FEDERAL UNIVERSITY OF RIO GRANDE DO SUL
FACULTY OF ECONOMIC SCIENCES
DEPARTMENT OF ECONOMICS AND INTERNATIONAL RELATIONS

PEDRO PABLO SKORIN URANGA

FORECASTING WITH HIGH-DIMENSIONAL DATA USING LINEAR AND P-SPLINES L2-BOOSTING: AN EXERCISE FOR THE UNEMPLOYMENT RATE IN BRAZIL

Porto Alegre
2022

PEDRO PABLO SKORIN URANGA

FORECASTING WITH HIGH-DIMENSIONAL DATA USING LINEAR AND P-SPLINES L2-BOOSTING: AN EXERCISE FOR THE UNEMPLOYMENT RATE IN BRAZIL

Work submitted to Faculty of Economics at UFRGS as partial requirement for obtaining the Bachelor's Degree in Economics.

Supervisor: Prof. Dr. Hudson da Silva Torrent

Porto Alegre

2022

CIP - Catalogação na Publicação

Skorin, Pedro Pablo
Forecasting with high-dimensional data using linear and p-splines L2-boosting: an exercise for the unemployment rate in Brazil / Pedro Pablo Skorin. -- 2022.
57 f.
Orientador: Hudson da Silva Torrent.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Curso de Ciências Econômicas, Porto Alegre, BR-RS, 2022.

1. Previsão do Desemprego. 2. Previsão em Alta-dimensionalidade. 3. Previsão Não-linear. 4. Boosting. 5. P-splines. I. Torrent, Hudson da Silva, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

PEDRO PABLO SKORIN URANGA

FORECASTING WITH HIGH-DIMENSIONAL DATA USING LINEAR AND P-SPLINES L2-BOOSTING: AN EXERCISE FOR THE UNEMPLOYMENT RATE IN BRAZIL

Work submitted to Faculty of Economics at UFRGS as partial requirement for obtaining the Bachelor's Degree in Economics.

Approved in: Porto Alegre, May 12, 2022.

EXAMINATION BOARD:

Prof. Dr. Hudson da Silva Torrent – Advisor
UFRGS

Prof. Dr. Cleiton Guollo Taufemback
UFRGS

Prof. Dr. Fernando Augusto Boeira Sabino da Silva
UFRGS

ACKNOWLEDGEMENTS

Altough this project is in English, não acho que faria sentido escrever este trecho em uma língua diferente do português. Os primeiros e principais agradecimentos vão para minha mãe, meu pai e meu irmão, pessoas que sempre me ajudaram na busca pelo conhecimento. Até por motivos fisiológicos, sem eles nada seria possível. Também deixo um agradecimento especial ao meu cachorro, membro da família que me acompanhou da infância até eu me tornar adulto e que infelizmente nos deixou pouco tempo antes do término deste projeto.

Após o agradecimento familiar, passamos para a parte mais difícil da monografia: ter que selecionar apenas alguns entre aqueles que me ajudaram nessa conquista. Primeiro agradeço o constante suporte e amor recebido por Vanessa, namorada que conheci na faculdade e que me acompanhou durante a graduação. Dentre os diversos amigos que andaram ao meu lado, agradeço ao Guilherme e ao Angelo pelo companheirismo ao longo do curso e ao Diego e ao Tuti, amigos da zona sul de Porto Alegre, com quem compartilhamos diversos trajetos até a faculdade. Não podem ficar de fora os agradecimentos a todo o grupo que entrou comigo em 2017 - *a nata* - com quem juntos enfrentamos a graduação; e também aos integrantes da empresa júnior Equilíbrio, que me incentivavam a ficar das 7 horas da manhã até as 9 horas da noite na FCE. Fecho este trecho agradecendo à Rílari, ao Vitor e aos outros companheiros do grupo de estudos de macroeconomia realizado durante o fatídico ano de 2020.

Em termos laborais, agradeço aqueles que confiaram no meu trabalho e me deram oportunidades de experiências profissionais. Tenho gratidão do meu período como estagiário na Melnick e como analista de risco no Sicredi. Com relação aos doscentes, deixo meu agradecimento ao professor Marcelo Portugal, pelo tempo de monitoria, ao professor Jorge Araújo, pelas horas de conversa após as aulas de economia matemática, e ao ilustre professor Hudson Torrent, meu orientador e incentivador, pela iniciação científica e por ter estado sempre presente para me ajudar. Como mensão honrosa, agradeço aos professores Carlos Horn, Miguel Caceiro, Carlos Hoppen e João Plínio pelas suas excepcionais capacidades na arte de educar.

Finalmente, agradeço a você leitor. Obrigado por se interessar pelo projeto e por ler meus agradecimentos até aqui!

Prediction is very difficult, especially about the future.

— Niels Bohr

RESUMO

Este trabalho tem por objetivo principal verificar a validade do algoritmo component-wise boosting para prever a taxa de desemprego mensal na Região Metropolitana de São Paulo. Para isto, desenvolvemos três modelos, um modelo completamente linear, apenas com learners lineares, um modelo completamente não-linear, apenas com learners p-splines, e um modelo misto. O modelo misto utiliza learners p-spline para os preditores relacionados a preços e learners lineares para os demais. Tal configuração é motivada por estudos que apontam uma relação não-linear da curva de Phillips; assim, o modelo misto aplica a modelagem não-linear com consciência das relações entre a variável dependente e seus preditores. Os modelos boosting utilizam 159 preditores e são testado entre 2013 e 2019, os anos em que ocorreu uma das maiores recessões da história brasileira. Para testar a validade do modelo, utilizamos o modelo SARIMA como benchmark. Os resultados indicaram uma superioridade em comparação com o benchmark dos três modelos para os horizontes de previsão $h = 2$ até $h = 11$. Além disso, o modelo misto se destaca com os melhores resultados em métricas de desempenho, seguido pelo modelo linear e depois pelo modelo completamente não-linear. O trabalho conclui que a modelagem mista tendeu a ser superior, especialmente para horizontes de previsão intermediários ($h = 5$ até $h = 11$), e também destaca a importância de usar a modelagem não linear com consciência das relações entre os preditores e a variável dependente.

Palavras-chave: Previsão do Desemprego, Previsão em Alta-dimensionalidade, Previsão Não-linear, Boosting, P-splines.

ABSTRACT

The main purpose of this paper is to verify the validity of the component-wise boosting algorithm to predict the monthly unemployment rate in the São Paulo Metropolitan Region. For this, we developed three models, a completely linear model, with only linear learners, a completely non-linear model, with only p-splines learners, and a mixed model. The mixed model uses p-spline learners for price-related predictors and linear learners for the rