Y10 term spread is lower or equal of the average value of this term spread and greater than the average value of M3-M6 term spread there is over 6 times more probability of economic recession than the probability of economic recession for the complementary conditions".

As a future work suggestion, several paths can be followed. On one accuracy side, the improvement of the model predictive accuracy as it is relevant to have tools with high quality and impact on predicting this phenomenon. On the interpretability side, as different exogenous variables can be added, more study on the variable interactions can be performed in order to understand the yield curve inversion with other variables that are relevant for generating policies in order to preventing and controlling. On the rules generation side, as rules are potentially changing over the time as variable importance also may variate as well, a predictive maintenance system could be proposed in order keep rules updated and valid over the time.

Chapter 3

Chapter 3. The importance of price, income and affordability in the demand for cigarettes: A Machine Learning approach for Spanish provinces.

3.1 Introduction

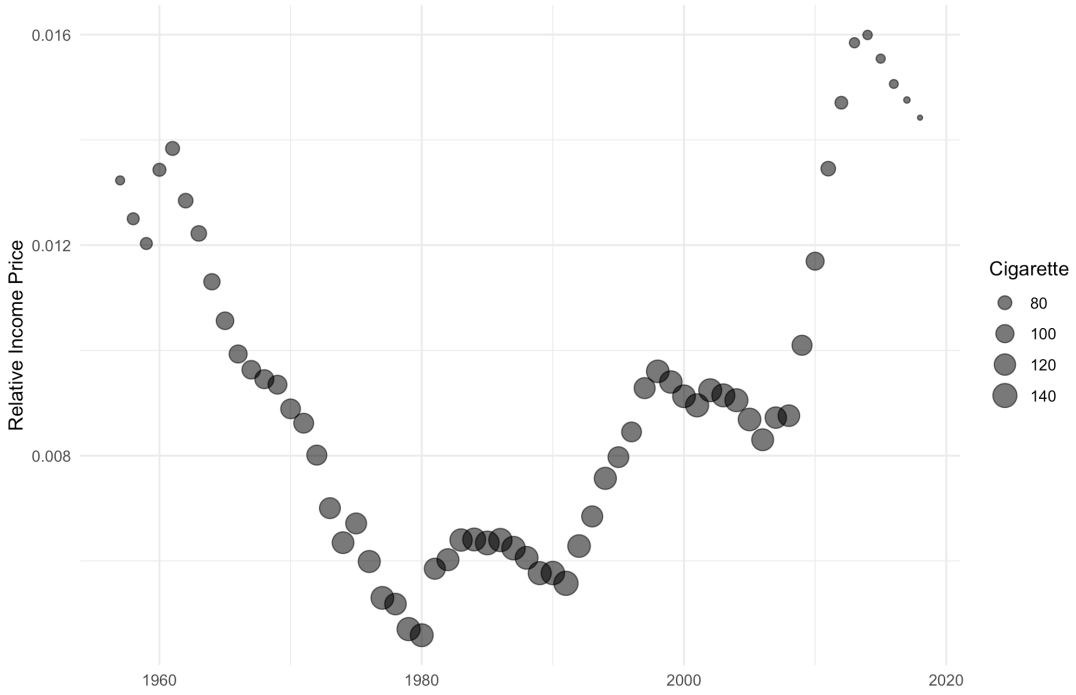
The main determinants in the demand for cigarettes are the price and the income of the consumers (Chaloupka, Yurekli, and Fong, 2012). Along these lines, some studies include price and income as explanatory variables in the cigarette demand function, however, the current trend is a combination of both as an explanatory variable and as a measure of the affordability of cigarettes. This measure of affordability is based on the idea that the purchasing power of consumers to buy cigarettes depends on the price of tobacco products in relation to consumer income (Nargis et al., 2021). In the literature it is commonly accepted that the best mechanism to control smoking is by increasing tobacco prices via taxes. However, there are some studies that indicate that the decrease in tobacco consumption when prices rise is because consumer income is not capable of counteracting said rise. In other words, they associate the decrease in tobacco consumption with a lower affordability of the products and not with the simple fact that prices rise. A recent study indicates that in Spain income is a determining factor that sometimes nullifies the effect of policies when governments use tax hikes to increase the price of cigarettes (JM et al., 2021).

The classical economic model of tobacco demand estimates price and income elasticity separately to measure the effects of price and income changes on tobacco demand. Price elasticity measures the sensitivity of tobacco demand to changes in the prices of tobacco products after adjusting for inflation (real prices), holding real income and other factors constant. Similarly, income elasticity measures the sensitivity of demand for tobacco products to changes in income, holding the real prices of tobacco products and other factors constant. A negative price elasticity of demand for tobacco products indicates that an increase in price causes a reduction in tobacco use, keeping everything else constant, including income. If income increases the demand for tobacco products, tobacco use is not guaranteed to decline after a price increase because the net effect of simultaneous changes in prices and income on the demand for tobacco products will depend on the relative strength of these two effects. In Spain, there is provincial heterogeneity in terms of price elasticity, in some cases exceeding unity in absolute terms in the long term (Almeida et al., 2021). On

the other hand, the income elasticity of cigarettes in Spain shows a marked asymmetry, while the 1% increase in income generates a 0.40% increase in the demand for cigarettes, an economic recession of 1% causes a fall in the demand for cigarettes of 3.60%, ceteris paribus (Álvarez et al., 2020).

The affordability elasticity measures the sensitivity of demand for tobacco products to changes in the price of tobacco products and income growth. Therefore, a negative affordability elasticity would imply that a price increase that exceeds the effect of income growth will lead to a reduction in tobacco use. In this sense, the extremely high values of the price and income elasticities found in Spain may be caused precisely by the lower affordability of tobacco products.

FIGURE 3.1: Evolution of affordability and per capita demand of cigarette in Spain (1957-2018)



There are many studies in which the price and income elasticities of the demand for tobacco are estimated. In addition, given the irruption in the market for new generation products, there are also studies that measure the price and income elasticities of products such as electronic cigarettes or smokeless tobacco (Chaloupka, Straif, and Leon, 2011; Weber and Wolters, 2013; Campbell, 1995). Although price elasticity and affordability elasticity may seem to be similar concepts, it is important to make a thorough analysis of both, because sometimes until a certain level of the relative income price (RIP hereafter) is reached, price policies are not effective.

Estimating the elasticity of the affordability of tobacco is currently a hot topic. However, the birth of this trend is motivated by a primitive study of men's demand for tobacco in Great Britain from 1946 to 1971. In this work, the elasticity of demand for cigarettes with respect to price was estimated as a percentage of annual per capita disposable personal income, known as the "price-income ratio", ranging from -0.44 to -0.58 (Russell, 1973). Although this early study used the "price-income relationship", currently the RIP is used as a measure of affordability (Blecher and

Van Walbeek, 2004). The higher the RIP value, the lower the affordability and vice versa. Blecher and Van Walbeek defined RIP as the percentage of income from the purchase of 100 packs of cigarettes (Blecher and Van Walbeek, 2004). They investigated the relationship between affordability and cigarette smoking by estimating the elasticity of demand for affordability, concluding that there is a negative relationship between RIP and demand for tobacco (FCTC, 2021).

Affordability analysis has become very important today to assess the effectiveness of tobacco tax increases as a measure to control tobacco consumption in accordance with the guidelines for the implementation on literature about price and tax measures to reduce the demand for tobacco (Chaloupka, Straif, and Leon, 2011) under the Framework Convention of the World Health Organization on tobacco control (WHO FCTC).

In this article we analyze the importance of the explanatory variables (price and income) of the basic cigarette demand function of all the Spanish provinces during the period 2002-2018. Next, the RIP is analyzed to find out from what levels of affordability the price policies are effective in Spain. To the best of our knowledge, it is the first time that machine learning techniques have been used to analyze the importance of price, income, and affordability in the demand for cigarettes.

3.2 Data and Methodology

3.2.1 Data

Our empirical analysis was developed using a panel of data from the Spanish provinces covering 2002 to 2018—the year in which the latest data on provincial GDP was published. For cigarette consumption, we used the official annual tobacco sales and the average price of a pack of 20 cigarettes in euros, as published by the Commission for the Trade of Tobacco. The real gross domestic product (GDP) is available from the National Institute of Statistics in Spain. All series employed here are per capita (18 years or older) and expressed in real terms using the consumer price index (CPI base 2016). For a descriptive satistics of the data, see Table B.1 at appendix B.

3.2.2 Empirical Methodology

Based on the theory of demand, cigarette consumption is a function of the real price and per capita of the real income.

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + \varepsilon_t \tag{3.1}$$

Where Q_t is cigarrette consumption, P_t is the real average price, and Y_t is the real GDP per capita. The objective of this work is to identify key factors in this demand function, which is widely used in previous literature, as well as to detect the relationship between these key factors and the affordability of cigarettes. Measuring the significance of variables is an important task, several approaches have been proposed in the literature for addressing this question (Wei, Lu, and Song, 2015; Yun et al., 2016; Yang et al., 2015; Vladislavleva et al., 2013b). To achieve this objective, the measurement of the importance of explanatory variables in a demand function will shed light on how the importance of GDP and price has evolved as explanatory variables of per capita cigarette consumption over time. The quotient between the

importance of price and the importance of GDP, will allow us to observe the evolution of the importance that affordability has had over time as a tobacco control tool.

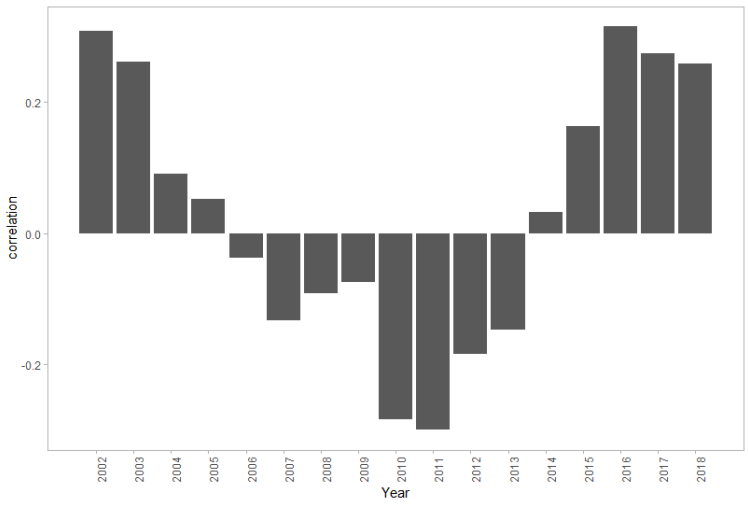
To measure the importance of price and GDP as tobacco control mechanisms, several tree-based ensemble methods have been fitted for finding the best model to interpret results. This model aims to not only predict tobacco sales by means of considering Price and GDP as independent variables, but also to study variable importance estimation through time in the Spanish territory. For this purpose, two datasets have been analyzed, one for a country-level analysis and the other for a provincial level analysis.[1]

In the provincial level dataset, the data is split into a training set and a test set. The training set consists of all the provinces available with the exception of the province whose tobacco sales are being predicted – namely the test set. That is, when possible, all provinces are used to predict sales in a selected province without including that chosen province. After model assessment, the variable importance metrics are computed and studied over the time and at a province level to obtain insights.

The data set at the country-level is divided into three periods, following a recent investigation that indicates that there are two structural breaks in the per capita demand for cigarettes in Spain (1969 and 2006) (JM et al., 2021). The variable importance for these three periods is analyzed by fitting the winning model for each period. To analyze the importance of variables, it must be taken into account the fact that the explanatory variables cannot be highly correlated (Dohoo et al., 1997). The collinearity issue has been proposed by means of a Correlation coefficient and Variance Inflation Factor(VIF), Dohoo et al., 1997 collinearity is present at the 0.9 level of a correlation coefficient or higher. Multicollinearity can assess by computing the VIF, which measures the inflation of the variance of a regression coefficient due to multicollinearity. Furthermore, VIF values variables should be higher than 5 or 10 (Lin, 2008; Gareth et al., 2013).

A correlation analysis by year between Price and GDP has been performed, Figure 3.2 shows the results of correlation coefficients:

FIGURE 3.2: Pearson correlation between Price and GDP at Spanish province level by year.



[1]For data time issues two datasets has been employed, as the province level dataset has information available from 2002-2018, the country level has a wider time range.

To complement this, the VIF test results show low values (1.41), so, with VIF and correlation analysis we assume that there is no correlation and collinearity between variables. This is relevant to the present study, which shows the methodology and confirms several requirements for correctly interpreting the variable importance.

Once the correlation between explanatory variables has been analyzed, the next methodological step is to present some joint tree-based methods to model the relationship between a dependent variable and the characteristic vector x. An estimation of models in which the quantiles of the response are modelled to depend on the features is presented by Koenker and Bassett Jr, 1978 which is the Quantile Regression. This method is based primarily on choosing a model for the conditional quantile, in contrast, the minimal squares estimate the conditional mean. Based on the imposed assumptions, the choice of parametric or non-parametric is available (Zhao, 2008; Engle and Manganelli, 2004). The conditional $\alpha - quantile\ q$ of a scalar variable Y, $P(Y \leq q \mid I) = \alpha$ where the probability $0 < \alpha < 1$ is given and I denotes an information set generated by independent variables X, for more details see Koenker and Bassett Jr, 1978.

In order to measure the importance of explanatory variables of the cigarette demand function, in this work two main methods based on trees are proposed. On the one hand, a hybrid of Random Forest (RF) and QR has been presented by Meinshausen and Ridgeway, 2006, this is the Quantile Regression Forest (QRF) approach. It should be remarked that a key difference between RF and QRF is that QRF for each node of each tree maintains the values of all observations of the node, but RF only keeps the mean of the observations found in the node (Meinshausen and Ridgeway, 2006). The quantreg Forest and ranger library, an implementation of RF (Breiman, 2001) in R software was used to fit a QRF and Ranger model respectively with the default settings (Wright and Ziegler, 2015). On the other hand, an implementation of Gradient Boosting Machine for quantile regression has been selected (Friedman, 2001) due to its flexibility and efficiency in performing regression tasks (Freund, Schapire, and Abe, 1999). A quantile version of the GBM has been selected which is the gradient boosted quantile regression (GBQR). This method has been applied in several fields (Ferreira and Figueiredo, 2012; Sun and Pfahringer, 2011b). The accurateness of GBM predictions comes from increasingly refined approximations, which are carried out by adding tree-based models together. For a better understanding of this method, in terms of its theory and formulation of GBM consult the following works: (Friedman, 2001; Freund, Schapire, and Abe, 1999). The R library GBM was used to fit the quantile version of GBM with the default settings.

To evaluate the model performance, a data partition was performed, the predictive accuracy of the models was measured by splitting the data into training and test sets. In the training set, the out-of-bag estimation error (OOB) was performed in order to get an unbiased estimate of the test set error in tree-based-ensemble methods (Breiman, 2001).

Every tree is made through bootstrap samples from data, in a one-third proportion, using cases out of the bootstrap sample for the kth tree construction. In about one-third of the trees a test set classification is obtained for each case. At the end of the run, take j to be the class that got most of the votes every time case n was OOB. The proportion of times that j is not equal to the true class of n averaged over all cases is the OOB error estimate. This has proven to be unbiased in many tests.

Given the $0.5 - quantile$ predicted responses ($\hat{y}_i$) and the actual values(y_i) for the training and test set, the error is computed as $e_i = y_i - y_i$, using the metrics for the following Table.

TABLE 3.1: Predictive performance metrics.

Metric	Formula
Prediction error	$e_i = y_i - \hat{y}_i$
Mean squared error	$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2$
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert e_i \rvert$

[a] For further understanding of these formulas and its statistical properties, see (Hyndman and Koehler, 2006a).

Tree-based ensemble methods were selected to quantify the importance of explanatory variables in predicting cigarette consumption. These methods were chosen due to their evident suitability for said objective. Specifically, these methods estimate the importance of an independent variable X_j due to the decrease in predictive precision when the values of X_j are randomly permuted.

Several variable importance metrics (VIM) based on tree-based-ensembles can be found in the literature, the more traditional Variance decrease VIM(VDVIM) and Gini VIM(GVIM)(Hyndman and Koehler, 2006a) and other metrics such as permutation VIM(PVIM) (Wei, Lu, and Song, 2015) and conditional permutation VIM (CPVIM)(Strobl et al., 2007b). Furthermore, GVIM, VDVIM and PVIM metrics have been benchmarked in several studies not only in simulated but also in real data (Wei, Lu, and Song, 2015; Strobl et al., 2007b). Specifically, when continuous explicative variables are mutually uncorrelated and collinearity is not present among them, GVIM/VDVIM is expected to produce better results than PVIM (Wei, Lu, and Song, 2015; Strobl et al., 2007b). As a result of the variable analysis, VDVIM may lead to better results over PVIM due to the lack of correlation and collinearity. Either for classification or regression models, VDVIM and GVIM are popular VIM methods, for further details of the GVIM method see literature (Hyndman and Koehler, 2006a).

The VDVIM computation at tree level is that at each node N, the choice of splitting variable X from a set of split-variable candidates, as well as the splitting criteria, are based on maximizing the decrease of the metric of this node N:

$$\text{VDVIM}(N) = \frac{1}{|S|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 - \left(\frac{1}{|S_t|^2}\sum_{i\in S_t}\sum_{j\in S_t}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|s_f|^2}\sum_{i\in_f}\sum_{j\in S_f}\frac{1}{2}(x_i - x_j)^2\right) \tag{3.2}$$

S is being the set of pre-split sample indices, S_t the set of sample indices for which the split test is true, and Sf the set of sample indices for which the split test is false... Each of the above summands are indeed variance estimates, though, written in a form that does not directly refer to the mean (Wright and Ziegler, 2015).

For GBM, it computes at each split in each tree the improvement in the split-criterion (MSE). Then, it averages the improvement made by each variable X_i across all the trees that the variable X_i is used. The variables with the largest average decrease in MSE are considered most important (Friedman, 2001; Freund, Schapire, and Abe, 1999).

3.3 Results

The results are presented in three parts. First, the results of the importance of variables in the cigarette demand function in Spain from 1957 to 2018 are shown. Next, the results of the provincial analysis of the importance of variables in the cigarette

demand function are shown. Finally, the relationship between affordability, measured by the RIP, and the importance of the explanatory variables of the demand function for cigarettes in Spain is shown.

FIGURE 3.3: Importance of price and GDP in the 3 sub-periods.

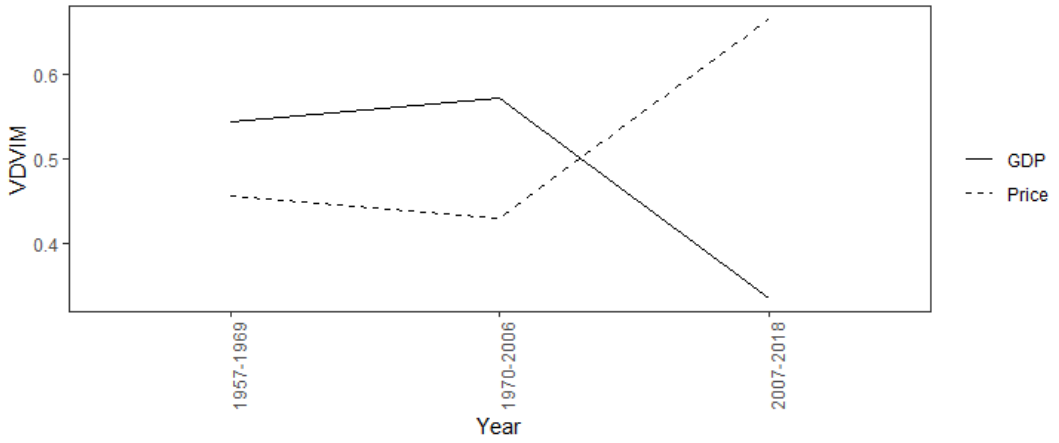


TABLE 3.2: Importance of price and GDP in the 3 sub-periods.

Period	Model	GDP importance	Price importance	MAE(OOB)	MSE(OOB)
1957-1969	QRF	54.43%	45.57%	0.31	0.14
1970-2006	QRF	57.14%	42.86%	0.64	0.49
2007-2018	QRF	33.50%	66.50%	0.51	0.22

For measuring, at national level, the changing importance of variables for the periods, as a result in Table B.3 it is detected the *QRF* as the winning model, which is applied in the 3 different periods for measuring variable importance, according to (JM et al., 2021). The results are shown at Figure 3.3 from Table 3.2 that comes from a mean from values at the density plot that can be found at the appendix B in Figure B.1. Variable importance of three fitted models for the established periods are plotted at Figure 3.3. Results show how variable importance changes over time. As can be seen, until 2006 the main driver of the demand for tobacco in Spain was income. However, during the period 2007-2018, price is the main explanatory variable of the demand function for cigarettes in Spain. These results are similar to what a recent study established; it is suggested that income in Spain canceled out the effect of increases in the price of cigarettes until 2006 (JM et al., 2021).

A province granularity level analysis is performed, year-averaged importance of Price and GDP metrics was computed for every model. We can see some patterns over time in Table 3.3. Results show a Price relevance increase from 2007 over the importance of the GDP variable, 2010 being the major switch over year. As can be seen, until 2010 (except in 2004) GDP is the variable that most determines cigarette consumption. However, as of 2010 (except in 2013 and 2014) it is Price that is the most relevant variable in determining cigarette consumption.

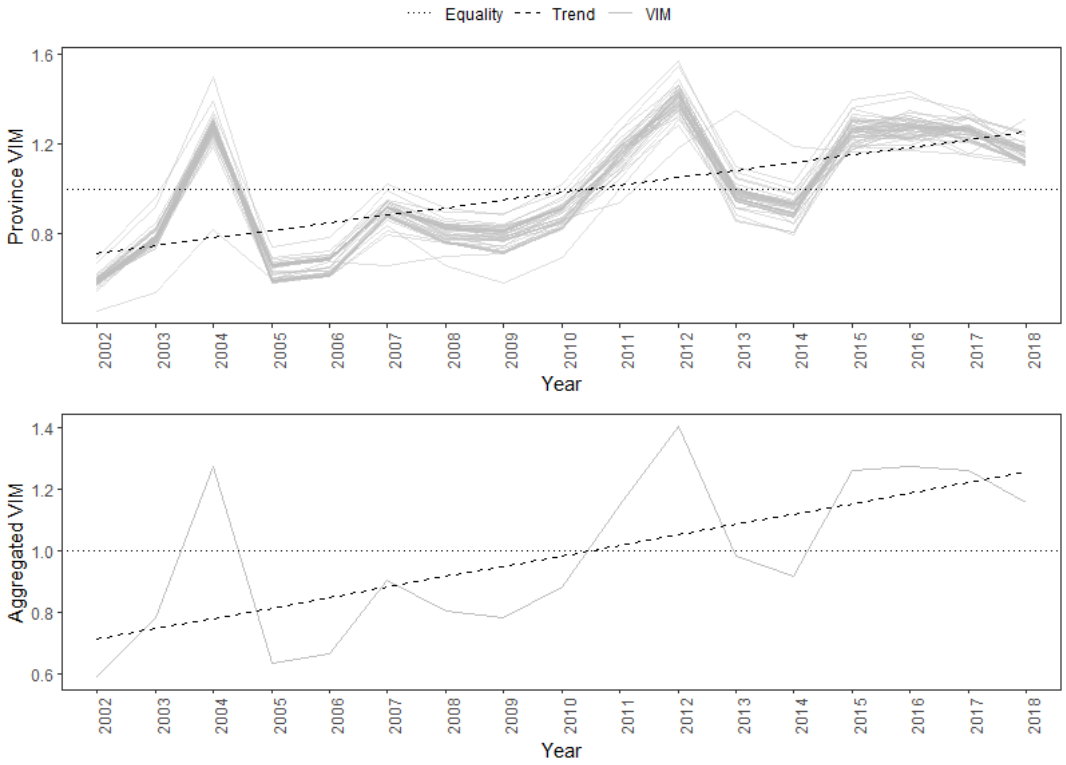
TABLE 3.3: Variable Importance of GDP, Price, and Affordability for Province Data.

Year	Model	GDP VIM	Price VIM	Affordability VIM	MAE(OOB)	MSE(OOB)
2002	QRF	62.88%	37.12%	0.59	0.1185	0.5135
2003	QRF	56.06%	43.94%	0.78	0.1547	0.6898
2004	QRF	44.05%	55.95%	1.27	0.1459	0.6685
2005	QRF	61.08%	38.92%	0.64	0.1515	0.5008
2006	QRF	60.04%	39.96%	0.67	0.1496	0.5879
2007	QRF	52.57%	47.43%	0.90	0.1742	0.8271
2008	QRF	55.42%	44.58%	0.80	0.1817	1.163
2009	QRF	56.08%	43.92%	0.78	0.1604	1.0105
2010	QRF	53.07%	46.93%	0.88	0.1453	0.9682
2011	QRF	46.56%	53.44%	1.15	0.1542	0.9481
2012	QRF	41.66%	58.34%	1.40	0.1576	1.1917
2013	QRF	50.50%	49.50%	0.98	0.1617	0.9383
2014	QRF	52.18%	47.82%	0.92	0.1391	0.8922
2015	QRF	44.27%	55.73%	1.26	0.1518	0.9684
2016	QRF	43.97%	56.03%	1.27	0.1478	0.9961
2017	QRF	44.29%	55.71%	1.26	0.154	1.0343
2018	QRF	46.36%	53.64%	1.16	0.1655	1.1451

To have a more precise vision of the evolution of the importance of the main explanatory variables of the cigarette demand function, the relationship between Price and GDP has been included in Table 3.3. This Price Importance/GDP Importance ratio (Affordability Importance) can be calculated to simplify the results and the results are more interpretable:

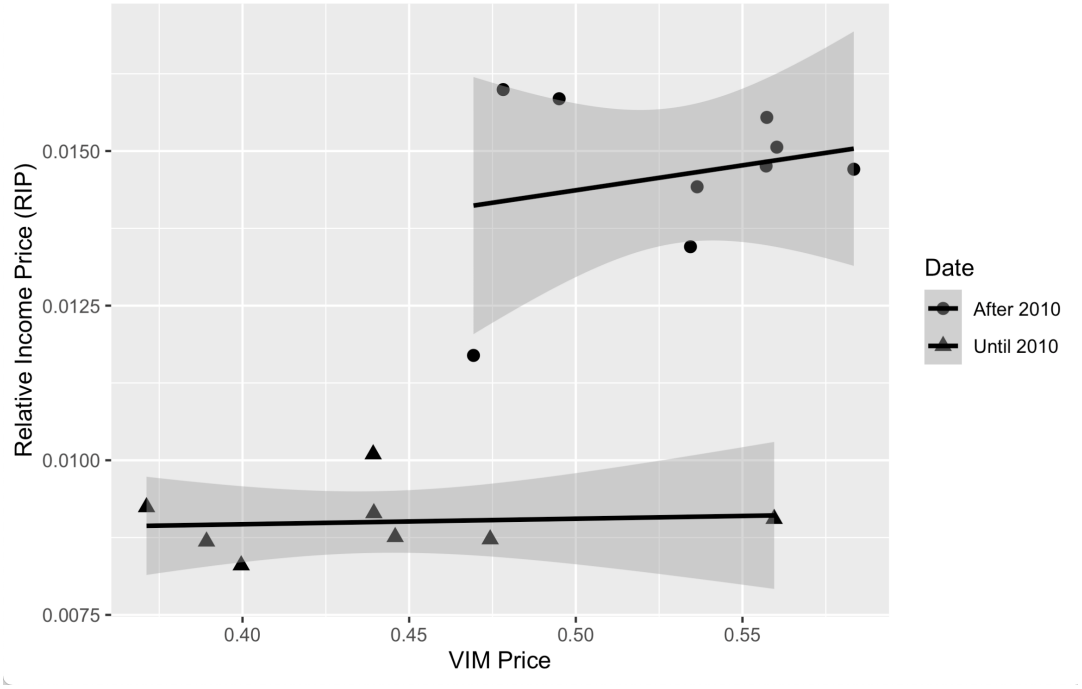
- $AffordabilityImportance < 1$: more relevance on GDP.
- $AffordabilityImportance = 1$: equally relevant.
- $AffordabilityImportance > 1$: more relevance on Price.

FIGURE 3.4: Evolution of Affordability Importante from 2002 to 2018 at Province Level and Aggregated Territory.



The Affordability Importance in Figure 3.4 shows a provincial granularity level at the top of the figure. In this sense, the patterns of all the provinces are aligned with a growing trend since 2007 and with a ratio greater than 1 as of 2010. On the other hand, Figure 3.5 at the bottom shows the aggregate mean of the Affordability Importance, which also shows an increasing trend at the national level of the ratio. These results seem to suggest that until 2010 price policies have not had the desired effect on public health. However, as of 2010, Price is more important than GDP in explaining the demand for cigarettes. Figure 3.1 shows how from 2010 there is a notable growth of the RIP, which may be related to the indicated behavior. In other words, it seems that from a certain threshold of RIP the price policies are more relevant to the aim of minimizing cigarette consumption.

FIGURE 3.5: Correlation plot between RIP and VIM Price.



Finally, Figure 3.5 shows the last relevant finding of this paper. Although it has been indicated that as of 2010 (when the RIP is around 1 percent) price is the explanatory variable that dominates the demand function for cigarettes in Spain, it seems reasonable to ask: is the RIP related to the importance of price in the demand function for cigarettes? In Figure 3.5 it is observed how until 2010 the relationship between RIP and VIM Price does not present a trend. However, as of 2010, a slight positive trend appears to be observed in the relationship between the RIP and the VIM. It seems that from a certain RIP threshold there is a slight positive relationship between affordability and the importance of price as an explanatory variable of demand for cigarettes.

3.4 Conclusions

In recent years, the concern about knowing which factors drive the demand for tobacco has grown due to the deleterious effect that tobacco has on public health. Based on data from a provincial panel on cigarette sales in Spain, the results of this article are adjusted to the existing evidence that price and income are variables capable of explaining the demand for cigarettes. The results also confirm that when measuring the effectiveness of tobacco price increases in reducing demand, it is important to consider the effect of income growth that can offset the effect of cigarette price increases. In other words, the findings of this work suggest that in times of economic growth, the price increases required to effectively reduce tobacco consumption in the population would be greater than the increases required under conditions of slow or no economic growth.

The main contribution of this article lies in indicating that the importance of affordability to control tobacco consumption in Spain has grown over time. Furthermore, until 2010, income has generally better explained the demand for cigarettes

in the Spanish provinces. However, as of 2010, price is the explanatory variable of the demand function that best explains the behavior of the demand for cigarettes. In these circumstances, the separate estimates of price and income elasticity that have been carried out in Spain so far must be interpreted considering that as of 2010, price is more important than income in explaining the demand for cigarettes. This means that, although the demand functions estimated so far are useful to make predictions about the behavior of cigarette demand, the government must consider that price is a good tool to control tobacco consumption from a certain point of affordability. In other words, for the Spanish government, the price is a more powerful way to control tobacco consumption from 2010 onwards. To our knowledge, this is the first attempt to obtain estimates of the explanatory power of the main elements of the tobacco demand function.

The recommendation for researchers that emerges from this analysis is that price and income elasticity must be contextualized to make effective decisions. In many cases, the conclusions are based on whether the equality of the price and income elasticity parameters is statistically rejected. Nevertheless, whether the suggestion is rejected that the price and income elasticity parameters are the same or not, up to certain levels of affordability the price is not a totally effective tool to control tobacco consumption.

Affordability, measured by the RIP, is a highly relevant concept in tobacco control. This work shows how price policies are more effective beyond a certain level of affordability. However, although the level of affordability does not allow price to have a significant effect on tobacco use, this concept is an important part in the formulation of tobacco control policies. The concept of affordability involves explaining the combined effects of simultaneous changes in prices and income to policy makers and the relevance of adjusting prices according to income growth and inflation. Price elasticity as a tobacco control tool is a very widespread concept, however, although the ceteris paribus clause is an appropriate starting point, the simultaneous effect of economic growth on tobacco consumption must be considered. The concept of affordability reinforces why policy makers have difficulty understanding why in some cases the effect of price on demand is more pronounced than in others.

Chapter 4

Chapter 4. Measuring anomalies in cigarette sales by using official data from Spanish provinces.

4.1 Introduction

Some theoretical and empirical works have questioned the Empty Pack Surveys (EPSs) because they are commissioned by the transnational tobacco companies (TTCs) and their methodology and validity are not certain (Rowell, Evans-Reeves, and Gilmore, 2014). In this context of non-independence of the EPS, it has generated a multitude of papers that have analysed the relationship between what the TTCs show regarding the illicit tobacco trade (ITT) and the official data published by the governments. In addition to the EPSs, the TTCs make reports, usually annually, about ITT. In this line, some studies conclude that the reports made by TTCs require greater transparency, external scrutiny, and the use of independent data (Gilmore et al., 2014). Another issue criticized by some studies is the funding and dissemination by ITT's research TTCs through corporate social responsibility initiatives. In this context, a study concludes that if TTCs data on ITT cannot meet the standards of accuracy and transparency established by high-quality research publications, a solution may be to tax the TTCs and administer the resulting funds to experts independent of the tobacco industry, using previously developed reliable models to measure ITT (Gallagher et al., 2019; Stoklosa, 2016).

In this context of non-independence, many studies have proposed methodologies, using official data, to measure the illicit tobacco market5. In this part of the literature there are many results achieved. Some studies conclude that industry-funded estimates inflate likely levels of illicit cigarette use (Chen et al., 2015; Miera Juarez et al., 2021). Other papers indicate that industry warnings against tax increases, based on illicit trade rates, in certain countries are not justified (Maldonado et al., 2020; Gallego et al., 2020; Paraje, 2019).

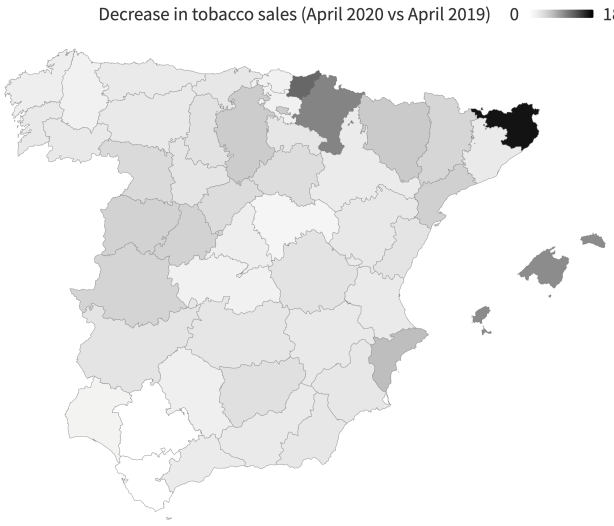
Given this academic trend that doubts about the suitability of using EPSs as an indicator of illicit tobacco trade, there are many studies that have focused on analyzing how illicit trade impacts the health of the population, as well as the policies implemented by the governments. Some studies suggest that tobacco tax policies to control the prevalence of smoking and national health care expenditures should be accompanied by a greater effort to curb smuggling activities across borders(Schafferer et al., 2018). On the other hand, some studies have analyzed the impact on the illicit trade of plain tobacco packaging (Haighton, Taylor, and Rutter, 2017; Gomajee et al., 2021). Finally, there are some studies that have carried out

a kind of EPSs parallel to the one commissioned by the TTCs to verify its veracity (Kurti et al., 2019; Chan et al., 2020).

Although there are many studies that have made an effort to contrast the EPSs with official data or by conducting parallel surveys, all of them have focused on analyzing whether the data provided by the TCCs regarding the rates of illicit trade are true. However, a recent study indicates that actual smoking prevalence sometimes exceeds the estimated actual consumption derived from aggregated data on official sales (Jakob, Cornuz, and Diethelm, 2017). This means that there are border areas in which the prevalence of smoking is underestimated because official data do not consider what smokers buy in areas with an attractive price differential. In this way, other work also indicates that the excessive production of cigarettes suggests a possible excess supply of cigarettes in some countries, probably diverted towards illicit trade (Leite Ribeiro and Conceição Pinto, 2020). Studies that contrast the veracity of EPSs focus on verifying whether the rates of illicit trade in a given country are real. However, what is indicated in the works cited in this paragraph highlights the need to also study excess sales rates of products at the borders of certain countries, which can be sold illegally in other countries with an attractive price differential.

Thus, to simultaneously study the veracity of EPSs and excess sales in border areas with another country that has a higher tobacco price, it is necessary to study a country with border countries with lower and higher prices. Spain is a border country with France and Gibraltar, two countries with which it maintains a price differential by excess and by default. In addition, a recent study indicate that distortions are observed in the borders of Spain with France and Gibraltar (Almeida, Golpe, and Álvarez, 2020; Shang, Dauchy, and Feldman, 2020). Although the cited study indicates that there are distortions in the border provinces with France and Gibraltar, it is important to know the magnitude of these distortions. The current health crisis caused by COVID-19 caused that the borders in Spain were closed during the month of April 2020, the effect on cigarette sales being shown in Figure 4.1. As can be seen, in some provinces sales decreased by up to 180%, while in other provinces tobacco sales did not decline. Therefore, focusing the study on Spain seems reasonable if we want to simultaneously analyze the veracity of EPSs and excess sales in border areas.

FIGURE 4.1: Year-on-year drop in tobacco sales (April 2019 - April 2020)



In this context, our study analyzed the two components of the anomalies in official tobacco sales, that is, both the provinces in which official sales are lower than expected and those in which sales are higher than fair value. There is no evidence in the literature that clarifies the territorial anomalies that are observed in both directions. To the best of our knowledge, this study is the first to analyse, simultaneously, whether the provisions of the EPSs are fulfilled, contrasting it with official data and, furthermore, which provinces show sales above reasonable values.

4.2 Data and Methodology

4.2.1 Data

Our empirical analysis was developed using a panel of data from the Spanish provinces from 2002 to 2017 - last data of the provincial GDP published correspond to the year 2017 -. For cigarette consumption, we used the annual tobacco official sales and the average price of a pack of 20 cigarettes in euros, as published by the Commission for Trade of the Tobacco. The real Gross Domestic Product (GDP) is available in the National Institute of Statistics from Spain. All series employed here are per capita (18 years or older), expressed in real terms using the consumer price index (CPI base 2016).For a descriptive statistics of the data, see Table B.1 at appendix B.

4.2.2 Empirical Methodology

Data-driven anomaly detection systems have been discussed in the literature as distortions detection systems in many fields of application (see Nagy et al., 2016; Harvey, 1993; Hamilton and Kim, 2002). Such systems aim to detect any abnormal deviations from the normal observations of any data set. Therefore, these methodologies provide a good opportunity to detect anomalies in tobacco sales. Furthermore, given the above characteristics, Spain seems a reasonable candidate country to quantify anomalies.

The aim of this work is the tobacco sales anomaly detection at province geographic level. A prediction of the upper and lower limits of tobacco sales at the provincial level is proposed as a methodology in order to identify any abnormal deviation from this behavior in tobacco sales. In this way, the methodology is proposed through a supervised learning method, adjusting a model to tobacco sales as a dependent variable from price and GDP as independent variables. On the other hand, the detection and estimation of anomalies is done through an unsupervised method, as mentioned before by means of the computation of upper and lower intervals. Several statistical and machine learning models were compared for finding the best model for predicting the tobacco sales of each province (these methods/models are presented in this section).

The main methodology consists on splitting the data into a training and test set of all the available province for the Spanish territory, where the training set consist on all the province available with the exception of the province to predict, which is on the test set. In other words, all province tends to be used to predict a chosen province without including the predicted one. As is common to explain the behavior of tobacco consumption in Spain (see Pinilla, 2002; Fernández et al., 2004; Álvarez et al., 2020), the dependent variable is the per capita tobacco sales for every province and the independent variables are the per capita GDP and price:

$$\text{Tobacco sales} = f\left(\text{price}, GDP, Pop^{18+}\right) \tag{4.1}$$

To model the relationship between the dependent variable and the independent variables (the characteristic vector x), two supervised learning methods have been used. In addition, in order to estimate the upper and lower limits of the prediction interval, quantile predictions will be used as intervals following the methods explained in this section.

The first method used to model the relationship between variables is Quantile Regression (QR). This method was introduced by Koenker and Bassett Jr, 1978 for the estimation of models in which the quantiles of the response are modelled to depend on the features. The τ-th quantile for a population is the sample where the $100/\tau\%$ proportion of the population lies. This model relationships between different quantile predictors and the dependent variable, in this case gives a good interpretability of anomaly detection results as is possible to identify an anomaly within a given range(Evangelou and Adams, 2020).

The conditional α-quantile q of a scalar variable Y, $P(Y \leq q \mid I) = \alpha$ where the probability $0 < \alpha < 1$ is given and I denotes an information set generated by independent variables X. For a complete justification of the method, (Koenker and Bassett Jr, 1978).

For the purposes of this work, two models were combined to build the intervals for detecting anomalies. This is, for the conditional $0.1 - quantile$ as a lower interval and $0.9 - quantile$ as an upper interval, for every province. By construction the probability that a value belongs to the interval between the upper and lower interval is:

$$\mathrm{P}(1 <= \mathrm{X} <= \mathrm{u}) = \mathrm{P}(\mathrm{X} <= \mathrm{u}) - \mathrm{P}(\mathrm{X} <= 1) = 0.9 - 0.1 = 0.8 \tag{4.2}$$

In contrast to the method of least squares that estimates the conditional mean, this method is based primarily on choosing a model for the conditional quantile. Depending on the strength of the assumptions imposed, a range of parametric or non-parametric options are available, (Komunjer, 2013).

For assessing the models, the conditional median response for each province was modelled, which means the $0.5 - quantile$. Not only the models are evaluated for punctual predictions but also the intervals for choosing the best model with a good performance in both tasks, this is discussed later in this section.

A bagging method is proposed in this work as an approach for estimating conditional quantiles. A combination of Random Forest (RF) and QR where proposed by Meinshausen and Ridgeway, 2006 giving as a result Quantile Regression Forest (QRF) approach. One of the main differences between RF and QRF is that QRF for each node of each tree maintains the values of all observations of the node, but RF only maintains the mean of the observations found in the node, (Meinshausen and Ridgeway, 2006). Ranger is a fast implementation of RF or recursive partitioning, (Breiman, 2001), particularly suited for high dimensional data. The R library *ranger* was used to fit a QRF model respectively with the default settings (Wright and Ziegler, 2015).

To detect anomalies, two methods were selected in order to build PIs trough conditional quantiles, for every new observation of the response variable there is a high probability that it lies within the prediction interval (PI), (Pinilla, 2002). Furthermore, an anomaly detection and quantification system is proposed by using an upper interval and lower interval computed through the fitted models.

As mentioned before the PIs are computed through the calculation of the chosen conditional $0.1 - quantile$ and $0.9 - quantile$ for lower interval and upper interval

respectively. One requirement for the decision of α for the intervals is using a symmetric range (i.e. you can not use the $0.1 - quantile$ as the lower interval and the $0.7 - quantile$ as the upper). It is not interest for this research to find the best intervals within a model but to provide a methodology for computing the ratios of abnormalities as shown in section 4.2.2.

The proposed method assumes a uniform distribution with endpoints as the lower and upper limits of the computed PIs. Every point outside this interval range is considered as abnormal, the intervals are also used to quantify the ratio of abnormality for the response variable.

To remark, the observed response for province p is abnormal if either case is true:

$$\begin{cases} y_i > y_t \\ y_i > y_{(100-n)} \end{cases} \tag{4.3}$$

where $\eta(0 \leq \text{t} \leq 100\%)$ presents the chosen quantile level being this symmetrical, limits y (100-t)% level and y t% level represent the upper and lower conditional quantiles, respectively. A small chosen value of t will lead to a larger number of provinces predicted as abnormal.

For training the model a data partition was performed, as explained in the aforementioned sections, the predictive accuracy of the models was measured by splitting the data into training and test sets.

The error assessment was performed either by using a $0.5 - quantile$ prediction for the quantile versions, the interval prediction were used to determine the quantity of the abnormality of tobacco sales that was evaluated with the results of surveys and with some metrics to assess the quality of this intervals.

The performance of the predicted responses ($\hat{y}_i$) in relation to the observed responses (y_i) of the training and test set were assessed by computing the following error metrics:

TABLE 4.1: Error assessment metrics for the fitted models.

Metric	Formula
Prediction error	$e_i = y_i - \hat{y}_i$
Mean squared error	$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2$
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert e_i \rvert$
Mean absolute percentage error	$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left\lvert \frac{e_i}{y_i} \right\rvert$
Median absolute error	$\text{MedAE}(y, \hat{y}) = \text{median}\left(\lvert y_i - \hat{y}_i \rvert, \ldots, \lvert y_n - \hat{y}_n \rvert\right)$
Median squared error	$\text{MedSE}(y, \hat{y}) = \text{median}\left((y_1 - \hat{y}_1)^2, \ldots, (y_n - \hat{y}_n)^2\right)$
Median absolute percentage error	$\text{MedAPE}(y, \hat{y}) = \text{median of sorted } \left\lvert \frac{\hat{y}_i - y_i}{y_i} \right\rvert$

[a] It is computed by ordering the absolute percentage error (APE) from the smallest to the largest and using its middle value (or the average of the middle two values if N is an even number) as the median.

[b] For further understanding of this formulas and its statistical properties, (Hyndman and Koehler, 2006a).

In addition to evaluating the punctual prediction, the prediction interval made has also been evaluated in this work. The academic literature has placed special emphasis on point prediction with respect to interval predictions and predictive densities, consequently there has been little work on the evaluation of PI, (Gneiting and Raftery, 2007).

A review of evaluating point prediction methods used at this work is in the Table 4.2, where the selected metrics assesses the accuracy into the train and test set. However, as the main idea of forecasting is in decreasing the uncertainty, an interval prediction evaluation is performed as well. Table 4.3 summarizes the metrics used to evaluate the prediction interval.

TABLE 4.2: Prediction interval accuracy for the fitted models.

Metric	Formula
Mean Internal Score	$MIS = (p_u - p_l) + \frac{2}{\alpha}(p_l - y)\,1(y < p_l) + \frac{2}{\alpha}(y - p_u)\,1(y > p_l)$
Range	$range = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{p_{u_i}}{p_{l_i}}\right\rvert$
Covareage	$covareage = \frac{1}{n}\sum_{i=1}^{n}\left(1\left(y_i < p_{l_i}\right) \times 1\left(y_i < p_{l_i}\right)\right)$
Pinball	$pinball = (1-\alpha)\sum_{\hat{y}_i<\hat{b}_i}\left\lvert\hat{y}_i - \hat{b}_i\right\rvert + \alpha\sum_{\hat{y}_i<\hat{b}_i}\left\lvert\hat{y}_i - \hat{b}_i\right\rvert$

[a] Where p_l is the lower PI, p_u the upper PI, α is the significance level, y the actual value and $1(\cdot)$ is the indicator function, for more details Breiman, 2001 .
[b] Where n is the number of sample size, p_u for upper PI, p_l for lower PI and y the actual value for each observation i.
[c] Where $\hat{b}_i$ is the predicted value of a interval (either an upper, or a lower).
[d] MIS balance coverage and range of the PI, the best choice is when a model has high coverage, but also short intervals.
[e] Pinball loss function show how well a quantile capture the data, the lower the value of pinball is, the closer the interval is to the specific quantile of the holdout distribution,(Wei et al., 2006).

TABLE 4.3: Quantification of anomalies in per capita tobacco consumption.

Metric	Formula
Upper anomaly ratio (UAR)	$UAR = \frac{y_i - p_u}{p_u}$
Lower anomaly ratio (LAR)	$LAR = \frac{y_i - p_l}{p_l}$

[a] Where p_l is the lower PI and p_u the upper PI.

As a final product of this work a model could discern between a province with abnormal tobacco sales and a province without abnormal tobacco sales, but also, when abnormality is detected this could abnormality due to a quantity is under a lower interval or abnormality due to a quantity is over a upper interval.

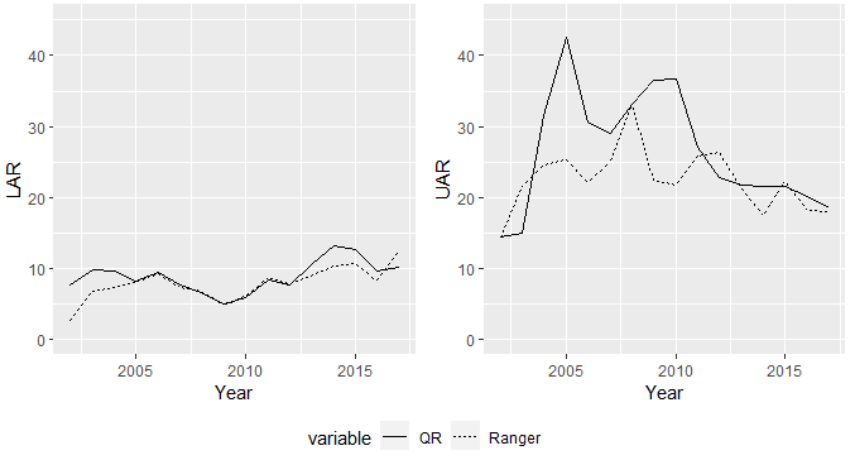
4.3 Results

The results of this article are shown in three parts. First, the evolution over time of the anomalies detected is shown (both the provinces in which sales are lower than fair values and those in which sales are higher than expected). Second, the temporal evolution of the regional anomalies detected in Spain is shown. Finally, the geographical distribution of the anomalies detected is shown.

As indicated, to quantify the anomalies in the Spanish provinces we will use the upper anomaly ratio (UAR) and the lower anomaly ratio (LAR). The average UAR and LAR for the Spanish territory have been represented, averaging the prediction ratio of tobacco sales per capita below the lower limit (the lower prediction interval) and also for the prediction ratio of tobacco sales per capita above the upper bound

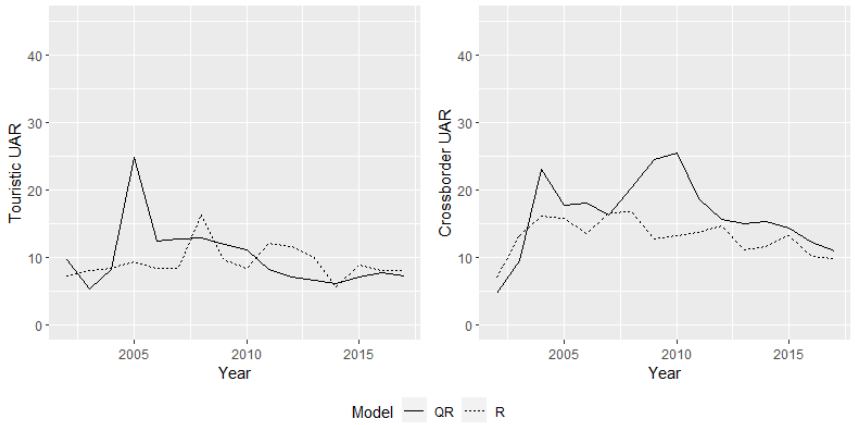
(the upper prediction interval). Figure 4.2 shows the aforementioned index and in it two important questions can be observed: (i) the magnitude of the average upper anomaly exceeds 40%, while the average lower anomaly does not reach 15%, (ii) the temporal evolution shows a upward trend in the lower anomaly and descending in the upper anomaly.

FIGURE 4.2: Average UAR and LAR for the Spanish territory.



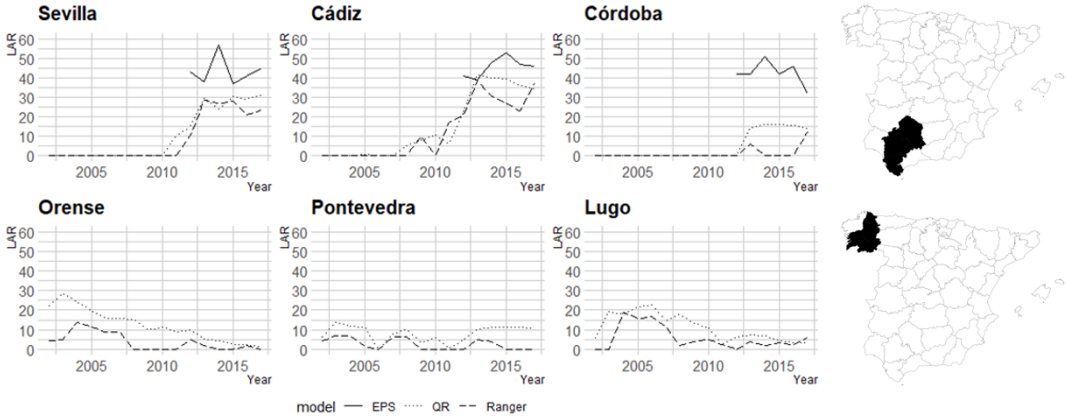
On the side of the magnitude of LAR and UAR, although, as will be seen in this section, the geographic pattern plays a key role, it seems that there is a significant difference between both indices. As for the average LAR, it represents the average percentage of those provinces that present observed sales below the estimated values. Therefore, given that this index is conditioned by crossborder and illegal trade, it seems that both activities have remained constant during the period studied. However, the average LAR has increased notably until 2005, showing a decreasing trend since then. The LAR index represents the average anomaly of the provinces in which the observed sales exceed those estimated by the model. For this reason, both the effect of tourism and that of the crossborder with countries where tobacco is more expensive than in Spain, determine the magnitude of the LAR index. In this line, according to Jakob, Cornuz, and Diethelm, 2017, the border provinces with France have been considered a crossborder effect, while the rest are provinces with a high influence of tourism. According to data from the National Institute of Statistics, the airports in these three provinces are the ones with the most arrivals, after Madrid and Barcelona, the largest provinces in Spain. Figure 4.3 shows the part of the average UAR that represents crossborder and tourism. Both effects seem to have a decreasing trend, the crossborder trend being more accentuated. This last highlight is consistent with the recent evidence on crossborder transactions between Spain and France (Schafferer et al., 2018).

FIGURE 4.3: Touristic and crossborder UAR in Spain.



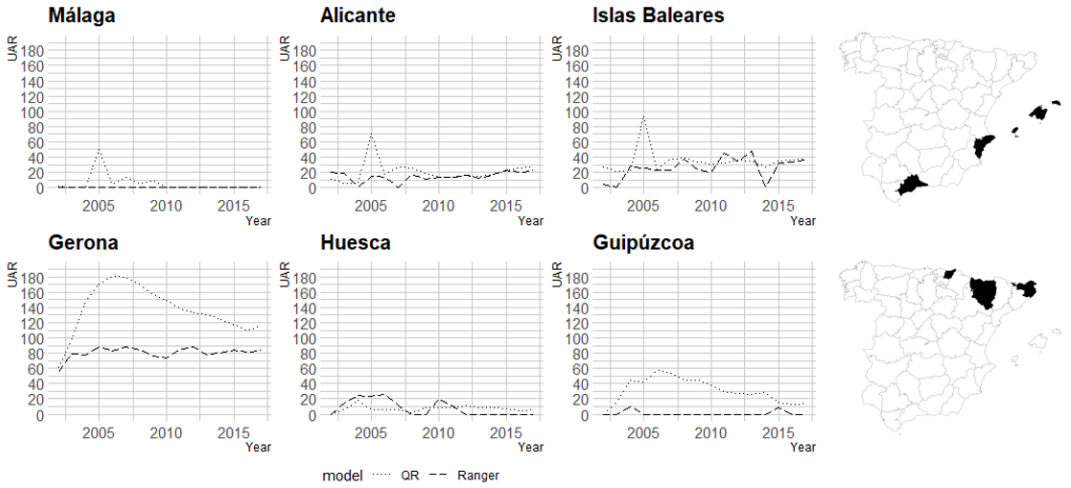
Once the magnitudes have been exposed at a global level, it is interesting to analyze the temporal evolution of UAR and LAR in the Spanish provinces. First, regarding the UAR, there are 6 provinces that stand out for their behavior. As shown in Figure 4, there are three provinces in the south of Spain (Sevilla, Cádiz and Córdoba) in which the UAR is observed for the first time in 2010 and grows notably until 2017, reaching values close to 40% in some cases. On the other hand, in three other provinces (Orense, Pontevedra and Lugo) the UAR has a decreasing trend, in addition, it rarely takes values close to 20%. In addition, Figure 4 also shows the values shown by the EPS performed by the TTCs from 2012 to 2017. As the aforementioned figure shows, in line with the previous literature, it seems that the use of independent data provides estimates of the illicit market lower than EPS. Specifically, while in Cádiz the EPS is slightly oversized, in Córdoba there are substantial differences.

FIGURE 4.4: Temporal evolution of UAR in the Spanish provinces.



Six provinces also stand out in the temporal evolution of the UAR in the Spanish territory. On the one hand, Malaga, Alicante and the Islas Baleares, provinces with a high influence of tourism , present a similar trend. On the other hand, in the border provinces with France, it is observed that in Gerona, Huesca and Guipúzcoa the UAR shows a decreasing trend, with the anomalies detected in Gerona of a much higher magnitude. As indicated in the introduction, Gerona is the Spanish region in which sales fell the most due to border closures due to the public health crisis of COVID-19. Therefore, the results are consistent with what is indicated.

FIGURE 4.5: Temporal evolution of LAR in the Spanish provinces.



Although the provincial evolution provides important information, the geographical distribution of the anomalies helps to understand the "contagion effect" of the UAR and the LAR, as well as the behavior of consumption at the borders with other countries. In this sense, Figures 4.6 and 4.7 show the geographic distribution of LAR and UAR according to the QR model, respectively. As can be seen, the anomalies in the provinces in which lower-than-estimated sales have been observed reach up to 35% and are concentrated in the Northwest in 2002 and in the South in 2017, something consistent with the previous literature. In addition, regarding the UAR, the anomalies in the provinces with sales above those estimated reach values of 190%. Finally, while in 2002 they were concentrated in tourist provinces and bordering Portugal, in 2017 the tourist provinces remain the same, but it is the border with France where the anomalies are located.

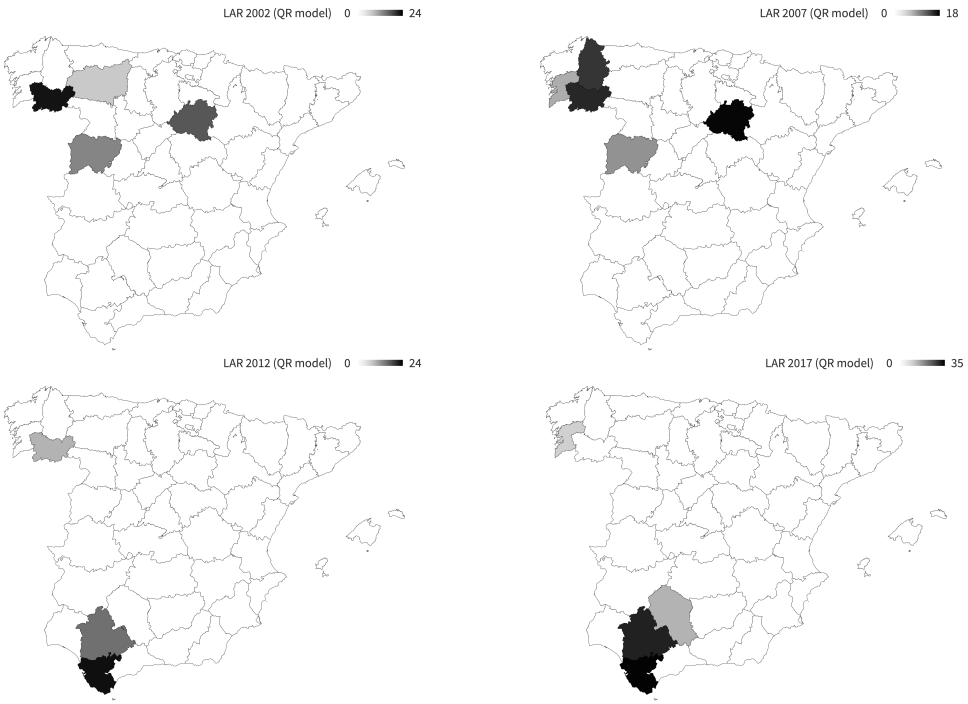


FIGURE 4.6: Geographical distribution of LAR in the Spanish provinces (QR model).

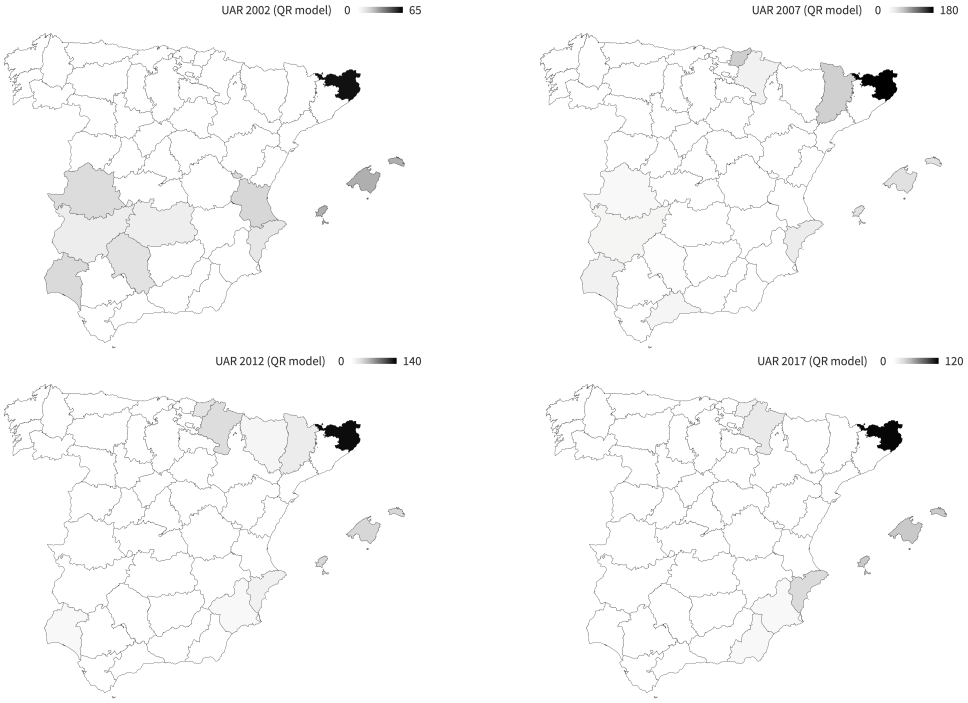


FIGURE 4.7: Geographical distribution of UAR in the Spanish provinces (QR model).

One of the main values of this article is the quantification of the LAR and UAR anomalies. Along these lines, the closure of Spain's borders with border countries due to the health crisis caused by COVID-19 during the months of April and May 2020 has made it possible to analyze the robustness of the results of this paper. As shown in Figures 4.8 and 4.9 , both the geographic pattern and the magnitude of the

UAR estimated in this work show robustness with the falls in tobacco sales in Spain in the months of April and May 2020. On the one hand, tobacco sales fell in April and May up to 180 and 160 percent, respectively, against. This magnitude is consistent with the anomalies shown in Figure 4.7. Furthermore, the geographical pattern is also coincident, with the greatest drops in tobacco sales having been observed in the border areas with France and in the tourist provinces. In addition, the border areas with Gibraltar are those in which tobacco sales have decreased the least during the border closures. It is precisely in these provinces that the highest UARs are observed in the last years analyzed.

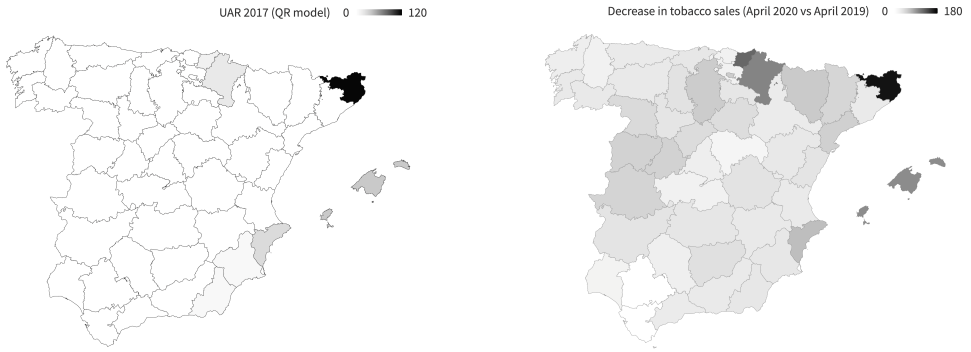


FIGURE 4.8: Comparison between the results of the model and the fall in sales of April 2020.

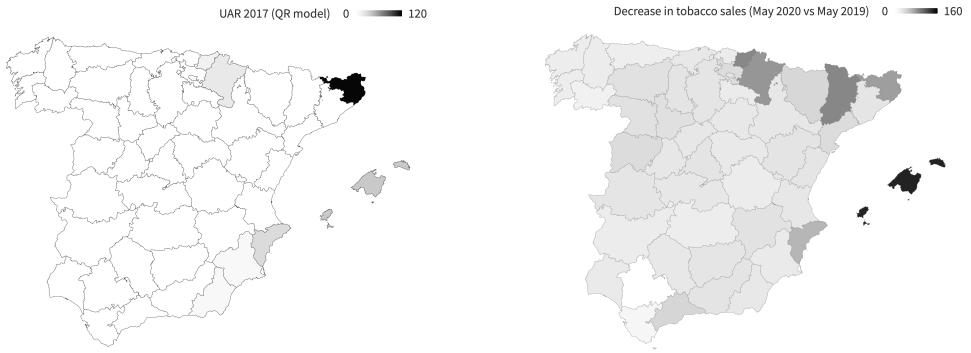


FIGURE 4.9: Comparison between the results of the model and the fall in sales of May 2020.

4.4 Conclusions

In recent years there has been a growing interest in knowing the mechanisms that can control cigarette consumption due to the great impact of tobacco consumption on public health. Along these lines, due to the free movement of people and illegal activities, sometimes legal tobacco sales are not a faithful representative of tobacco consumption. For this, there are multiple studies commissioned by the TTCs to demonstrate that there is an illicit and crossborder activity that generates a greater consumption of tobacco in the population than the governments believe. Although there are many initiatives commissioned by TTCs, EPSs are the most widespread studies. These EPSs are in charge of detecting what appears to be illegal trade in provinces where there is less than reasonable tobacco consumption. This study has shown, in line with previous literature, that in Spain the EPSs that are performed

to estimate the illicit market (mainly in border areas with Gibraltar) are oversized. In addition as a contribution to the literature, in this work anomalies have been detected in provinces where sales are higher than fair values, information that TTCs ignore.

To the best of our knowledge, this study is the first to quantify anomalies in regional tobacco sales in Spain, including anomalies in provinces where more than fair values are sold. In particular, the results found on the provinces in which the observed tobacco sales are below fair values are similar to those found in the previous literature: the EPSs overestimate the illicit trade values. In reference to the anomalies detected in the provinces in which tobacco sales above reasonable values are observed, something that is rarely found in the literature, the findings are novel. Specifically, the provinces where tobacco sales are highest relative to fair values are those where sales have fallen the most with border closures driven by the COVID-19 public health crisis. This finding undoubtedly confirms that tobacco sales in Spain are conditioned by the effect of tourism and the price differential with France. Furthermore, it is found that cross-border tobacco purchases between Spain and France show a decreasing trend in recent years, somewhat in line with the previous literature. Although the effect of cross-border tobacco purchases between Spain and France has decreased in recent years, the deviation of sales in the provinces where more tobacco is sold than is reasonable (tourist provinces and the border with France) is much higher on average to deviations from those provinces where sales are below fair values (border with Gibraltar). This result is novel, given that the anomalies of tourist provinces and border provinces with France had never been quantified. Therefore, when the TTCs show results of EPSs in Spain, it must be taken into account that there is an inverse effect to what the EPS detects that must be considered by governments.

Our results show provinces in which smoking control policies cannot be evaluated using official sales, since these sales are altered. In this sense, we find that the provinces in which sales are most affected are border and tourist areas, evidencing the existence of large-scale illegal trade and cross-border purchases. The results support that for some years no anomalies have been found in border areas with Portugal. Therefore, the results reveal the effectiveness of the common policies implemented by the governments of Spain and Portugal, which consists of maintaining a low-price differential between the countries.

All these results provide recipes for the agendas of academics and governments. The academic community should bear in mind that there is more evidence about the overestimation of illicit trade in EPSs and that the average of the excess anomalies is much higher than the average of the default anomalies. In addition, policy makers should consider that there are provinces where the evaluation of the effectiveness of anti-smoking policies cannot be evaluated using official sales. The allocation of resources to control smoking must consider the abnormalities identified in this report. If not, the provinces in which there is excessive consumption distorted will have more recourse to control a tobacco habit that is not real. On the contrary, the border provinces with Gibraltar will have fewer resources to control smoking if official sales are used, when the reality is that there is hidden consumption due to illicit trade.

In conclusion, the results seem to show that EPSs overestimate the value of illicit trade. Furthermore, in Spain, the provinces with sales volumes above fair values have higher ratios than those in which sales are below fair values. Therefore, it seems that the sum of the effect of tourism and cross-border purchases between Spain and France is higher than the cross-border purchases between Spain and Gibraltar

detected by the EPS. Finally, the anomalies prevent the Spanish government from knowing the total benefit for public health generated by the policies against smoking.

Chapter 5

Chapter 5. Short-term prediction of Time Series based on bounding techniques.

5.1 Introduction

The goal of this chapter is to introduce a new non-parametric model for time series predictions based on its observed past values. It is a well known fact that in parametric time series analysis that the relationship between the observed past values of the time series and the prediction is obtained by specifying a functional cost function and a fixed finite number of parameters. Widely studied parametric models include auto-regressive (AR), moving average (MA), and different combinations as ARMA or ARIMA models Box and Jenkins, 1976; Hamilton, 1994.

In nonlinear time series analysis, some commonly studied parametric approaches include threshold auto-regressive (TAR) Tong, 1983, the exponential auto-regressive (EXPAR) and smooth-transition auto-regressive (STAR) Haggan and Ozaki, 1981; Chang and Tong, 1986 models being some important examples. In these approaches the performance of the parametric predictor is a consequence of a chosen cost function.

By contrast, in non-parametric approaches a more flexible methodology that avoids choosing a specific cost function.

The local conditional mean or median method provides a prediction using the mean or the median of a neighborhood of the interest point Truong, 1993. The Nadaraya-Watson estimator averages past observations by a kernel function obtaining a prediction Nadaraya, 1964; Watson, 1964. Local regression by means of either linear or polynomial functions of past observations can be used to approximate nonlinear relationships between variables Härdle, 1990; Fan and Gijbels, 1996.

Semi-parametric models as nonlinear additive auto-regressive (NAAR) models or functional coefficient auto-regressive (FAR) models have also been proposed Hastie and Tibshirani, 1990; Härdle, Lütkepohl, and Chen, 1997. An extensive review of non-parametric methods applied to time series prediction has been compiled by many researchers Fan and Yao, 2003; Gao, 2007; Gooijer and Gannoun, 2000; Yin and Shang, 2016.

In this chapter a new non-parametric prediction method is proposed. The prediction is obtained by a weighted sum of past observations. An upper bound of the prediction error is computed under some deterministic and stochastic assumptions. A constrained optimization problem is formulated both to minimize the upper bound of the prediction error and to obtain the set of optimal weights used to compute the

prediction. The main novelty of our method is that the optimization problem includes a hyperparameter to balance the deterministic-stochastic assumptions. This hyperparameter can be tuned with training data and a cross-validation scheme in order to improve the predictor performance Bergmeir, Hyndman, and Koo, 2018. The proposed predictor provides a general framework that includes as particular cases some relevant non-parametric predictors as the Nadaraya-Watson predictor Nadaraya, 1964; Watson, 1964 and predictors based on local linear regression Fan and Gijbels, 1996, models that have been widely used in the literature Mangalova and Shesterneva, 2016; Hyndman et al., 2005.

The chapter is organized as follows. In Section 5.2, the problem formulation is addressed. The deterministic and stochastic assumptions are introduced in Section 5.3. The new predictor is proposed in Section 5.4. Benchmark results are illustrated in Section 5.5. Finally, Section 5.6 reports some conclusions.

We consider a discrete time series process $\{Z_t\}$ with time $t \in \{0, 1, 2, \dots k\}$.

5.2 Formulation

We assume that at instant $t = k$ past data $\{Z_t\}$ with $t \leq k$ has been observed and we are interested a forecast for predicting Z_{k+1}. Once a detrend is applied to the time series,[1]

we can write $Z_t = y_t + \mu_t$ with $t \in \{0, \dots, k\}$, being $\{\mu_t\}$ the trend component, $\{y_t\}$ the detrended component and y_{k+1} the desired detrended future value.

We also denoted by $\{z_j\}$ with $j = 0, 1, ..., k$ the set of the vectors consisting of the detrended observed past values of the time series, that is $z_j^T = [y_{j-p-1}, \dots, y_{j-1}, y_j]$.

Henceforth this p-dimensional vector set will be called *embedding vector*. This set of data, z_j is used to forecast future values for the time series. At this point let us be specific about the meaning of the parametric and non-parametric models used in this article. A parametric model is characterized by the use of the training set for estimating the parameters of the model and once this is done the training data set is not used again. On the contrary, the non-parametric model considered in this work is of local nature, since each forecast is obtained by using all the available data set but selecting a time neighborhood of the data point of interest. In this sense, it is assumed that the time series can be generated by an unknown local linear modelHärdle, 1990; Fan and Gijbels, 1996.

Assumption 1 *Considering the forecast of y modeled as:*

$$y_{k+1} = r(z_k)^T \Phi_k + e_k \tag{5.1}$$

where it is assumed that the existence of an unknown vector of parameters $\Phi_k \in \mathcal{R}^n$, a known function $r(\cdot)$ valuated at the embedding set and an unknown error term e_k.[2]

In order to complete the presentation of the model it should be discussed in more detail the so called *regressor generator function* $r(\cdot)$. This function allows transform the original values into vectors of dimension n_r by means of the vectors belonging to the embedding set. A formal definition of this regressor generator function is as follows.

[1] It should be noted that in coherence with the prediction system and in order to estimate μ_{k+1}, only the past observations can be used, independently of the detrending method used.

[2] This modeling is flexible enough to admit alternative assumptions about the error term. As discussed later, the model is presented by using both deterministic and stochastic bounds for the error term e_k.

Definition 1 (Regressor generator function) *The function $r(\cdot) : \mathcal{R}^p \rightarrow \mathcal{R}^{n_r}$ specifies the regressor vector components. This function admits any kind of auto-regressive representation, nonlinear expression of past components and different functional forms for decomposing the different components of the time series.*[3]

Definition 2 (Linear Prediction) *For an instant k, a forecast of $y_{k+1} \in \mathbb{R}$ can be derived through a linear combination of past data, that is:*

$$\begin{aligned} \hat{y}_{k+1}(\Psi) &= b_Y^T \Psi \\ &= \sum_{j=1}^{v} \Psi_j y_j \end{aligned} \tag{5.2}$$

where $1 \leq v \leq k$, $\Psi \in \mathbb{R}^v$ is a weight vector and $b_Y = [y_1, \ldots, y_v]^T$.

When $v = k$, all data is used to forecast y_{k+1}. Then, the forecast error can be explained as the difference between y_{k+1} and the linear prediction $\hat{y}_{k+1}(\Psi)$.

Definition 3 (Prediction error) *It is defined the prediction error $\hat{e}_k(\Psi)$, being k the time instant:*

$$\hat{e}_k(\Psi) = y_{k+1} - \hat{y}_{k+1}(\Psi). \tag{5.3}$$

Thus, the crux of the matter is how to get not only the weight vector Ψ but also an outer limit of the prediction error. This outer limit is estimated by using the assumed relationship between z_{j-1} and y_j, with $j = 1, 2, \ldots, k$ in expression (5.1). Then, a set of past components z_j with $j = 0, 1, ..., k$ should be available. Section 5.3 formulates these key ideas.

5.3 Assumptions

In this section the assumptions are based on some local affine approximations. In order to construct the proposed predictor, the definition of approximation error is used. This is, the result of using the vectors $r(z_{j-1})$ and Φ_k to infer y_j.
[4]

Definition 4 (Approximation error) *For a vector Φ_k, the approximation error e_{j-1} with the pair (z_{j-1}, y_j) being $j = 1, 2, ..., k$ can be defined as:*

$$e_{j-1} = e_{j-1}(\Phi_k) = y_j - r(z_{j-1})^T \Phi_k. \tag{5.4}$$

From now on the dependency of $e_{j-1}(\Phi_k)$ with Φ_k is omitted. It should be noted that the value of Φ_k is unknown. The prediction error $\hat{e}_k(\Psi)$ may be biased by the selected vector Ψ. The theorem 1 suggests an approach to define the prediction error $\hat{e}_k(\Psi)$ as a function of the vector Ψ and the aforementioned approximation errors e_j .

Theorem 1 proposes an expression to characterize the prediction error $\hat{e}_k(\Psi)$ as a function of vector Ψ and approximation errors e_j previously defined.

[3]For instance suppose a set $z_k = [y_k, y_{k-1}, y_{k-2}]$. Then $r(z_k)$ could be the function $r(z_k) = z_k$ that is, an auto-regressive model. There exist also alternative configurations such as a nonlinear auto-regressive model $r(z_k) = [y_k^2, y_{k-1}, y_{k-2}, y_k \cdot y_{k-2}]$ or any possible combination.

[4]The reader should note that the point is to relate the *k-th* prediction error e_k and the prediction errors generated by using the *k-th* vector of unknown parameters Φ_k with the *i-th* regressors $r(z_i)$, with $i = 0, \ldots, k-1$.

Theorem 1 *For either vector* $\Psi \in \mathbb{R}^v$ *so that*

$$\sum_{j=1}^{v} \Psi_j r(z_{j-1}) = r(z_k), \tag{5.5}$$

then, the prediction error

$$\hat{e}_k(\Psi) = y_{k+1} - \hat{y}_{k+1}(\Psi)$$

is set as a linear combination of the approximation errors e_j, this is

$$\hat{e}_k(\Psi) = -\sum_{j=1}^{v} \Psi_j e_{j-1} + e_k.$$

Notice that it is necessary to know the vector Φ_k to get an error value e_{j-1}. Alternatively, other properties of e_{j-1} can also be assumed. Both deterministic and stochastic options are available in the literature. In a deterministic view, an upper bound of $|e_{j-1}|$ is considered. This idea is discussed in the section 5.3.1.

5.3.1 Deterministic error

In methods with bounded-error (Milanese et al., 1996), a parametric model and an unknown but bounded-error are regarded. An upper limit of this error is expected to estimate a set of consistent parameters. Similar assumptions are presumed in this work in order to develop a predictor with deterministic assumptions.

Assumption 2 *Constants* $\sigma, L \geq 0$ *are set such that approximation errors* e_{j-1} *and* e_k *are delimited by expressions*

$$|e_{j-1}| \leq \sigma + L||z_{j-1} - z_k|| \tag{5.6}$$

with $j = 1, ..., k$ *and*

$$|e_k| \leq \sigma \tag{5.7}$$

being $||\cdot||$ *a norm.*

The error term is bounded by $|e_k| \leq \sigma$. The assumption 2 has been broadly used in the bounded-error system identification's context (Milanese et al., 1996). Remark that σ is the tunning parameter that adds the minimum level of noise considered and L the tunning parameter of uncertainty due to the local affine approximation.

Remark 1 *Historical data can be used to estimate an approximate value of* σ *and* L *when no prior knowledge of these constants is available. In (Bravo et al., 2017) a method based on bounded-error and non-counterfeit data is provided.*

Lemma 1 *Considering Assumptions 1 and 2, for any* Ψ *such that* $A^T\Psi = r(z_k)$*, then, the prediction error* $\hat{e}_k(\Psi) = y_{k+1} - \hat{y}_{k+1}(\Psi)$ *is bounded by:*

$$|\hat{e}_k(\Psi)| \leq \sum_{j=1}^{v} |\Psi_j|(\sigma + L||z_{j-1} - z_k||) + \sigma. \tag{5.8}$$

Proof.Through a straightforward application of Theorem 1 and bound $|e_i| \leq \sigma + L||z_j - z_k||$ *is obtained the expression (5.8).* QED

At this point the possibility of considering how to obtain the vector Ψ is established. A wise option is to use the vector that minimizes an upper bound of $|\hat{e}_k(\Psi)|$ using the expression (5.8).

Definition 5 (Deterministic predictor) *The deterministic prediction $\hat{y}_{k+1}(\Psi^D)$ is defined by*

$$\hat{y}_{k+1}(\Psi^D) = \sum_{j=1}^{v} \Psi_j^D y_j,$$

where vector Ψ^D adresses the problem of constrained linear optimization as follows

$$\begin{aligned} \Psi^D = \quad & arg\min_{\Psi} \quad ||W_k\Psi||_1 \\ & s.t. \quad A^T\Psi = r(z_k) \end{aligned} \tag{5.9}$$

where W_k is a diagonal matrix with central items $w_{j,j}^k = \sigma + L||z_{j-1} - z_k||$ with $j = 1, ..., v$. Then, an upper bound of the absolute value of the prediction error is minimized by the vector Ψ^D.

It is important to note that the notation Ψ^D refers to the deterministic nature of the estimate. Expression (5.9) use L_1-norm to obtain the vector solution Ψ^D. In this case, Ψ^D is sparse, that is, most of number of components Ψ_i^D of vector Ψ^D are zero. As Ψ^D is a sparse matrix and considering Definition 5 then it is deduced that $\hat{y}_{k+1}(\Psi^D)$ use a relatively short number of measurements y_i.

5.3.2 Stochastic error

The stochastic view consider the approximation error e_j as a random variable. So there are some assumptions about the mean and the variance of e_j. Specifically there are assumptions in the dimension of variance of e_j.

Assumption 3 *The independent variables, approximation error e_{j-1} and error term e_k have zero mean and variances bounded by $var(e_{j-1}) \leq \sigma + L||z_{j-1} - z_k||$ and $var(e_k) \leq \sigma$ accordingly. Positive values of constants σ and L is taken as prior knowledge.*

As indicated in Remark 1, if not available previous knowledge of the constants σ and L, historical data may be used to obtain an estimation. The variance of error e_{j-1} consists of a minimum value defined by σ and a term depending of the local approximation, i.e. $||z_{j-1} - z_k||$. it is possible to extend that as e_{j-1} and e_k are random variables then $\hat{e}_k(\Psi)$ is also random and therefore other properties can be derived.

Assumption 4 *Taking into account the previous Assumptions 1 and 3, for any Ψ such that $A^T\Psi = r(z_k)$, prediction error $\hat{e}_k(\Psi) = y_{k+1} - \hat{y}_{k+1}(\Psi)$ is a random variable with zero mean and variance, it is defined by:*

$$\begin{aligned} var(\hat{e}_k(\Psi)) \quad &= \quad \sum_{j=1}^{v} \Psi_j^2 var(e_{j-1}) + \sigma \\ &\leq \quad \sum_{j=1}^{k} \Psi_j^2(\sigma + L||z_{j-1} - z_k||) + \sigma. \end{aligned} \tag{5.10}$$

At this point,it is possible to formulate a predictor that minimize the outer bound of the variance prediction error.

Definition 6 (Stochastic prediction) *The stochastic prediction $\hat{y}_{k+1}(\Psi^S)$ is defined by:*

$$\hat{y}_{k+1}(\Psi^S) = \sum_{j=1}^{v} \Psi_j^S y_j,$$

being Ψ^S a vector that solves a constrained linear optimization problem as follows:

$$\begin{aligned} \Psi^S = \quad & \underset{\Psi}{arg\,min} \quad \Psi^T W_k \Psi \\ & s.t. \quad A^T\Psi = r(z_k). \end{aligned} \tag{5.11}$$

An explicit notation of this optimization problem is:

$$\Psi^S = W_k^{-1}A(A^T W_k^{-1} A)^{-1} r(z_k). \tag{5.12}$$

In the same way, Ψ^S highlights the stochastic assumptions considered to get the estimate. The following equality is satisfied

$$\hat{y}_{k+1}(\Psi^S) = b_Y^T \Psi^S = r(z_k)^T \Phi^*,$$

where $\Phi^* = (A^T W_k^{-1} A)^{-1} A^T W_k^{-1} b_Y$ is the argument of which minimizes the quadratic prediction-error, with the following cost function:

$$\begin{aligned} J(\Phi) = \quad & (b_Y - A\Phi)^T W_k^{-1} (b_Y - A\Phi) \\ = \quad & \sum_{j=1}^{k} \frac{(y_j - r(z_{j-1})^T \Phi)^2}{(\sigma + L||z_{j-1} - z_k||)}. \end{aligned} \tag{5.13}$$

In this way, the stochastic prediction is equivalent to solve a weighted least-squares problem where the weights are set by the items of the diagonal of W_k squared. Commonly, Ψ^S is not a sparse vector, since most items are non zero numbers. So, in order to get the prediction y_{k+1} a great number of y_j would be used.

The goal of this paper is to bring a predictor that combines the two predictions based on the different assumptions obtained from Ψ^D and Ψ^S respectively. Section 5.4 introduces the key points of this paper.

5.4 Proposed predictor

In this work, we aim to obtain an estimation of the output y_{k+1} by a linear combination of past data y_j, with $j = 1, 2, ..., v$ where $v \leq k$ (Roll, Nazin, and Ljung, 2005). Next, a formal definition of the proposed predictor is provided by a constant $\gamma \geq 0$ to balance the deterministic or stochastic nature of the prediction.

Definition 7 *Given a constant $\gamma \geq 0$, the predictor $\hat{y}_{k+1}(\Psi^*)$ is defined by $\hat{y}_{k+1}(\Psi^*) = \sum_{j=1}^{k} \Psi_j^* y_j$ where Ψ^* is the optimal solution of:*

$$\begin{aligned} \Psi^*(\gamma) = \quad & \underset{\Psi}{arg\,min} \quad ||W_k\Psi||_1 \\ & s.t. \quad A^T\Psi = r(z_k) \\ & \quad\quad ||\Psi - \Psi^S||_1 \leq \gamma \end{aligned} \tag{5.14}$$

and vector Ψ^S is defined in (5.12).

Some qualitative properties of the proposed predictor can be clarified. Note that expression (5.14) is a constrained linear convex optimization problem and can be solved in an efficient way Boyd and Vandenberghe, 2004. Assuming that (5.14) has a bounded solution, there is a constant $\bar{\gamma}$ such that if $\gamma \geq \bar{\gamma}$ then equality $\Psi^* = \Psi^D$ is obtained. Term $||\Psi - \Psi^S||_1$ of expression (5.14) takes into account the stochastic

Assumption explained in Section 3 to obtain the optimal solutions Ψ^*. If $\gamma = 0$ then $\Psi^* = \Psi^S$. So, constant γ can be seen as a tuning parameter to balance the deterministic or stochastic nature of the approximation error.

Remark 2 *It is important to remark that the proposed predictor encompasses some relevant non-parametric predictors. If $\gamma = 0$ and $r(z_k) = 1$ the proposed predictor is equivalent to the Nadaraya-Watson predictor Nadaraya, 1964; Watson, 1964. On the other hand if $\gamma = 0$ and $r(z_k) = [z_k^T \ 1]$ a predictor based on Local Linear Regression is obtained. Besides, if $\gamma = 0$ and if $L = 0$ a parametric auto-regressive linear regression is performed.*

Remark 3 *It is important to remark that following similar reasoning it is possible to obtain different forecasting horizons. This is represented by the expression $y_{k+h}(\Psi^*)$ being $h \geq 1$ the number of steps ahead.*

5.5 Results

Results are shown in this section, several time series are selected which comes from different areas and have different statistical properties, so they are suitable benchmark to test time series predictors. Four well known time series are studied, so the Monthly airline passenger numbers, the Canadian lynx data, the Monthly critical radio frequencies in Washington, D.C. and the Monthly pneumonia and influenza deaths time series. On the other hand, a real-world time-series is selected in order to show an applied case of the proposed predictor. We use this benchmark in order to demonstrate the appropriateness and effectiveness of the proposed predictor.

Subsection 5.5.1 explain the characteristics of the study performed, so hyper-parameterization, kernels, error measures and more details are exposed below.

5.5.1 Considerations

- To simplify the study, the proposed predictor (denoted CP) is considered with values $\sigma = 0$ and $L = 1$ in all cases. Note that σ and L could be considered hyper-parameters in order to improve the results obtained by the proposed predictor in this study.

- The proposed predictor (CP) is compared to three Nadaraya-Watson predictors (denoted by the acronyms as $NW1$, $NW2$ and $NW3$) and three local linear regression models (denoted with by acronyms LL_1, LL_2 and LL_3). For the aformentioned methods, the acronym subscript refers to the use of Epanechnikov, Gaussian and Tricube kernel functions respectively to define the local weights. Table 5.1 shows the expression of weights $w_{i,i}$ with $i = 1, ..., N$ for the aforementioned kernel functions. A bandwidth γ is considered in the non-parametric predictors. Also, an auto-regressive linear regression (AR) is indirectly included in the benchmark, as the proposed predictor also includes this model according to the hyperparameter combinations as explained in remark 2 of section 7.

- Two different forecast consistency measures are used in order to compare the predictor performances with the aforementioned models: Mean Absolute Error (MAPE) and Symmetric Mean Absolute Error (SMAPE) that have been studied by several authors (Armstrong, 1985). Mean Absolute Error is defined by

TABLE 5.1: Kernel functions

Epanechnikov	Gaussian	Tricube
$w_{i,i} = \begin{cases} 1 - v_i^2 & if\ \|v_i\| \leq 1 \\ 0 & if\ \|v_i\| > 1 \end{cases}$	$w_{i,i} = e^{-\frac{1}{2}v_i^2}$	$w_{i,i} = \begin{cases} (1 - \|v_i\|^3)^3 & if\ \|v_i\| \leq 1 \\ 0 & if\ \|v_i\| > 1 \end{cases}$
$v_i = \frac{\|z_i - z_k\|}{\gamma}$		

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \tag{5.15}$$

where $\hat{y}_t$ and y_t are the predicted and observed data, respectively, and n is the number of data. The second criterion is the Symmetric Mean Absolute Percentage error (SMAPE), which is

$$SMAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|\hat{y}_t - y_t|}{(|y_t| + |\hat{y}_t|)/2}, \tag{5.16}$$

where $\hat{y}_t$ and y_t are the predicted and observed data, respectively, and n is the number of data.

- In order to train the predictors, the time series has been splitted into a training and test set with a percentage between 70-90 for a training set and a percentage between 10-30 of data for a test set. The aforementioned error measures described in formulas (5.15) and (5.16) are selected to benchmark the predictors, not only because of its interpretability but also because are scale-independent (Hyndman and Koehler, 2006b). An optimal value for hyper-parameter-γ is performed by leave-one-out cross validation at the training set and it is validated in the test set to evaluate and avoid overfitting of the model.

- As explained in Section 5.4, the estimation can be used by a linear combination of the past data. Different models are evaluated by using only the training set to infer the prediction; this is, when past data y_j is used with $v < k$. Besides, three different forecast horizons are computed by predictor (1 step-ahead, 2 step-ahead and 3 step-ahead).

In order to test the proposed predictor, two different type of data are used to perform the benchmark, the first one in subsection 5.5.2 where some famous and more academic time series are used to compare the models, and the second in subsection 5.5.3 where a real life time series is used to test the predictor.

5.5.2 Academic Time series

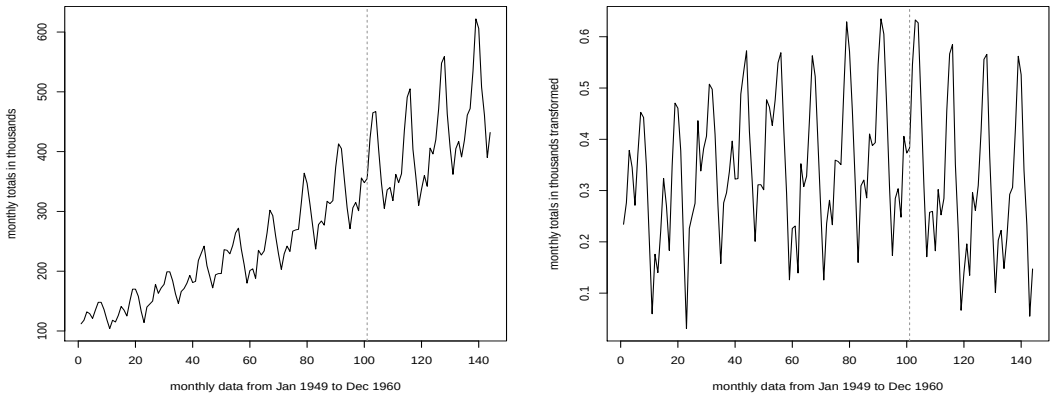
This subsection compare 6 non-parametric models against the proposed model in four different time series, performing forecasts in three different predicting horizon lengths. Each time series provides the results in bar-plots by $MAPE$ and $SMAPE$ errors in the test set for the proposed forecasting horizons. A final sub-subsection averages all time series results in order to extract general conclusions for the different benchmarked models.

5.5.2.1 Airline passengers dataset

The classic Box and Jenkins airline data contains monthly totals of international airline passengers from 1949 to 1960 (Box and Jenkins, 1976).

This time series plotted in Figure 5.1a has 144 observations, the first 101 observations were used as training set and the last 43 as test set. It has also been very analyzed in the time series literature.

As explained in Section 5.2, a detrended time series is considered. In this sense, a log with base 10 and a linear detrend function are applied to transform the data, the data set transformed is shown in Figure 5.1b.



(A) Original Time series (B) Transformed Time series

FIGURE 5.1: Monthly totals of international airline passengers (1949 – 1960).

The auto-correlation plot of the transformed data set shown in figure 5.2 at the appendix exhibits a high correlation between observations of this time series that are separated by $k = 12$ time units. Consequently, the predictor can be represented as $r(z_k) = z_k = [y_{k-1}\ y_{k-2}\ ...\ y_{k-12}]^T$.

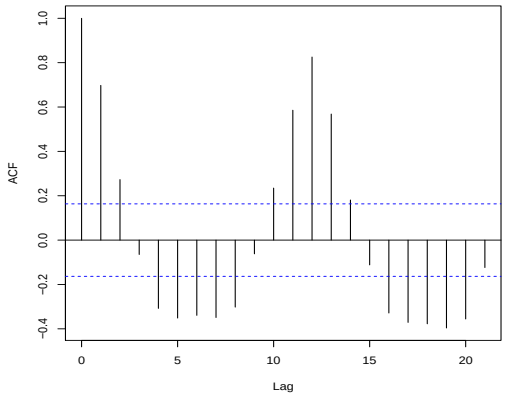


FIGURE 5.2: Auto-correlation function of Airline passengers transformed time series.

Figure 5.3 shows the forecasts of the proposed predictor by forecasting horizons with the hyper-parameters selected in both error measures.

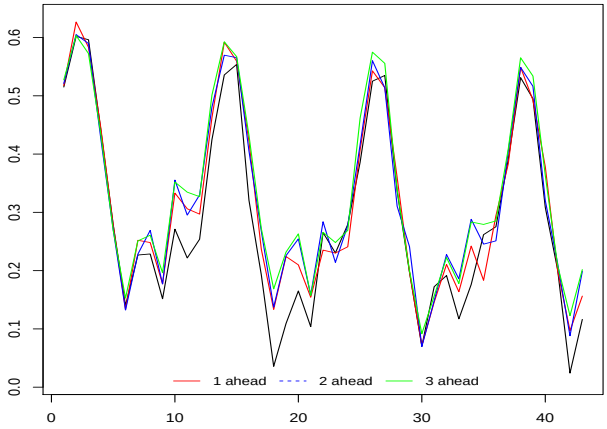


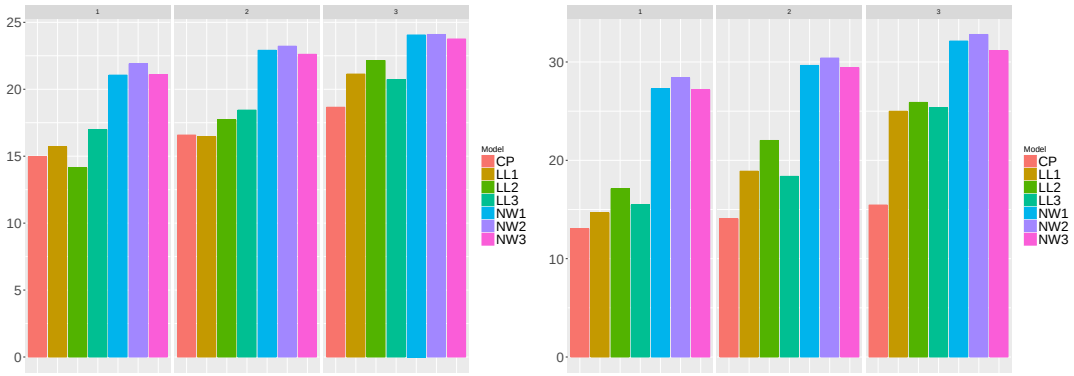
FIGURE 5.3: International airline passengers predictions by forecasting horizon in the test set.

The hyper-parameter γ is selected in the training set where the error is minimum, the value of γ is inferred to perform forecasts in the test set. Besides, depending on the error measure selected in the training set the results may vary. Specifically, same optimal hyper-parameters are found in both error measures, these results are in table 5.2.

TABLE 5.2: Airline passengers time series optimal gamma.

Ahead	γ_mape	γ_smape
1.00	0.12	0.12
2.00	0.14	0.14
3.00	0.00	0.00

Results of this time series are shown on Table D.1 in the appendix. To sum up the aforementioned table in a graphical way, figure 5.4 show the error measures by predictor and prediction horizon.



(A) SMAPE (B) MAPE

FIGURE 5.4: Mean of errors by forecasting horizon in airline passengers time series in test set.

At figure 5.4, the aggregated mean error is plotted by forecasting horizon and model, results show how the proposed predictor outperforms in the error metrics. However, the $SMAPE$ error metric shows similar results to LL results. Besides, the selected kernels makes no significant variation of results.

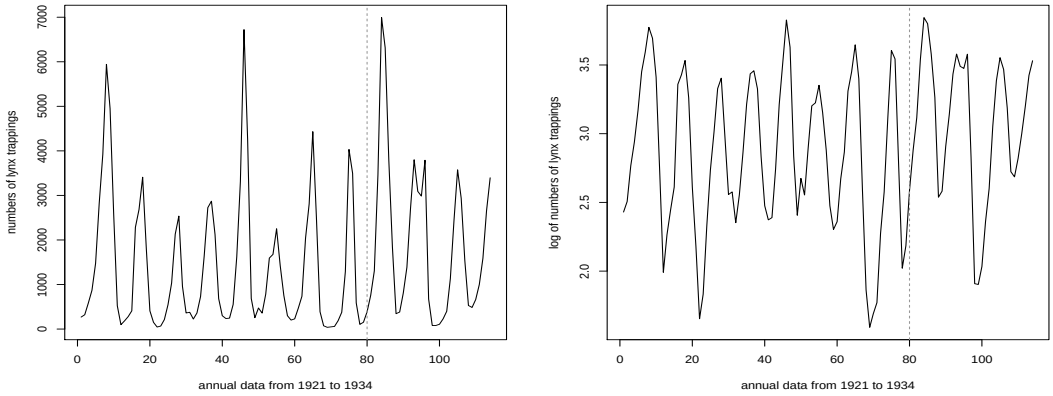
5.5.2.2 Canadian Lynx

The Annual numbers of lynx trappings in Canada, contains the number of lynx trapped per year in the Mackenzie River district of Northern Canada from 1821 to 1934 (Campbell, 1977).

It has been extensively analyzed in the time series literature with a focus on the nonlinear modeling. The lynx series plotted in Figure 5.5a shows a periodicity of approximately 10 years. The lynx series was studied by many researchers and found the best-fitted model is AR(12) model (Zhang, 2003). In this way, the predictor is based on a auto-regressive model of order $p = 12$, this is $r(z_k) = z_k = [y_{k-1}\ y_{k-2} \cdots y_{k-12}]^T$.

The lynx series plotted in Figure 5.5a has 114 observations, the first 80 observations of this data set were used as training set and the last 34 as test set.

A log with base 10 was applied to the series in order to make a symmetrical data set, the plot of the data set is at Figure 5.5b:



(A) Original Time series (B) Transformed Time series

FIGURE 5.5: Annual number of lynx trappings in Canada from 1821 to 1934.

Figure 5.5 shows the forecasts of the proposed predictor by forecasting horizons with the hyper-parameters selected in both error measures. As a result, different optimal hyper-parameters are found in the error measure selection on the training set.

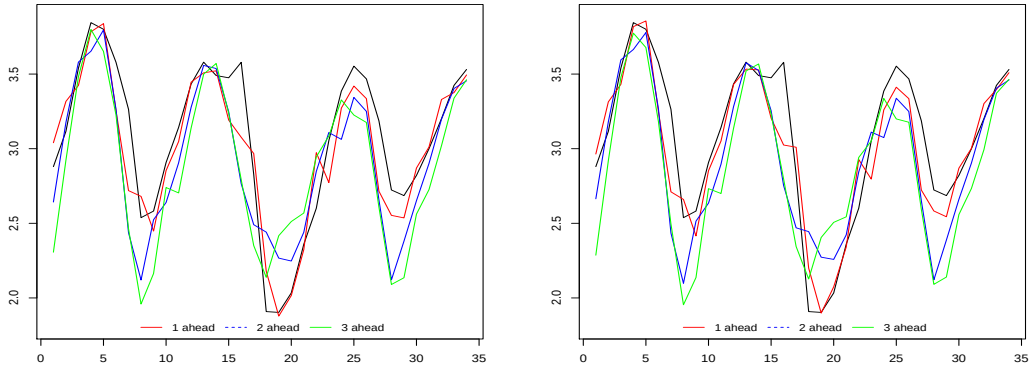
Table 5.3 shows results that comes from different optimal gamma selections by different error criteria and forecasting horizons.

TABLE 5.3: Canadian Lynx time series optimal gamma.

Ahead	γ_mape	γ_smape
1.00	0.01	0.06
2.00	0.00	0.02
3.00	0.02	0.04

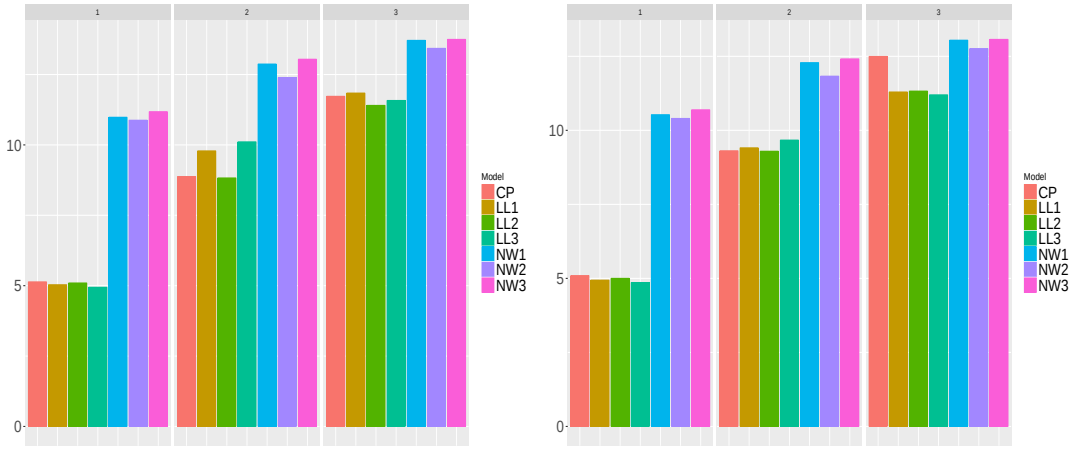
Corresponding to this predictions, the error measures are shown on Table D.2 in the appendix. To sum up this table in a graphical way, the figure 5.7 plot the error measures by predictor and prediction horizon.

Results in figure 5.7 show that by selecting an appropriate kernel on Local Linear regression helps to get closer results in a short prediction term to the proposed predictor CP. However, in a three step-ahead prediction horizon the results mark a tie



(A) MAPE selection criterion (B) SMAPE selection criterion

FIGURE 5.6: Canadian lynx time series predictions by forecasting horizon in the test set.



(A) SMAPE (B) MAPE

FIGURE 5.7: Mean of test set errors by forecasting horizon in Canadian lynx time series.

between LL and CP in the $SMAPE$, on the contrary by selecting $MAPE$ criterion the LL outperforms, this results supports the use of the proposed predictor for a short term prediction.

5.5.2.3 Monthly critical radio frequencies

Monthly critical radio frequencies in Washington, D.C., contains the highest radio frequency that can be used for broadcasting from May 1934 to April 1954 (Newton, 1988).

This time series plotted in Figure 5.8 has 240 observations, the first 216 observations were used as training set and the last 24 as a test set.

According to auto-correlation plot attached in the Figure 5.9 at the appendix, the established model is based on an auto-regressive model of order twelve, which has also been used by many researchers (Rao and Gabr, 1984; Zhang, 2003). The auto-regressive model has the shape like $r(z_k) = z_k = [y_{k-1}\ y_{k-2}\ ...\ y_{k-12}]^T$.

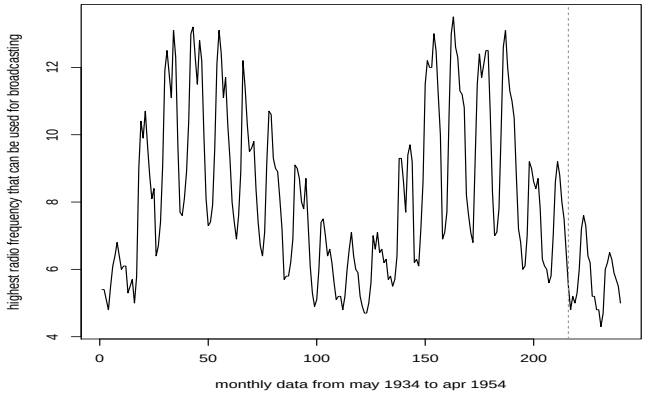


FIGURE 5.8: Monthly critical radio frequencies (1934−1954).

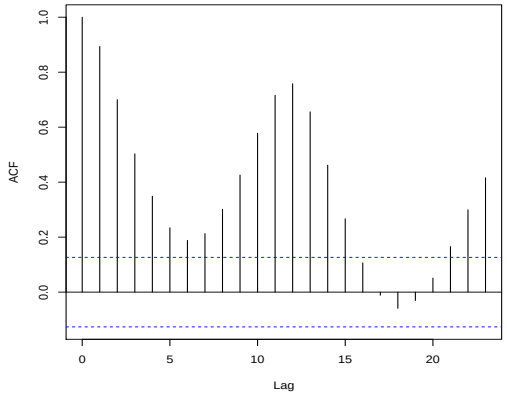


FIGURE 5.9: Auto-correlation function of Monthly critical radio frequencies time series.

Prediction are shown in figures 5.10 for the proposed predictor by forecasting horizons with the hyper-parameters selected in both error measures. In this case, different optimal hyper-parameter are found in the different error measure criteria selected on the training set.

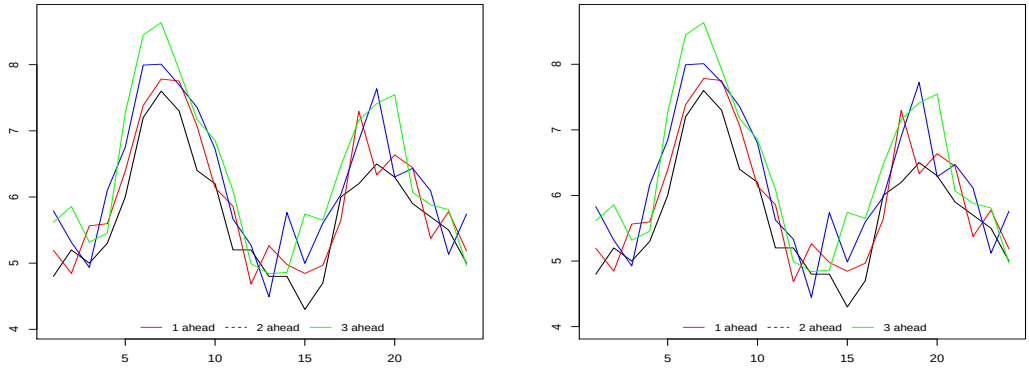
Prediction error results of this time series are gathered on Table D.3 in the appendix, and table 5.4 shows the optimal selected gamma.

TABLE 5.4: Monthly critical radio frequencies time series optimal gamma.

Ahead	γ_mape	γ_smape
1.00	0.00	0.00
2.00	0.03	0.00
3.00	0.00	0.00

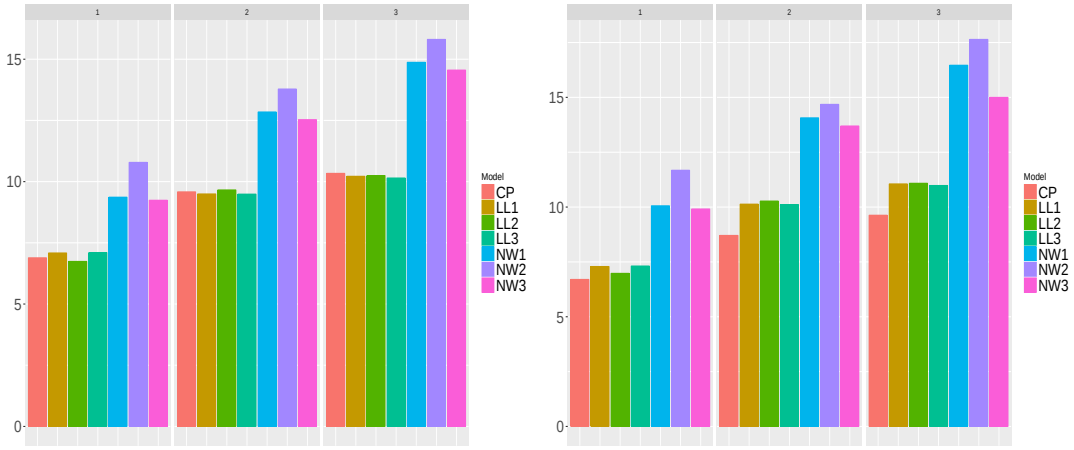
To sum up the table D.3 in a graphical way, figure 5.11 show the error measures by predictor and forecasting horizon.

The proposed predictor get similar results to Local Linear regression in SMAPE measure in the proposed forecast horizons, on the contrary by selecting $MAPE$ measure the Proposed predictor CP outperforms. Besides, results shows that there is not a high significant variation of results with the selection of the different kernels in Local Linear Regression.



(A) MAPE selection criterion (B) SMAPE selection criterion

FIGURE 5.10: Monthly critical radio frequencies time series prediction by forecasting horizon in the test set.



(A) SMAPE (B) MAPE

FIGURE 5.11: Mean of test set errors by forecasting horizon in monthly critical radio frequencies time series.

5.5.2.4 Monthly pneumonia and influenza deaths

Monthly pneumonia and influenza deaths per 10.000 people in the United States for 11 years, 1968 to 1978.

This time series plotted in Figure 5.12 has 132 observations, the first 84 observations were used as training set and the last 24 as a test set.

Figure 5.13 of auto-correlation at the appendix shows a seasonality of approximately 12 months. In this line the predictor is based on a auto-regressive model of order $p = 12$, this is $r(z_k) = z_k = [y_{k-1}\ y_{k-2}\ ...\ y_{k-12}]^T$.

Figure 5.14 shows the forecasts of the proposed predictor by forecasting horizons with the hyper-parameters selected in both error measures.

Results of this time series are shown on Table D.4 at the appendix. The hyper-parameter γ is selected in the training set where the error is minimum, the value of γ is inferred to perform forecasts in the test set as shown in table 5.5.

To sum up table D.4 in a graphical way, figure 5.15 plots the error measures by predictor and prediction horizon.

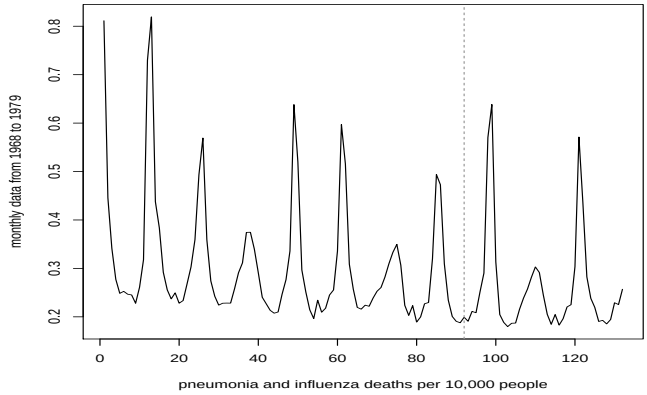


FIGURE 5.12: Monthly pneumonia and influenza deaths (1968−1978).

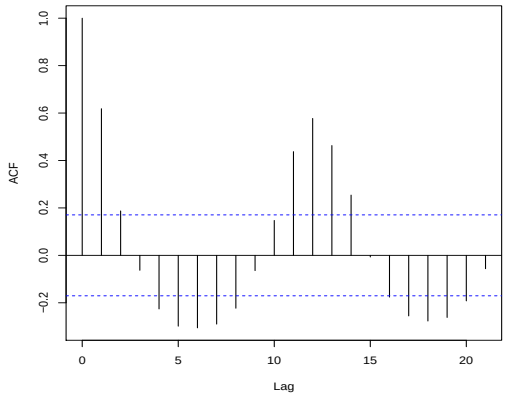


FIGURE 5.13: Auto-correlation function of Monthly pneumonia and influenza deaths time series.

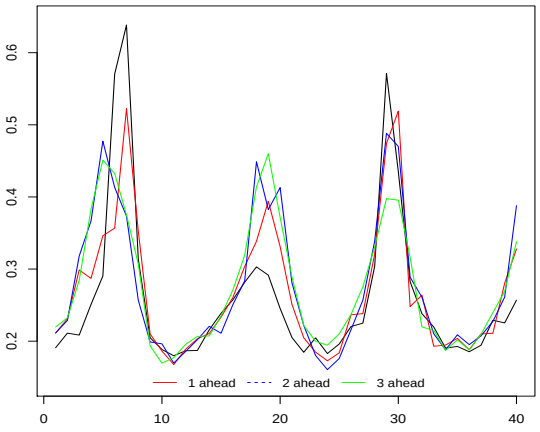
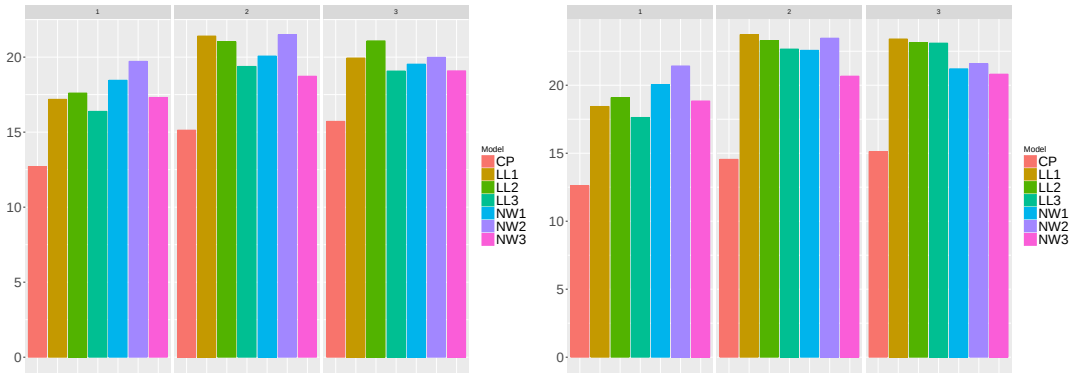


FIGURE 5.14: Monthly pneumonia and influenza deaths time series predictions by forecasting horizon in the test set.

TABLE 5.5: Monthly pneumonia and influenza deaths time series optimal gamma.

Ahead	γ_mape	γ_smape
1.00	0.50	0.50
2.00	0.22	0.22
3.00	0.08	0.08



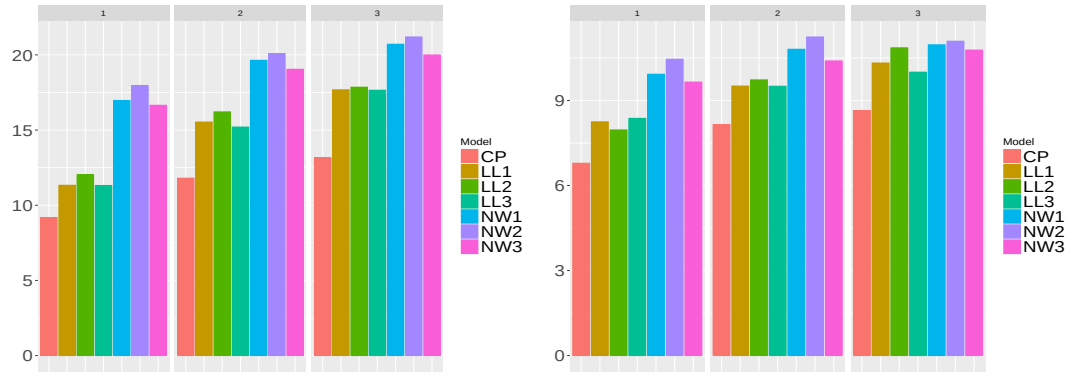
(A) SMAPE (B) MAPE

FIGURE 5.15: Mean of test set errors by forecasting horizon in monthly pneumonia and influenza deaths time series.

Results shows by forecasting horizon that the proposed predictor outperforms in both error measures in the proposed forecast horizons. Besides, the results shows that tricube kernel is a suitable option that outperforms between the other selected non parametric methods for a short term forecast.

5.5.2.5 Averaged results

In order to provide an overview of the results, this subsection provides an averaged results in tables D.1, D.2, D.3 and D.4 attached at the appendix. This summary is an aggregated mean of out of sample errors for the aforementioned time series. As a result, two aggregated error measures are obtained. Results shows that in average the proposed predictor outperforms the selected methods in the different results.



(A) SMAPE (B) MAPE

FIGURE 5.16: Mean of test set errors by forecasting horizon

Figures from 5.17 to 5.19 shows aggregated mean by forecasting horizon. results shows that the proposed predictor is outperforming.

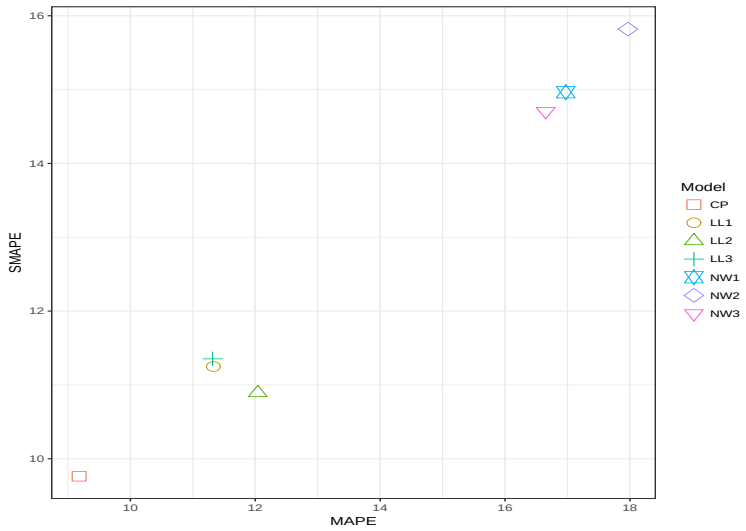


FIGURE 5.17: Mean of SMAPE and MAPE results for 1 step-ahead forecasts

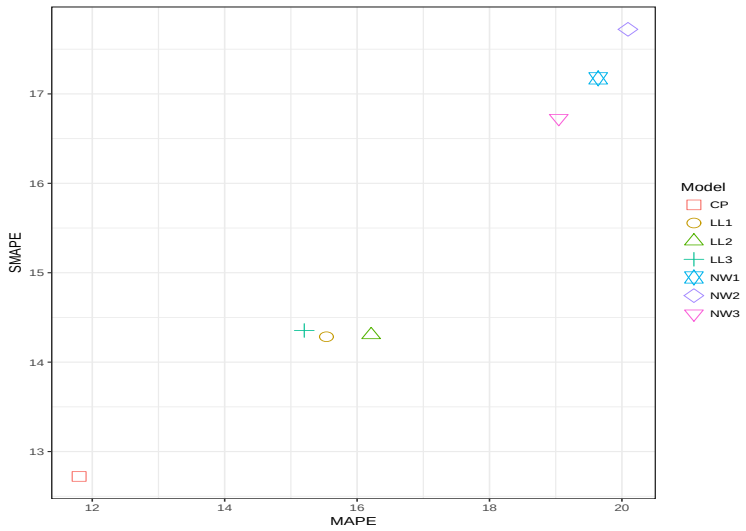


FIGURE 5.18: Mean of SMAPE and MAPE results for 2 step-ahead forecasts

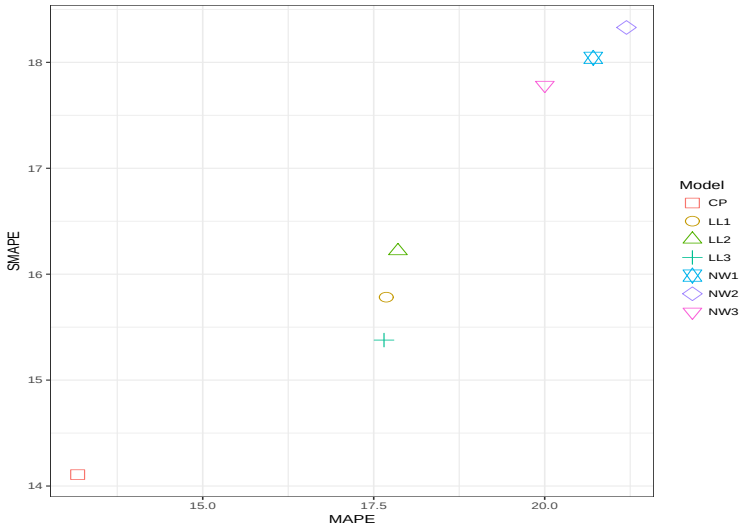


FIGURE 5.19: Mean of SMAPE and MAPE results for 3 step-ahead forecasts

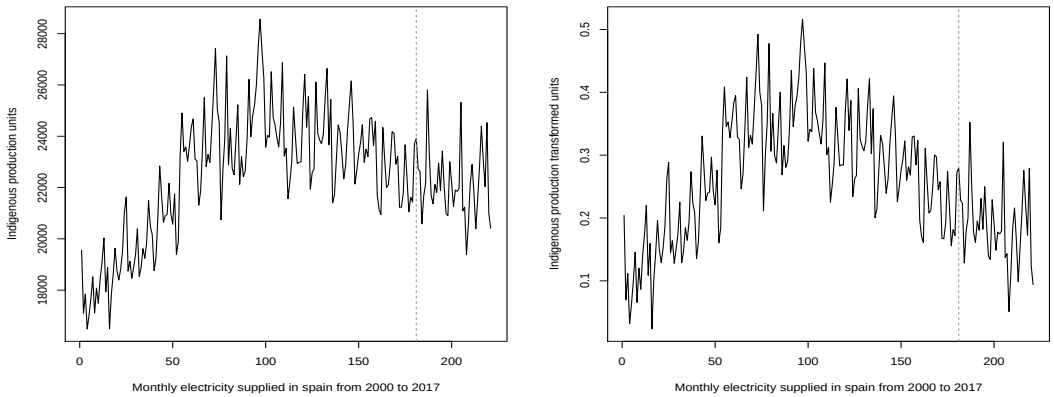
5.5.3 Monthly electricity supplied

In this subsection we select a dataset a from the International Energy Association (IEA), this institution provides monthly statistics with timely and consistent oil, oil price, natural gas and electricity data for all Organization for Economic Co-operation and Development member countries.

Countries submitted monthly data is adjusted proportionately to maintain consistency with the most recent annual data for each generation source.

The time series is the electricity supplied for Spain from January of 2000 to May of 2017, defined as the Indigenous production plus Imports minus Exports. It includes transmission and distribution losses.

Figure 5.20 plots the aforementioned data set which has 221 observations, the first 181 observations were used as training set and the last 40 as a test set.



(A) Original Time series (B) Transformed Time series

FIGURE 5.20: Monthly electricity supplied in Spain (2000−2017).

Figure 5.21 of auto-correlation at the appendix shows a seasonality of approximately 12 months. In this line the predictor is based on a auto-regressive model of

order $p = 12$, this is $r(z_k) = z_k = [y_{k-1}\ y_{k-2}\ ...\ y_{k-12}]^T$.

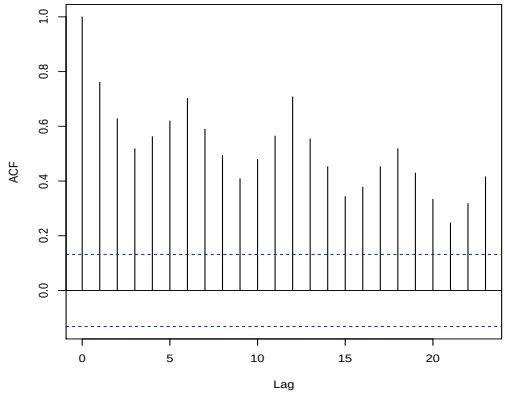
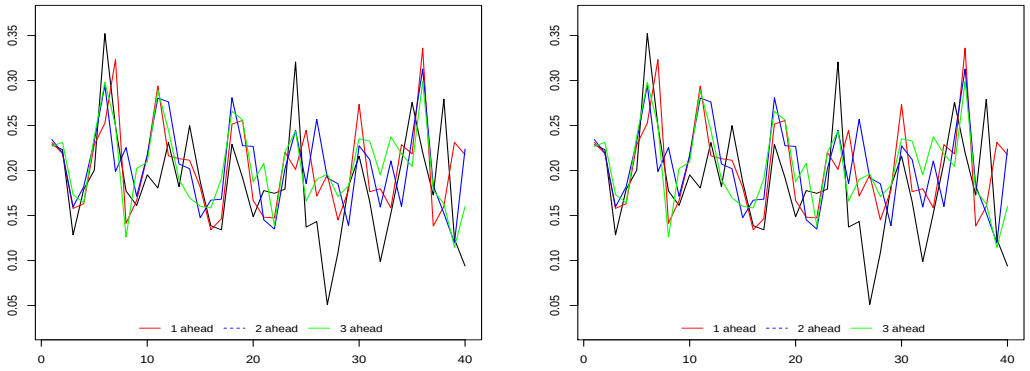


FIGURE 5.21: Auto-correlation function of Monthly electricity supplied in Spain.

The forecasts for the proposed predictor are shown in figure 5.22 and are plotted by forecasting horizons with the hyper-parameters selected in both error measures.



(A) MAPE selection criterion

(B) SMAPE selection criterion

FIGURE 5.22: Monthly electricity supplied time series predictions by forecasting horizon in the test set.

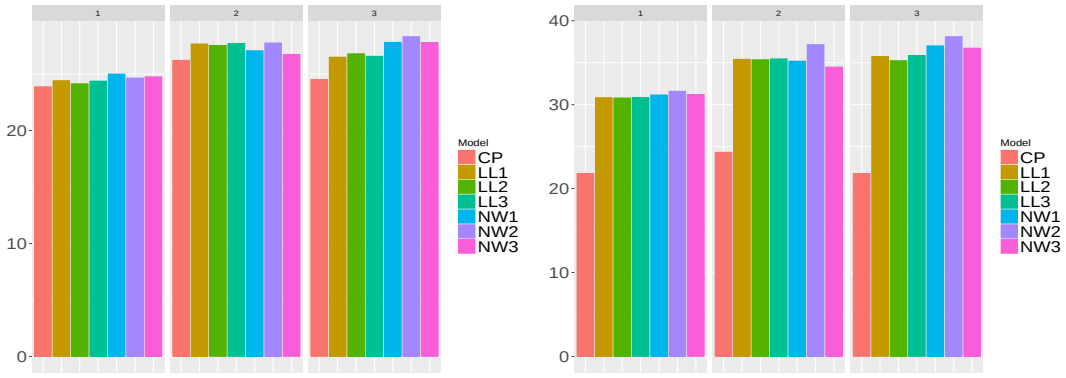
Results of this time series are shown on Table D.5 at the appendix. The hyper-parameter γ is selected in the training set where the error is minimum and the value of γ is inferred to perform forecasts in the test set as shown in table 5.6.

TABLE 5.6: Monthly electricity supplied time series optimal gamma.

Ahead	γ_mape	γ_smape
1.00	0.06	0.07
2.00	0.26	0.25
3.00	0.09	0.09

To sum up table D.5 in a graphical way, figure 5.23 plots the error measures by predictor and prediction horizon.

Results from Figure 5.23 shows that the proposed predictor outperforms in both error measures in the proposed forecast horizons. The proposed predictor is a suitable option to consider in a real life problem.



(A) SMAPE (B) MAPE

FIGURE 5.23: Mean of test set errors by forecasting horizon in Monthly electricity supplied time series.

5.6 Conclusions

Time series are often used to support statistical models. The purpose of these models is twofold: to explain the past and to forecast the future. In the supply chain, forecasting future demand is necessary to determine purchase and production orders and to minimize the risks of stock surpluses.

Time series are, therefore, a fairly versatile abstraction and a basic statistical tool. However, their apparent simplicity can be misleading, as several factors can alter the way in which the data are presented or account for important variations in the data. Knowing how the data are collected and accounting for the factors that affect trends is essential, being especially simplified in univariate approaches since forecasts assume that the future is modeled solely by past trends of the same study series without external information.

A novel non-parametric Time Series forecasting method has been proposed. The prediction is obtained by a weighted sum of past observations. A combination of deterministic and stochastic assumptions are used to obtain an expression of the outer bound of the prediction error. The weights are obtained solving a convex optimization problem that minimizes the upper bound of the prediction error. The method includes a tuning hyper-parameter. This hyper-parameter may balance the deterministic and stochastic considered assumptions. By a cross-validation scheme, a suitable hyper-parameter can be obtained. The performance of the proposed predictor is exposed by some datasets.

Chapter 6

Conclusions and limitations

In this work, we have explored different applications to accurately model reality and take advantage of its interpretability. Current approaches to machine learning and artificial intelligence like deep learning are black boxes. These systems generate predictions based on billions of calculations with no explanations. It is impossible for a human to follow the logic and understand the rationale behind the prediction. There are many open questions in creating the formal links between applications, the science of human understanding, and modern machine learning models. Besides, it is not just in academia, policymakers and businesses as well have realized that interpretability is a key to warding off potential dangers arising from the notoriously instable ML models being deployed in businesses, public health, criminal justice, and so on.

In the economic and financial sector, regarding the term structure, long-term rates could explain changes in future short-term rates. Understanding the term structure and hence, the yield curve, our goal is to create an interpretable forecasting model that is able enough accurately inform us about future recessions, which could be a useful tool for practitioners, researchers, governments and central banks. for three main groups, the public sector, the private sector which are the households, banks and investors, and the Federal Reserve. From investors point of view, this information could be useful to take right decisions for investing considering different strategies regarding this information, as the expanding economic activity is correlated with the stock market expansion[66]. Federal Reserve by using the term spread to know in advance a possible economic recession could modify the interest rates in order to trying reduce the effect of this phenomenon.

In the economic field of economic recession detection, this thesis provides a novel contributions about relevant term spreads, 3-month - 6-month, 2-year–5-year,..., among others. Furthermore, with respect to these variables are computed some recession detection rules, metrics to detect intervals with higher recession probability and SHAP contribution values analysis. Implementing the necessary policy mix they can dampen the effects of the recession, minimize its duration, or steer the economy away from it all together. With this technique several descriptive conditions allow the user not only to understand this phenomenon but also to have indicators with the goal of detecting in order to minimize the magnitude of the effect of the recession.

As a future work suggestion, several paths can be followed. On one accuracy side, the improvement of the model predictive accuracy as it is relevant to have tools with high quality and impact on predicting this phenomenon. On the interpretability side, as different exogenous variables can be added, more study on the variable interactions can be performed in order to understand the yield curve inversion with other variables that are relevant for generating policies in order to preventing and controlling. On the rules generation side, as rules are potentially changing over the

time as variable importance also may variate as well, a predictive maintenance system could be proposed in order keep rules updated and valid over the time.

In recent years in the consumer economic sector, the concern about knowing which factors drive the demand for tobacco has grown due to the deleterious effect that tobacco has on public health. Based on data from a provincial panel on cigarette sales in Spain, the results of this article are adjusted to the existing evidence that price and income are variables capable of explaining the demand for cigarettes. The results also confirm that when measuring the effectiveness of tobacco price increases in reducing demand, it is important to consider the effect of income growth that can offset the effect of cigarette price increases. In other words, the findings of this work suggest that in times of economic growth, the price increases required to effectively reduce tobacco consumption in the population would be greater than the increases required under conditions of slow or no economic growth. Besides, a contribution of this thesis lies in indicating that the importance of affordability to control tobacco consumption in Spain has grown over time. Furthermore, until 2010, income has generally better explained the demand for cigarettes in the Spanish provinces. However, as of 2010, price is the explanatory variable of the demand function that best explains the behavior of the demand for cigarettes. In these circumstances, the separate estimates of price and income elasticity that have been carried out in Spain so far must be interpreted considering that as of 2010, price is more important than income in explaining the demand for cigarettes. This means that, although the demand functions estimated so far are useful to make predictions about the behavior of cigarette demand, the government must consider that price is a good tool to control tobacco consumption from a certain point of affordability. In other words, for the Spanish government, the price is a more powerful way to control tobacco consumption from 2010 onwards. To our knowledge, this is the first attempt to obtain estimates of the explanatory power of the main elements of the tobacco demand function.

Additionally, this thesis provides an innovative tool using machine learning in the economic field of detecting illicit trade, specifically in the tobacco industry. Thesis publication findings are important because they show that cigarette sales in Spain are conditioned by the effect of tourism and by the price differential with border countries. Along these lines, cooperation between countries in tobacco control policies can have better effects than policies developed based on information from a single country. The lack of control over the transactions of tourists and inhabitants of border countries can cause important anomalies that distort the vision that governments have on tobacco consumption based on official data.

Finally, a novel non-parametric Time Series forecasting method has been proposed. The prediction is obtained by a weighted sum of past observations. A combination of deterministic and stochastic assumptions are used to obtain an expression of the outer bound of the prediction error. The weights are obtained solving a convex optimization problem that minimizes the upper bound of the prediction error. The method includes a tuning hyper-parameter. This hyper-parameter may balance the deterministic and stochastic considered assumptions. By a cross-validation scheme, a suitable hyper-parameter can be obtained. The performance of the proposed predictor is exposed by some datasets.

Appendix A

Appendix Chapter 2

A.1 Descriptive statistics

TABLE A.1: Pearson correlation coefficient for the most correlated variable.

Variable	Correlated	Correlation
Y1-Y10	Y20-M6	-0.98
Y20-M6	Y1-Y10	-0.98
Y1-Y7	Y20-M6	-0.97
Y1-Y5	Y10-M6	-0.97
Y10-M6	Y1-Y5	-0.97
Y1-Y20	Y20-M6	-0.97
Y7-M6	Y1-Y5	-0.97
Y2-Y5	Y20-M6	-0.96
Y20-M3	Y1-Y7	-0.96
Y2-Y7	Y20-M6	-0.96
Y10-M3	Y1-Y5	-0.96
Y1-Y3	Y7-M6	-0.96
Y5-M6	Y1-Y3	-0.96
Y7-M3	Y1-Y3	-0.95
Y1-Y2	Y5-M6	-0.94
Y2-Y10	Y20-M6	-0.94
Y2-Y3	Y20-M6	-0.94
Y5-M3	Y1-Y3	-0.94
Y3-Y5	Y20-M6	-0.93
Y3-Y7	Y20-M6	-0.93
Y3-M6	Y1-Y2	-0.92
Y2-Y20	Y20-M6	-0.91
Y3-Y10	Y20-M6	-0.90
Y3-M3	Y1-Y2	-0.90
Y5-Y7	Y20-M6	-0.87
Y3-Y20	Y20-M6	-0.87
Y5-Y10	Y20-M6	-0.83
Y2-M6	Y1-Y2	-0.81
Y5-Y20	Y20-M6	-0.80
Y2-M3	Y1-Y2	-0.79
Y1-M3	M3-M6	-0.75
M3-M6	Y1-M3	-0.75
Y7-Y20	Y20-M6	-0.73
Y7-Y10	Y20-M6	-0.73
Y10-Y20	Y20-M6	-0.65
Y1-M6	M3-M6	-0.44

TABLE A.2: Term spread descriptive statistics.

Feature	mean	median	min	max	sd
Y1-Y2	-0.29	-0.31	-1.06	0.95	0.34
Y1-Y3	-0.46	-0.51	-1.63	1.77	0.55
Y1-Y5	-0.75	-0.77	-2.50	2.35	0.81
Y1-Y7	-0.98	-1.02	-2.87	2.82	0.99
Y1-Y10	-1.14	-1.18	-3.40	3.07	1.15
Y1-Y20	-1.42	-1.33	-4.15	3.33	1.38
Y1-M3	0.52	0.43	-0.94	2.93	0.44
Y1-M6	0.39	0.31	-0.39	1.60	0.32
Y2-Y3	-0.17	-0.17	-0.59	0.83	0.22
Y2-Y5	-0.49	-0.46	-1.55	1.41	0.53
Y2-Y7	-0.74	-0.71	-2.28	1.88	0.72
Y2-Y10	-0.93	-0.85	-2.83	2.13	0.91
Y2-Y20	-1.30	-1.14	-3.67	2.39	1.19
Y2-M3	0.79	0.72	-1.76	3.86	0.66
Y2-M6	0.69	0.61	-0.82	2.44	0.54
Y3-Y5	-0.30	-0.27	-0.99	0.58	0.31
Y3-Y7	-0.53	-0.50	-1.72	1.05	0.52
Y3-Y10	-0.69	-0.60	-2.36	1.30	0.71
Y3-Y20	-1.01	-0.82	-3.27	1.56	1.00
Y3-M3	0.97	0.98	-2.01	4.11	0.80
Y3-M6	0.85	0.86	-1.20	2.74	0.69
Y5-Y7	-0.23	-0.21	-0.76	0.47	0.22
Y5-Y10	-0.39	-0.31	-1.46	0.72	0.42
Y5-Y20	-0.73	-0.60	-2.47	1.25	0.72
Y5-M3	1.27	1.33	-2.25	4.33	0.99
Y5-M6	1.15	1.20	-1.56	3.12	0.89
Y7-Y10	-0.16	-0.11	-0.74	0.38	0.22
Y7-Y20	-0.50	-0.42	-1.80	0.84	0.52
Y7-M3	1.50	1.56	-2.49	4.46	1.12
Y7-M6	1.37	1.43	-2.03	3.31	1.03
Y10-Y20	-0.34	-0.34	-1.06	0.87	0.34
Y10-M3	1.66	1.74	-2.65	4.42	1.24
Y10-M6	1.53	1.59	-2.28	3.64	1.16
Y20-M3	1.93	2.01	-3.00	4.44	1.41
Y20-M6	1.80	1.79	-2.54	4.36	1.35
M3-M6	-0.12	-0.10	-1.45	1.01	0.19

A.2 Variable importance results

TABLE A.3: SHAP values for train and test set.

Feature	Training	Test
M3-M6	0.2095	0.2217
Y2-Y5	0.1571	0.1986
Y5-Y10	0.1674	0.1394
Y3-Y7	0.0837	0.1012
Y3-M3	0.0939	0.1002
Y2-M6	0.1022	0.0949
Y1-M6	0.1745	0.0824
Y1-Y20	0.0877	0.0778
Y2-Y10	0.071	0.0778
Y1-Y10	0.0422	0.0699
Y2-Y3	0.0442	0.0474
Y1-Y2	0.0827	0.0445
Y5-M3	0.0355	0.0432
Y3-Y10	0.0337	0.0431
Y10-Y20	0.0731	0.0403
Y5-Y7	0.0536	0.0403
Y7-Y10	0.0363	0.0403
Y1-Y3	0.0292	0.0381
Y5-Y20	0.0545	0.0321
Y2-Y7	0.0272	0.0278
Y3-Y5	0.0213	0.0262
Y7-Y20	0.0585	0.025
Y2-M3	0.0294	0.0248
Y1-M3	0.0675	0.0244
Y1-Y7	0.0134	0.0194
Y10-M6	0.0145	0.0188
Y1-Y5	0.0209	0.0186
Y10-M3	0.011	0.017
Y2-Y20	0.0081	0.0116
Y7-M3	0.0065	0.0096
Y20-M3	0.0173	0.0093
Y3-M6	0.0036	0.0087
Y7-M6	0.0076	0.0078
Y3-Y20	0.0137	0.0067
Y20-M6	0.0108	0.0066
Y5-M6	0.0015	0.0007

Appendix B

Appendix Chapter 3

B.1 Descriptive statistics

TABLE B.1: Descriptive statistics of the data used.

Province	Years	Per capita cigarette sales*					Price*					Per capita GDP*				
		Mean	SD	Quartile			Mean	SD	Quartile			Mean	SD	Quartile		
				Q1	Q2	Q3			Q1	Q2	Q3			Q1	Q2	Q3
Albacete	16	93.45	25.84	64.12	101.18	115.86	3.07	1.19	1.93	2.93	4.39	19.72	3.34	18.01	21.27	21.47
Alicante	16	135.80	55.05	78.74	127.64	184.19	3.10	1.15	2.02	2.93	4.36	19.75	2.21	19.37	20.19	21.01
Almería	16	114.20	38.21	72.19	118.10	146.96	3.14	1.17	2.03	2.98	4.41	21.59	2.59	20.89	21.92	23.40
Álava	16	81.18	22.34	56.98	84.07	101.80	3.06	1.21	1.92	2.91	4.37	36.26	6.96	32.92	38.29	40.87
Asturias	16	88.74	21.88	64.93	95.44	107.41	3.06	1.19	1.91	2.90	4.31	20.90	3.56	19.37	22.41	23.12
Ávila	16	93.47	25.53	65.35	99.97	115.41	3.08	1.20	1.94	2.92	4.37	18.93	3.27	16.96	20.30	21.09
Badajoz	16	98.68	28.85	66.72	110.11	123.38	3.04	1.20	1.87	2.91	4.29	17.53	2.96	16.05	18.74	19.34
Islas Baleares	16	168.33	72.05	95.93	150.77	228.62	3.13	1.15	2.05	2.93	4.37	27.29	3.66	26.06	28.11	29.15
Barcelona	16	87.59	27.11	58.16	89.84	109.03	3.04	1.20	1.89	2.86	4.34	28.75	5.25	26.19	30.21	31.28
Vizcaya	16	77.93	18.21	58.77	81.43	92.03	3.06	1.21	1.92	2.91	4.37	28.92	5.62	25.51	30.90	31.83
Burgos	16	87.44	23.79	61.89	91.81	109.79	3.05	1.21	1.89	2.88	4.33	26.61	4.62	24.13	28.33	29.22
Cáceres	16	99.28	26.80	69.04	109.19	121.22	3.04	1.19	1.89	2.90	4.34	17.25	3.14	15.54	18.33	18.86
Cádiz	16	75.60	33.69	37.81	82.92	107.68	3.04	1.19	1.89	2.90	4.32	18.82	2.40	18.64	19.65	20.27
Cantabria	16	93.99	30.05	65.44	102.02	118.87	3.08	1.20	1.93	2.91	4.41	22.51	3.54	20.86	23.99	24.55
Castellón	16	102.53	33.67	66.20	103.11	133.72	3.07	1.17	1.94	2.92	4.36	25.61	3.68	24.76	26.19	27.02
Ciudad Real	16	96.62	26.99	66.04	105.98	120.09	3.06	1.20	1.91	2.91	4.37	20.66	3.31	19.17	21.80	22.65
Córdoba	16	88.33	32.20	50.46	98.48	115.96	3.03	1.21	1.88	2.88	4.35	17.91	2.97	16.83	18.99	19.45
La Coruña	16	81.50	20.58	58.94	87.07	98.53	3.05	1.20	1.89	2.89	4.35	21.68	4.30	19.17	23.58	23.98
Cuenca	16	98.42	25.96	68.50	106.47	121.25	3.08	1.20	1.94	2.93	4.42	20.53	3.97	18.54	21.77	22.65
Guipúzcoa	16	146.87	51.79	95.44	144.96	193.39	3.03	1.21	1.85	2.86	4.32	31.07	5.47	28.36	33.18	33.92
Gerona	16	267.40	97.35	169.15	261.90	358.11	3.05	1.20	1.87	2.86	4.36	29.22	4.39	27.97	30.58	31.27
Granada	16	99.06	31.07	63.37	105.28	126.55	3.08	1.19	1.94	2.93	4.38	18.28	2.97	17.18	19.32	20.06
Guadalajara	16	93.49	28.64	61.79	96.66	116.66	3.08	1.18	1.97	2.92	4.38	20.93	2.69	20.53	21.98	22.47
Huelva	16	113.75	41.13	65.68	125.89	150.66	3.04	1.19	1.88	2.89	4.31	19.40	2.70	18.84	19.91	21.07
Huesca	16	116.51	33.40	79.16	124.19	147.17	3.07	1.20	1.92	2.90	4.35	26.90	5.23	23.35	28.87	30.21
Jaén	16	95.73	27.76	64.02	106.05	118.55	3.05	1.19	1.91	2.93	4.31	17.70	2.77	16.33	18.62	19.38
León	16	84.99	21.14	61.87	92.17	103.14	3.08	1.21	1.92	2.91	4.35	19.93	3.27	18.45	21.61	21.93
Lleida	16	140.99	52.46	85.33	144.39	188.71	3.02	1.20	1.84	2.84	4.34	29.59	4.90	26.65	31.15	33.19
Lugo	16	73.08	15.41	56.07	78.52	87.09	3.06	1.20	1.92	2.89	4.37	20.07	4.33	17.87	20.89	22.89
Madrid	16	88.05	27.11	59.71	90.11	108.47	3.07	1.19	1.93	2.91	4.35	33.79	5.86	30.85	35.65	36.87
Málaga	16	113.73	50.43	60.59	114.07	160.94	3.10	1.17	2.01	2.92	4.34	19.02	2.68	18.81	20.02	20.43
Murcia	16	107.57	32.88	71.37	111.70	136.50	3.09	1.17	1.99	2.94	4.35	21.52	3.36	20.33	22.57	23.18
Navarra	16	139.86	40.06	97.19	148.17	174.95	3.05	1.19	1.89	2.88	4.29	30.88	4.81	28.90	32.48	33.60
Orense	16	73.21	14.59	57.04	80.84	86.36	3.07	1.20	1.92	2.90	4.38	18.59	3.54	16.47	19.68	20.82
Palencia	16	89.41	23.05	64.70	96.01	108.03	3.07	1.20	1.92	2.90	4.35	24.18	4.14	21.85	25.57	26.43
Pontevedra	16	78.35	21.73	53.76	84.15	99.32	3.05	1.19	1.89	2.89	4.36	20.62	3.67	19.11	21.86	22.74
La Rioja	16	87.69	22.32	63.63	90.91	106.82	3.06	1.19	1.93	2.90	4.37	26.56	4.36	24.49	28.18	28.94
Salamanca	16	84.75	24.26	57.80	94.72	106.90	3.08	1.20	1.92	2.92	4.29	19.64	2.85	18.45	20.62	20.98
Segovia	16	86.75	25.42	58.20	91.45	109.43	3.08	1.20	1.93	2.91	4.36	22.86	3.11	22.15	23.92	24.44
Sevilla	16	86.20	38.24	42.25	95.66	120.93	3.05	1.19	1.90	2.90	4.28	20.76	3.22	19.61	22.07	22.61
Soria	16	83.82	20.09	61.58	89.94	99.13	3.11	1.20	1.96	2.96	4.39	23.91	3.88	21.42	25.48	26.35
Tarragona	16	115.17	41.05	71.29	112.86	153.18	3.11	1.16	2.01	2.94	4.40	30.05	4.15	27.91	31.13	31.66
Teruel	16	89.73	21.70	65.42	95.51	108.95	3.09	1.21	1.95	2.94	4.40	25.17	3.99	23.31	26.90	27.96
Toledo	16	97.11	30.58	62.88	103.27	124.87	3.07	1.18	1.94	2.90	4.35	19.55	2.65	19.44	20.30	20.93
Valencia	16	98.39	31.02	64.17	102.72	123.52	3.01	1.18	1.88	2.86	4.30	23.34	3.54	21.68	24.57	25.53
Valladolid	16	85.01	24.65	58.01	89.36	105.15	3.05	1.20	1.90	2.88	4.36	24.50	4.12	22.49	26.03	26.62
Zamora	16	79.91	19.41	58.84	87.73	95.88	3.06	1.20	1.91	2.89	4.33	18.27	3.48	16.13	19.31	20.69
Zaragoza	16	94.87	26.92	64.64	99.35	118.11	3.05	1.18	1.91	2.90	4.35	26.60	4.46	24.66	28.25	29.05
Spain	62	109.90	23.47	89.69	117.86	129.85	1.82	1.08	1.05	1.20	2.42	19.39	7.01	14.21	19.72	26.45

* Per capita sales are measured in packs of 20 cigarettes per year. The price is measured in real euros of 2016. GDP per capita is expressed in thousands of real euros of 2016.

B.2 Variable importance results

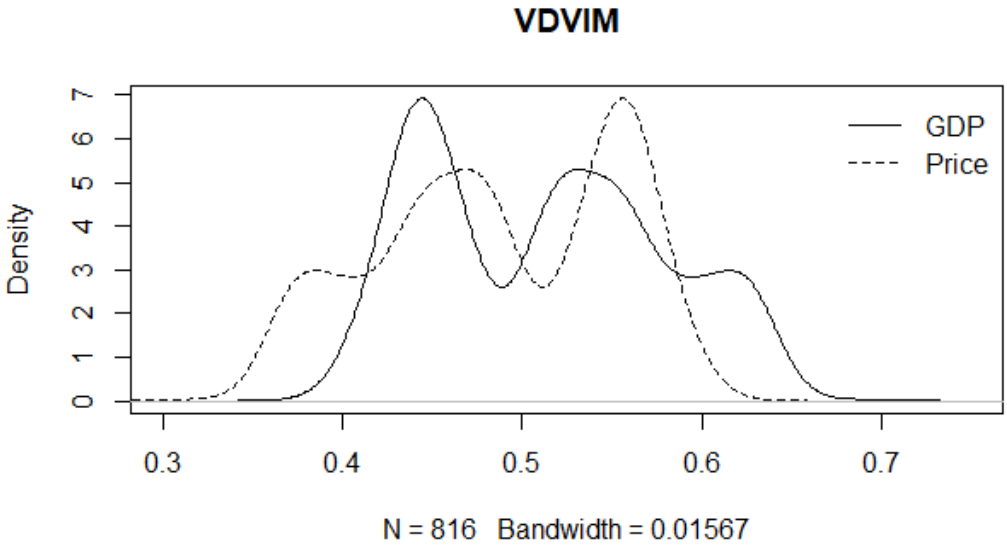
TABLE B.2: VIM statistics comparison for every model.

Model	VIM	Feature	mean	median	sd	mad
Ranger	Relative VDVIM	GDP	0.5124	0.5168	0.0666	0.0883
Ranger	Relative VDVIM	Price	0.4876	0.4832	0.0666	0.0883
QRF	Relative VDVIM	GDP	0.5201	0.5160	0.0791	0.0889
QRF	Relative VDVIM	Price	0.4799	0.4840	0.0791	0.0889
GBM	Relative VDVIM	GDP	0.5783	0.6069	0.2460	0.2712
GBM	Relative VDVIM	Price	0.4217	0.3931	0.2460	0.2712

TABLE B.3: Results of training and test set for error assessment.

Set	Model	MAE	MSE
Training	Ranger	0.1537	0.8849
Training	QRF	0.1561	0.0509
Training	GBM	0.1515	0.0581
Test	Ranger	0.1539	0.0495
Test	QRF	0.1527	0.0526
Test	GBM	0.1553	0.0594

FIGURE B.1: Density plots for VIM metrics for QRF (winning model).



Appendix C

Appendix Chapter 4

C.1 Model error measurements

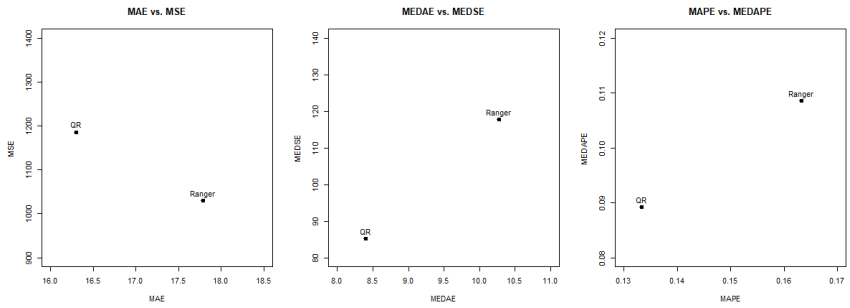
C.1.1 Training set error

The training set comprises the data without the province to predict for every year, at the following Figure C.1, it shows the averaged results (Table C.1) of the metrics presented in subsection 4.2.2.

TABLE C.1: Average metrics for the prediction at the training set.

MAE	MSE	MAPE	MEDAE	MEDSE	MEDAPE	MAE
QR	16.3	1185.9	0.13	8.4	85.19	0.09
Ranger	17.79	1029.36	0.16	10.27	117.78	0.11

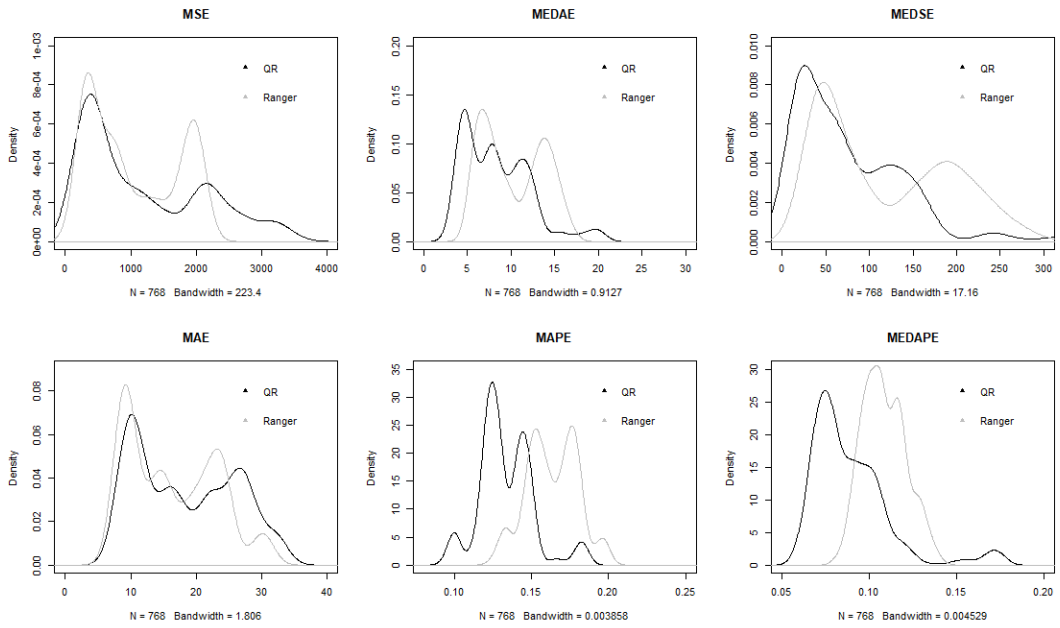
FIGURE C.1: Scatter plots for the error of fitted models at the training set.



As discussed on some research of error measurement, Meinshausen and Ridgeway, 2006, different statistical properties reflects every metric, with the errors shown on the previous table QR shows minimum error on training set.

In order to avoid the bias of the averaged metrics, the following density plots are shown in the Figure C.2. With this we can confirm the superiority of QR over Ranger at the training set.

FIGURE C.2: Density plots for the errors of fitted models at the training set.



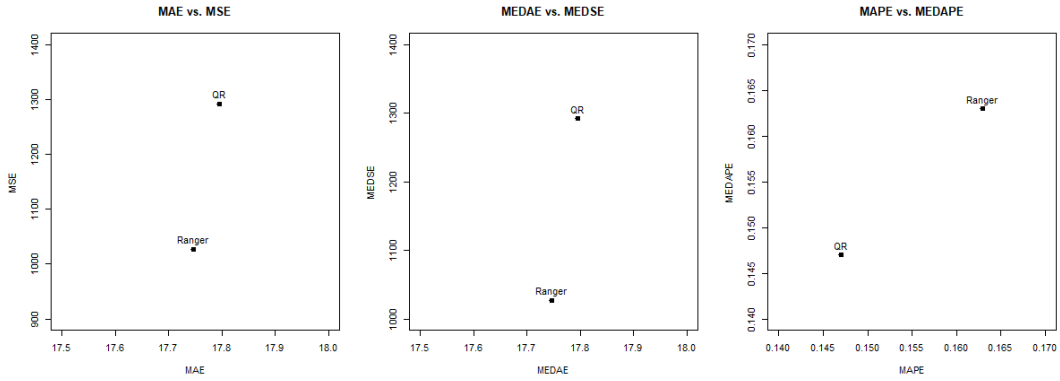
C.1.2 Test set error

The test set comprises the data with the province to predict for every year, at the Figure C.3, it shows the averaged results (Table C.2) of the metrics presented in subsection 4.2.2.

TABLE C.2: Average metrics for the prediction at the test set.

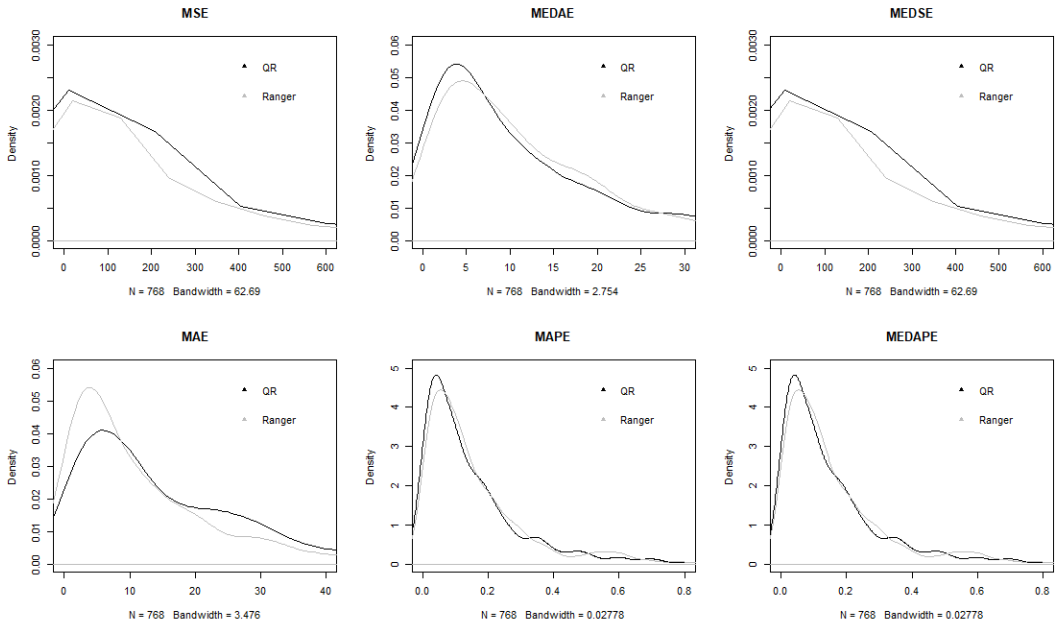
MAE	MSE	MAPE	MEDAE	MEDSE	MEDAPE	MAE
QR	17.8	1292.18	0.15	17.8	1292.18	0.15
Ranger	17.75	1027.63	0.16	17.75	1027.63	0.16

FIGURE C.3: Scatter plots for the fitted models at the training set.



At the test set the predictions errors are very similar, but the square metrics (MSE and MEDSE) penalizes the QR showing a slightly superiority of Ranger. Besides, the density plots shows the distribution of errors at the test set (Figure C.4), where the similarity of test errors are present with the difference at the MAE Error.

FIGURE C.4: Density plots for errors of the fitted models at the test set.



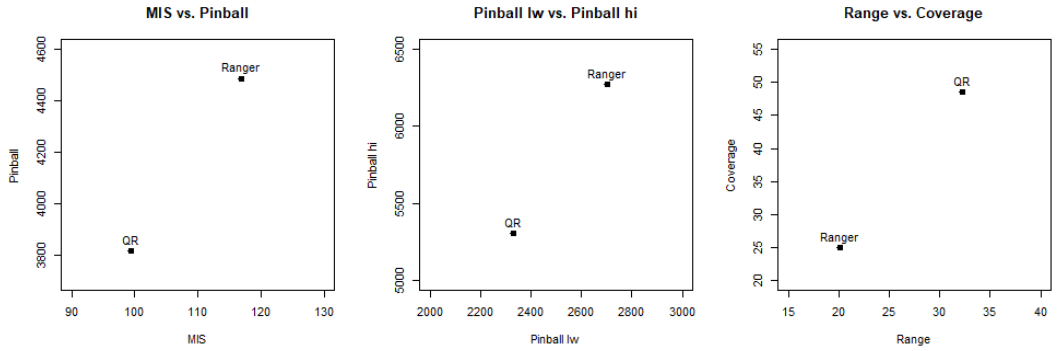
C.1.3 Interval Score Metrics

This subsection shows at Table C.3 metrics for assessing the prediction intervals which are the main novelty use for this work. By using these metrics, a wide overview of how intervals are fitted is potentially used to discard a method for abnormality detection and quantification as this work propose. The results of Table C.3 are shown visually at Figure C.5, where the superiority of QR over Ranger is present. The MIS and Pinball (average Pinball lw and Pinball hi) shows better performance of the intervals, also the pinball score for bot metrics. The Range and coverage of Ranger are smaller than QR, in this case is the intervals are wide enough to cover the regular points and having a better fit of QR intervals.

TABLE C.3: Averaged Interval metrics for the predicted intervals.

Model	MIS	Coverage	Range	$Pinball_{lw}$	$Pinball_{hi}$
QR	116.81	25.04	20.14	2702.23	6269.05
Ranger	99.41	48.52	32.22	2328.82	5305.85

FIGURE C.5: Density plots for errors of the fitted models at the test set.



Appendix D

Appendix Chapter 5

This appendix contains two sections containing mathematical derivations and results from tables and plots.

D.1 Mathematical derivations

Taking into account Assumption 1 and Definitions 3 and 2 the following equalities can be inferred

$$\begin{aligned}
\hat{e}_k(\Psi) &= y_{k+1} - \hat{y}_{k+1}(\Psi) && \text{(D.1)}\\
&= y_{k+1} - \Psi^T b_Y && \text{(D.2)}\\
&= r(z_k)^T \Phi_k - \Psi^T b_Y + e_k && \text{(D.3)}\\
&= (A^T \Psi)^T \Phi_k - \Psi^T b_Y + e_k && \text{(D.4)}\\
&= \Psi^T (A\Phi_k - b_Y) + e_k && \text{(D.5)}\\
&= \sum_{j=1}^{k} \Psi_j (r(z_{j-1})^T \Phi_k - y_j) + e_k && \text{(D.6)}\\
&= -\sum_{j=1}^{k} \Psi_j e_{j-1} + e_k. && \text{(D.7)}\\
& && \text{(D.8)}
\end{aligned}$$

QED

D.2 Tables

Model	Ahead	γ	tr_MAPE	te_MAPE	γ	tr_SMAPE	te_SMAPE
CP	1.00	**0.12**	15.75	13.04	**0.12**	15.46	14.97
LL1	1.00	1.57	15.19	14.65	1.84	14.34	15.69
LL2	1.00	1.55	14.82	17.11	1.62	13.77	14.13
LL3	1.00	1.89	16.42	15.47	2.33	14.04	16.98
NW1	1.00	1.08	17.69	27.26	1.08	13.82	21.05
NW2	1.00	1.02	17.36	28.38	1.02	13.48	21.90
NW3	1.00	1.24	17.91	27.16	1.26	13.98	21.09
CP	2.00	**0.14**	17.36	14.05	**0.14**	16.88	16.58
LL1	2.00	1.58	13.69	18.89	1.58	13.56	16.45
LL2	2.00	1.54	14.22	22.02	1.69	13.24	17.71
LL3	2.00	1.67	14.79	18.37	2.09	14.38	18.44
NW1	2.00	1.12	17.06	29.65	1.14	13.29	22.91
NW2	2.00	0.98	16.71	30.40	0.98	13.05	23.22
NW3	2.00	1.26	17.33	29.41	1.28	13.40	22.61
CP	3.00	**0.00**	17.43	15.45	**0.00**	16.50	18.66
LL1	3.00	1.64	16.06	24.99	2.15	15.59	21.13
LL2	3.00	1.61	17.89	25.89	2.01	15.37	22.14
LL3	3.00	1.78	16.49	25.33	2.34	15.72	20.72
NW1	3.00	1.12	17.16	32.13	1.14	13.23	24.05
NW2	3.00	0.98	16.88	32.80	0.98	13.13	24.09
NW3	3.00	1.24	17.55	31.12	1.30	13.37	23.75

TABLE D.1: Airline passengers time series results

Model	Ahead	γ	tr_MAPE	te_MAPE	γ	tr_SMAPE	te_SMAPE
CP	1.00	**0.01**	6.75	5.09	**0.06**	6.73	5.13
LL1	1.00	13.01	6.66	4.94	13.01	6.62	5.03
LL2	1.00	12.01	6.54	5.00	12.01	6.50	5.09
LL3	1.00	16.01	6.72	4.86	16.01	6.68	4.94
NW1	1.00	4.33	8.42	10.53	4.47	7.99	10.98
NW2	1.00	4.03	8.28	10.40	4.03	7.87	10.88
NW3	1.00	5.11	8.42	10.69	5.27	7.99	11.18
CP	2.00	**0.00**	10.08	9.31	**0.02**	10.07	8.88
LL1	2.00	11.01	9.94	9.41	11.01	9.85	9.79
LL2	2.00	11.01	9.82	9.29	12.01	9.66	8.83
LL3	2.00	12.01	9.99	9.67	12.01	9.89	10.11
NW1	2.00	4.76	9.45	12.29	4.77	8.91	12.87
NW2	2.00	4.31	9.33	11.83	4.31	8.79	12.39
NW3	2.00	5.49	9.46	12.41	5.66	8.92	13.04
CP	3.00	**0.02**	11.18	12.49	**0.04**	11.27	11.73
LL1	3.00	12.01	11.24	11.29	12.01	10.92	11.84
LL2	3.00	10.01	10.98	11.32	12.01	10.72	11.40
LL3	3.00	14.01	11.36	11.20	16.01	11.00	11.58
NW1	3.00	4.96	9.88	13.04	4.96	9.30	13.71
NW2	3.00	4.35	9.58	12.76	4.35	9.03	13.43
NW3	3.00	5.70	9.92	13.07	5.70	9.34	13.75

TABLE D.2: Canadian Lynx time series results

Model	Ahead	γ	tr_MAPE	te_MAPE	γ	tr_SMAPE	te_SMAPE
CP	1.00	**0.00**	7.33	6.70	**0.00**	7.26	6.89
LL1	1.00	50.00	7.25	7.29	50.00	7.23	7.09
LL2	1.00	30.00	7.17	6.98	30.00	7.18	6.74
LL3	1.00	60.00	7.26	7.31	60.00	7.24	7.10
NW1	1.00	13.00	8.92	10.06	13.00	8.92	9.36
NW2	1.00	13.00	9.09	11.68	13.00	9.04	10.79
NW3	1.00	15.00	8.86	9.91	15.00	8.85	9.24
CP	2.00	**0.03**	11.16	8.71	**0.00**	10.92	9.58
LL1	2.00	40.00	10.75	10.13	40.00	10.74	9.50
LL2	2.00	40.00	10.71	10.27	40.00	10.69	9.66
LL3	2.00	50.00	10.74	10.11	50.00	10.74	9.49
NW1	2.00	14.00	10.51	14.07	14.00	10.45	12.84
NW2	2.00	12.00	10.60	14.68	13.00	10.52	13.78
NW3	2.00	16.00	10.49	13.69	16.00	10.44	12.53
CP	3.00	**0.00**	12.42	9.63	**0.00**	12.25	10.34
LL1	3.00	50.00	12.26	11.05	50.00	12.14	10.22
LL2	3.00	40.00	12.18	11.08	40.00	12.07	10.25
LL3	3.00	50.00	12.24	10.98	50.00	12.14	10.15
NW1	3.00	14.00	11.65	16.46	14.00	11.53	14.87
NW2	3.00	13.00	11.74	17.64	13.00	11.57	15.81
NW3	3.00	15.00	11.61	15.00	16.00	11.52	14.56

TABLE D.3: Monthly critical radio frequencies time series results

Model	Ahead	γ	tr_MAPE	te_MAPE	γ	tr_SMAPE	te_SMAPE
CP	1.00	**0.50**	10.30	11.89	**0.50**	9.70	12.06
LL1	1.00	1.47	10.47	18.44	1.48	10.31	17.19
LL2	1.00	1.41	10.69	19.08	1.41	10.56	17.60
LL3	1.00	1.61	10.80	17.63	1.61	10.68	16.38
NW1	1.00	0.92	10.87	20.04	0.92	10.96	18.46
NW2	1.00	0.92	11.38	21.41	0.92	11.45	19.72
NW3	1.00	0.92	10.58	18.84	0.92	10.69	17.31
CP	2.00	**0.22**	12.81	15.14	**0.22**	12.04	15.85
LL1	2.00	1.95	13.99	23.72	1.48	13.95	21.40
LL2	2.00	1.90	13.54	23.27	1.90	13.77	21.02
LL3	2.00	1.61	14.70	22.66	1.55	14.08	19.38
NW1	2.00	1.01	11.22	22.56	0.96	11.24	20.07
NW2	2.00	0.95	11.65	23.46	0.95	11.67	21.50
NW3	2.00	0.92	10.84	20.67	0.92	10.78	18.73
CP	3.00	**0.08**	13.62	15.11	**0.08**	12.97	15.71
LL1	3.00	1.91	13.52	23.40	1.49	13.38	19.94
LL2	3.00	1.88	13.09	23.12	1.88	13.11	21.08
LL3	3.00	2.00	13.88	23.09	1.63	13.43	19.07
NW1	3.00	0.92	11.08	21.20	0.92	11.17	19.53
NW2	3.00	0.92	11.65	21.58	0.92	11.84	19.99
NW3	3.00	0.94	10.92	20.81	0.94	10.96	19.09

TABLE D.4: Monthly pneumonia and influenza deaths time series results

Model	Ahead	γ	tr_MAPE	te_MAPE	γ	tr_SMAPE	te_SMAPE
CP	1.00	**0.06**	13.88	21.81	**0.07**	14.16	23.88
LL1	1.00	1.90	17.08	30.84	2.00	14.52	24.42
LL2	1.00	1.87	16.93	30.80	1.50	14.40	24.14
LL3	1.00	2.00	17.14	30.86	2.00	14.62	24.38
NW1	1.00	0.68	17.76	31.16	0.73	15.00	25.00
NW2	1.00	0.65	17.58	31.60	0.65	15.04	24.65
NW3	1.00	0.78	17.81	31.21	0.81	14.98	24.76
CP	2.00	**0.26**	15.22	24.32	**0.25**	15.23	26.22
LL1	2.00	2.00	18.01	35.40	2.00	15.94	27.67
LL2	2.00	1.98	17.77	35.36	1.85	15.85	27.53
LL3	2.00	2.00	18.18	35.46	2.00	16.09	27.71
NW1	2.00	0.70	19.01	35.19	0.72	16.14	27.07
NW2	2.00	0.68	18.96	37.16	0.70	16.24	27.75
NW3	2.00	0.78	19.04	34.48	0.80	16.19	26.74
CP	3.00	**0.09**	15.95	21.80	**0.09**	16.13	24.51
LL1	3.00	1.96	18.62	35.73	1.92	16.30	26.50
LL2	3.00	1.79	18.33	35.24	1.59	16.02	26.80
LL3	3.00	2.00	18.76	35.87	2.00	16.48	26.58
NW1	3.00	0.73	19.92	37.01	0.73	16.87	27.81
NW2	3.00	0.65	19.24	38.11	0.65	16.70	28.31
NW3	3.00	0.83	20.01	36.74	0.85	16.92	27.79

TABLE D.5: Monthly electricity supplied in Spain time series results

Bibliography

Almeida, Alejandro, Antonio A Golpe, and JM Martín Álvarez (2020). "A spatial analysis of the Spanish tobacco consumption distribution: Are there any consumption clusters?" In: *Public Health* 186, pp. 28–30.

Almeida, Alejandro et al. (2021). "The price elasticity of cigarettes: new evidence from Spanish regions, 2002–2016". In: *Nicotine and Tobacco Research* 23.1, pp. 48–56.

Álvarez, JM Martín et al. (2020). "Price and income elasticities of demand for cigarette consumption: what is the association of price and economic activity with cigarette consumption in Spain from 1957 to 2016?" In: *Public Health* 185, pp. 275–282.

Ang, Andrew, Monika Piazzesi, and Min Wei (2006). "What does the yield curve tell us about GDP growth?" In: *Journal of econometrics* 131.1-2, pp. 359–403.

Armstrong, J. S. (1985). *Long-range Forecasting: From Crystal Ball to Computer*. Vol. 2. Wiley. ISBN: 9780471822608.

Berge, Travis J (2015). "Predicting recessions with leading indicators: Model averaging and selection over the business cycle". In: *Journal of Forecasting* 34.6, pp. 455–471.

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo (2018). "A note on the validity of cross-validation for evaluating autoregressive time series prediction". In: *Computational Statistics & Data Analysis* 120, pp. 70 –83.

Bernanke, Ben, A Blinder, et al. (1992). "l992. the federal funds rate and the channels of monetary transmission". In: *American Economic Review* 82.4, pp. 90–2.

Blecher, EH and CP Van Walbeek (2004). "An international analysis of cigarette affordability". In: *Tobacco Control* 13.4, pp. 339–346.

Bluwstein, Kristina et al. (2020). "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach". In.

Box, G.E.P. and G.M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Bravo, J. M. et al. (June 2017). "A General Framework for Predictors Based on Bounding Techniques and Local Approximation". In: *IEEE Transactions on Automatic Control* 62.7, pp. 3430–3435. ISSN: 0018-9286. DOI: 10.1109/TAC.2016.2612538.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Cadahia, Pedro et al. (2021). "Measuring anomalies in cigarette sales using official data from Spanish provinces: Are the anomalies detected by the Empty Pack Surveys (EPSs) used by Transnational Tobacco Companies (TTCs) the only anomalies?" In: *Tobacco Induced Diseases* 19.December, pp. 1–12. DOI: 10.18332/tid/143321. URL: https://doi.org/10.18332/tid/143321.

Cadahía, Pedro and Jose Manuel Bravo Caro (2021). *Short-term prediction of Time Series based on bounding techniques*. arXiv: 2101.10719 [stat.ML].

Cameron, David R (1978). "The expansion of the public economy: A comparative analysis". In: *American political science review* 72.4, pp. 1243–1261.

Campbell, John Y (1995). "Some lessons from the yield curve". In: *Journal of economic perspectives* 9.3, pp. 129–152.

Campbell, M. J.and A. M. Walker (1977). "A Survey of statistical work on the Mackenzie River series of annual Canadian lynx trappings for the years 1821?1934 and a new analysis". In: *Journal of the Royal Statistical Society series* A, 140, pp. 411–431.

Chaloupka, Frank J, Kurt Straif, and Maria E Leon (2011). "Effectiveness of tax and price policies in tobacco control". In: *Tobacco control* 20.3, pp. 235–238.

Chaloupka, Frank J, Ayda Yurekli, and Geoffrey T Fong (2012). "Tobacco taxes as a tobacco control strategy". In: *Tobacco control* 21.2, pp. 172–180.

Chan, Tiffany et al. (2020). "Identifying Counterfeit Cigarettes Using Environmental Pollen Analysis: An Improved Procedure". In: *Journal of Forensic Sciences* 65.6, pp. 2138–2145.

Chang, K.S. and H. Tong (1986). "On estimating thresholds in autoregressive models". In: *Journal of Time Series Analysis* 7, pp. 179–190.

Chatfield, Chris and Alexander Collins (1981). *Introduction to multivariate analysis*. Vol. 1. CRC Press.

Chen, Jing et al. (2015). "Did the tobacco industry inflate estimates of illicit cigarette consumption in Asia? An empirical analysis". In: *Tobacco control* 24.e2, e161–e167.

Chen, Nai-Fu (1991). "Financial investment opportunities and the macroeconomy". In: *The Journal of Finance* 46.2, pp. 529–554.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chinn, Menzie and Kavan Kucko (2015). "The predictive power of the yield curve across countries and time". In: *International Finance* 18.2, pp. 129–156.

Davidian, Marie and Thomas A Louis (2012). *Why statistics?*

Deng, Houtao (2019). "Interpreting tree ensembles with intrees". In: *International Journal of Data Science and Analytics* 7.4, pp. 277–287.

Dodge, Yadolah and Daniel Commenges (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Dohoo, Ian R et al. (1997). "An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies". In: *Preventive veterinary medicine* 29.3, pp. 221–239.

Döpke, Jörg, Ulrich Fritsche, and Christian Pierdzioch (2017). "Predicting recessions with boosted regression trees". In: *International Journal of Forecasting* 33.4, pp. 745–759.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.

Dotsey, Michael (1998). "The predictive content of the interest rate term spread for future economic growth". In: *FRB Richmond Economic Quarterly* 84.3, pp. 31–51.

Dueker, Michael J et al. (1997). "Strengthening the Case for the Yield Curve as a Predictor of US Recessions". In: *Federal Reserve Bank of St. Louis Review* Mar, pp. 41–51.

Engle, Robert F and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". In: *Journal of business & economic statistics* 22.4, pp. 367–381.

Estrella, Arturo (2005a). "The yield curve as a leading indicator: frequently asked questions". In: *New York Fed*.

— (2005b). "Why does the yield curve predict output and inflation?" In: *The Economic Journal* 115.505, pp. 722–744.

Estrella, Arturo and Gikas Hardouvelis (1990). "Possible roles of the yield curve in monetary policy". In: *Federal Reserve Bank of New York (Ed.), Intermediate targets and indicators for monetary policy*, pp. 339–362.

Estrella, Arturo and Gikas A Hardouvelis (1991). "The term structure as a predictor of real economic activity". In: *The journal of Finance* 46.2, pp. 555–576.

Estrella, Arturo and Frederic S Mishkin (1996). "The yield curve as a predictor of US recessions". In: *Current issues in economics and finance* 2.7.

— (1998). "Predicting US recessions: Financial variables as leading indicators". In: *Review of Economics and Statistics* 80.1, pp. 45–61.

Estrella, Arturo, Anthony P Rodrigues, and Sebastian Schich (2003). "How stable is the predictive power of the yield curve? Evidence from Germany and the United States". In: *Review of Economics and Statistics* 85.3, pp. 629–644.

Estrella, Arturo and Mary Trubin (2006). "The yield curve as a leading indicator: Some practical issues". In: *Current issues in Economics and Finance* 12.5.

Evangelou, Marina and Niall M Adams (2020). "An anomaly detection framework for cyber-security data". In: *Computers & Security* 97, p. 101941.

Evgenidis, Anastasios, Stephanos Papadamou, and Costas Siriopoulos (2020). "The yield spread's ability to forecast economic activity: What have we learned after 30 years of studies?" In: *Journal of Business Research* 106, pp. 221–232.

Evgenidis, Anastasios, Athanasios Tsagkanos, and Costas Siriopoulos (2017). "Towards an asymmetric long run equilibrium between stock market uncertainty and the yield spread. A threshold vector error correction approach". In: *Research in International Business and Finance* 39, pp. 267–279.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric Methods and Parametric Methods*. Springer Series in Statistics. New York: Springer.

Fan, Jianqing and Irène Gijbels (1996). *Local polynomial modelling and its applications*. Monographs on statistics and applied probability series 66. London [u.a.]: Chapman and Hall. ISBN: 0412983214.

FCTC, WHO (Mar. 2021). *Guidelines for Implementation of Article 6 of the WHO FCTC: Price and Tax Measures to Reduce the Demand for Tobacco*. World Health Organization Framework Convention on Tobacco Control. http:// www.who.int/fctc/guidelines/adopted/Guid

Fernández, E et al. (2004). "Price and consumption of tobacco in Spain over the period 1965–2000". In: *European Journal of Cancer Prevention* 13.3, pp. 207–211.

Ferreira, Artur J and Mário AT Figueiredo (2012). "Boosting algorithms: A review of methods, theory, and applications". In: *Ensemble machine learning*, pp. 35–85.

Freund, Yoav, Robert Schapire, and Naoki Abe (1999). "A short introduction to boosting". In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780, p. 1612.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Gallagher, Allen WA et al. (2019). "Tobacco industry data on illicit tobacco trade: a systematic review of existing assessments". In: *Tobacco control* 28.3, pp. 334–345.

Gallego, Juan M et al. (2020). "Tobacco taxes and illicit cigarette trade in Colombia". In: *Economics & Human Biology* 39, p. 100902.

Gao, J. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. Chapman and Hall/CRC.

Gareth, James et al. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gebka, Bartosz and Mark E Wohar (2018). "The predictive power of the yield spread for future economic expansions: Evidence from a new approach". In: *Economic Modelling* 75, pp. 181–195.

Gilmore, Anna B et al. (2014). "Towards a greater understanding of the illicit tobacco trade in Europe: a review of the PMI funded 'Project Star'report". In: *Tobacco control* 23.e1, e51–e61.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378.

Gogas, Periklis, Theophilos Papadimitriou, and Efthymia Chrysanthidou (2015). "Yield curve point triplets in recession forecasting". In: *International Finance* 18.2, pp. 207–226.

Gogas, Periklis et al. (2015). "Yield curve and recession forecasting in a machine learning framework". In: *Computational Economics* 45.4, pp. 635–645.

Gomajee, Ramchandar et al. (2021). "Decrease in cross-border tobacco purchases despite intensification of antitobacco policies in France". In: *Tobacco Control* 30.4, pp. 428–433.

Gooijer, Jan G.De and Ali Gannoun (2000). "Nonparametric conditional predictive regions for time series". In: *Computational Statistics & Data Analysis* 33.3, pp. 259 –275.

Granter, Scott R, Andrew H Beck, and David J Papke Jr (2017). "AlphaGo, deep learning, and the future of the human microscopist". In: *Archives of pathology & laboratory medicine* 141.5, pp. 619–621.

Haggan, V. and T. Ozaki (1981). "Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model". In: *Biometrika* 68, pp. 186–196.

Haighton, Catherine, Catherine Taylor, and Ailsa Rutter (2017). "Standardized packaging and illicit tobacco use: A systematic review". In: *Tobacco Prevention & Cessation* 3.

Hamilton, James D and Dong H Kim (2002). "A re-examination of the predictability of the yield spread for real economic activity". In: *Journal of Money, Credit, and Banking* 34.2, pp. 340–360.

Hamilton, James Douglas (1994). *Time series analysis*. Princeton, NJ: Princeton Univ. Press.

Härdle, W., H. Lütkepohl, and R. Chen (1997). "A Review of Nonparametric Time Series Analysis". In: *International Statistical Review* 65.1, pp. 49–73.

Härdle, Wolfgang (1990). *Applied nonparametric regression*. Econometric Society monographs 19. Cambridge u. a.: Cambridge University Pr.

Harvey, Campbell R (1989). "Forecasts of economic growth from the bond and stock markets". In: *Financial Analysts Journal* 45.5, pp. 38–45.

— (1993). "Term structure forecasts economic growth". In: *Financial Analysts Journal* 49.3, pp. 6–8.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Heider, Fritz and Marianne Simmel (1944). "An experimental study of apparent behavior". In: *The American journal of psychology* 57.2, pp. 243–259.

Ho, Tin Kam (1995). "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Holmes, Mark J, Jesús Otero, and Theodore Panagiotidis (2015). "The expectations hypothesis and decoupling of short-and long-term US interest rates: A pairwise approach". In: *The North American Journal of Economics and Finance* 34, pp. 301–313.

Hyndman, Rob J and Anne B Koehler (2006a). "Another look at measures of forecast accuracy". In: *International journal of forecasting* 22.4, pp. 679–688.

Hyndman, Rob J. and Anne B. Koehler (2006b). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4, pp. 679 –688.

Hyndman, Rob J. et al. (2005). "Local Linear Forecasts Using Cubic Smoothing Splines". In: *Australian & New Zealand Journal of Statistics* 47.1, pp. 87–99.

Jakob, Julian, Jacques Cornuz, and Pascal Diethelm (2017). "Prevalence of tobacco smoking in Switzerland: do reported numbers underestimate reality?" In: *Swiss medical weekly* 147.

JM, Martín Álvarez et al. (2021). "The influence of cigarette price on the cigarette consumption in Spain: a Logarithmic Mean Divisia Index analysis from 1957 to 2018." In: *Revista Espanola de Salud Publica* 95.

Kim, Been, Rajiv Khanna, and Oluwasanmi O Koyejo (2016). "Examples are not enough, learn to criticize! criticism for interpretability". In: *Advances in neural information processing systems* 29.

Kleinberg, EM (1990). "Stochastic discrimination". In: *Annals of Mathematics and Artificial intelligence* 1.1-4, pp. 207–239.

Kleinberg, Eugene M (2000). "On the algorithmic implementation of stochastic discrimination". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.5, pp. 473–490.

Koenker, Roger and Gilbert Bassett Jr (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Komunjer, Ivana (2013). "Quantile prediction". In: *Handbook of economic forecasting*. Vol. 2. Elsevier, pp. 961–994.

Kurti, Marin et al. (2019). "Presence of counterfeit Marlboro gold packs in licensed retail stores in New York City: evidence from test Purchases". In: *Nicotine and Tobacco Research* 21.8, pp. 1131–1134.

Lahiri, Kajal, George Monokroussos, and Yongchen Zhao (2013). "The yield spread puzzle and the information content of SPF forecasts". In: *Economics Letters* 118.1, pp. 219–221.

Laurent, Robert D (1988). "An interest rate-based indicator of monetary policy". In: *Economic Perspectives* 12.Jan, pp. 3–14.

Leite Ribeiro, Livio Santos de and Vilma da Conceição Pinto (2020). "Discrepancies in the Brazilian tobacco production chain: raw inputs, international trade and legal cigarette production". In: *Tobacco Control* 29.Suppl 5, s310–s318.

Lin, Feng-Jenq (2008). "Solving multicollinearity in the process of fitting regression model using the nested estimate procedure". In: *Quality & Quantity* 42.3, pp. 417–426.

Lipovetsky, Stan and Michael Conklin (2001). "Analysis of regression in game theory approach". In: *Applied Stochastic Models in Business and Industry* 17.4, pp. 319–330.

Liu, Weiling and Emanuel Moench (2016). "What predicts US recessions?" In: *International Journal of Forecasting* 32.4, pp. 1138–1150.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

Maldonado, Norman et al. (2020). "Measuring illicit cigarette trade in Colombia". In: *Tobacco control* 29.Suppl 4, s260–s266.

Mangalova, Ekaterina and Olesya Shesterneva (2016). "Sequence of nonparametric models for GEFCom 2014 - probabilistic electric load forecasting". In: *International Journal of Forecasting* 32.3, pp. 1023 –1028.

Meinshausen, Nicolai and Greg Ridgeway (2006). "Quantile regression forests." In: *Journal of Machine Learning Research* 7.6.

Miera Juarez, Belen Saenz de et al. (2021). "Measuring the illicit cigarette market in Mexico: a cross validation of two methodologies". In: *Tobacco control* 30.2, pp. 125–131.

Milanese, M. et al. (1996). *Bounding Approaches to System Identification*. Plenum Press, New York.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267, pp. 1–38.

Mottaqi, Mohammad Sadeq, Fatemeh Mohammadipanah, and Hedieh Sajedi (2021). "Contribution of machine learning approaches in response to SARS-CoV-2 infection". In: *Informatics in Medicine Unlocked*, p. 100526.

Nadaraya, E. A. (1964). "On Estimating Regression". In: *Theory of Probability & Its Applications* 9.1, pp. 141–142.

Nagy, Gábor I et al. (2016). "GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach". In: *International Journal of Forecasting* 32.3, pp. 1087–1093.

Nargis, Nigar et al. (2021). "Price, income, and affordability as the determinants of tobacco consumption: A practitioner's guide to tobacco taxation". In: *Nicotine and Tobacco Research* 23.1, pp. 40–47.

Newton (1988). *Monthly critical radio frequencies in Washington*. data retrieved from datamarket. URL: http://datamarket.com/data/list/?q=provider:tsdl.

Ng, Serena (2017). *Opportunities and challenges: Lessons from analyzing terabytes of scanner data*. Tech. rep. National Bureau of Economic Research.

Nyberg, Henri (2010). "Dynamic probit models and financial variables in recession forecasting". In: *Journal of Forecasting* 29.1-2, pp. 215–230.

Otte, Clemens (2013). "Safe and interpretable machine learning: a methodological review". In: *Computational intelligence in intelligent data analysis*, pp. 111–122.

Paraje, Guillermo (2019). "Illicit cigarette trade in five South American countries: a gap analysis for Argentina, Brazil, Chile, Colombia, and Peru". In: *Nicotine and Tobacco Research* 21.8, pp. 1079–1086.

Pinilla, Jaime (2002). "Análisis comparado del impacto de las políticas impositivas vía precio en el consumo de tabaco". In: *Gaceta sanitaria* 16.5, pp. 425–435.

Poole, William, Robert H Rasche, Daniel L Thornton, et al. (2002). "Market anticipations of monetary policy actions". In: *Review-Federal Reserve Bank of Saint Louis* 84.4, pp. 65–94.

Rao, T. S. and M.M. Gabr (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Vol. 24. New York: Springer-Verlag. ISBN: 9781468463187.

Rimol, Meghan and Katie Costello (Mar. 10, 2021). "Gartner Says Tech Investors Will Prioritize Data Science and Artificial Intelligence Above "Gut Feel" for Investment Decisions By 2025". In: *Gartner*. URL: https://www.gartner.com/en/newsroom/press-releases/2021-03-10-gartner-says-tech-investors-will-prioritize-data-science-and-artificial-intelligence-above-gut-feel-for-investment-decisions-by-20250.

Roll, J., A. Nazin, and L. Ljung (2005). "Nonlinear system identification via direct weight optimization". In: *Automatica* 41.3, pp. 475–490.

Rowell, Andrew, Karen Evans-Reeves, and Anna B Gilmore (2014). "Tobacco industry manipulation of data on and press coverage of the illicit tobacco trade in the UK". In: *Tobacco control* 23.e1, e35–e43.

Rudebusch, Glenn D and John C Williams (2009). "Forecasting recessions: the puzzle of the enduring power of the yield curve". In: *Journal of Business & Economic Statistics* 27.4, pp. 492–503.

Russell, MA (1973). "Changes in cigarette price and consumption by men in Britain, 1946-71: a preliminary analysis." In: *British journal of preventive & social medicine* 27.1, p. 1.

Saltelli, Andrea et al. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Schafferer, C et al. (2018). "A simulation impact evaluation of a cigarette excise tax increase on licit and illicit cigarette consumption and tax revenue in 36 European countries". In: *Public health* 162, pp. 48–57.

Shang, Ce, Estelle P Dauchy, and Naomi Feldman (2020). "The price elasticity of demand for heated tobacco products". In: *Tobacco Prevention & Cessation* 6.Supplement.

Shiller, Robert J and J Huston McCulloch (1990). "The term structure of interest rates". In: *Handbook of monetary economics* 1, pp. 627–722.

Stock, James H and Mark W Watson (1989). "New indexes of coincident and leading economic indicators". In: *NBER macroeconomics annual* 4, pp. 351–394.

Stoklosa, Michal (2016). "Is the illicit cigarette market really growing? The tobacco industry's misleading math trick". In: *Tobacco control* 25.3, pp. 360–361.

Strobl, Carolin et al. (2007a). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, pp. 1–21.

— (2007b). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, pp. 1–21.

Strobl, Carolin et al. (2008). "Conditional variable importance for random forests". In: *BMC bioinformatics* 9.1, pp. 1–11.

Sun, Quan and Bernhard Pfahringer (2011a). "Bagging ensemble selection". In: *Australasian Joint Conference on Artificial Intelligence*. Springer, pp. 251–260.

— (2011b). "Bagging ensemble selection". In: *Australasian Joint Conference on Artificial Intelligence*. Springer, pp. 251–260.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*. Vol. 21. Lecture Notes in Statistics. Heidelberg: Springer.

Truong, Y.K. (1993). *A nonparametric framework for time series analysis*. New Directions in Time Series Analysis. New York: Springer.

Tufféry, Stéphane (2011). *Data mining and statistics for decision making*. John Wiley & Sons.

Vetzal, Kenneth R (1994). "A survey of stochastic continuous time models of the term structure of interest rates". In: *Insurance: Mathematics and Economics* 14.2, pp. 139–161.

Vides, José Carlos, Jesús Iglesias, and Antonio A Golpe (2018). "The term structure under non-linearity assumptions: New methods in time series". In: *New methods in fixed income modeling*. Springer, pp. 117–136.

Vladislavleva, Ekaterina et al. (2013a). "Predicting the energy output of wind farms based on weather data: Important variables and their correlation". In: *Renewable energy* 50, pp. 236–243.

— (2013b). "Predicting the energy output of wind farms based on weather data: Important variables and their correlation". In: *Renewable energy* 50, pp. 236–243.

Vrontos, Spyridon D, John Galakis, and Ioannis D Vrontos (2021). "Modeling and predicting US recessions using machine learning techniques". In: *International Journal of Forecasting* 37.2, pp. 647–671.

Watson, Geoffrey S. (1964). "Smooth regression analysis". In: *Sankhyā Ser.* 26, pp. 359–372.

Weber, Enzo and Jürgen Wolters (2012). "The US term structure and central bank policy". In: *Applied Economics Letters* 19.1, pp. 41–45.

— (2013). "Risk and policy shocks on the us term structure". In: *Scottish Journal of Political Economy* 60.1, pp. 101–119.

Wei, Pengfei, Zhenzhou Lu, and Jingwen Song (2015). "Variable importance analysis: a comprehensive review". In: *Reliability Engineering & System Safety* 142, pp. 399–432.

Wei, Ying et al. (2006). "Quantile regression methods for reference growth charts". In: *Statistics in medicine* 25.8, pp. 1369–1382.

Weyn, Jonathan A, Dale R Durran, and Rich Caruana (2019). "Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data". In: *Journal of Advances in Modeling Earth Systems* 11.8, pp. 2680–2693.

Wheelock, David C, Mark E Wohar, et al. (2009). "Can the term spread predict output growth and recessions? A survey of the literature". In: *Federal Reserve Bank of St. Louis Review* 91.5 Part 1, pp. 419–440.

Wright, Marvin N and Andreas Ziegler (2015). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *arXiv preprint arXiv:1508.04409*.

Yang, Song et al. (2015). "Variable importance analysis for urban building energy assessment in the presence of correlated factors". In: *Procedia Engineering* 121, pp. 277–284.

Yin, Yi and Pengjian Shang (2016). "Forecasting traffic time series with multivariate predicting method". In: *Applied Mathematics and Computation* 291, pp. 266 –278. ISSN: 0096-3003.

Yun, Yong-Huan et al. (2016). "Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery". In: *Analytica chimica acta* 911, pp. 27–34.

Zhang, G.Peter (Jan. 2003). "Time series forecasting using a hybrid ARIMA and neural network model". In: *Neurocomputing* 50, pp. 159–175.

Zhao, Zhibiao (2008). "Parametric and nonparametric models and methods in financial econometrics". In: *Statistics Surveys* 2, pp. 1–42.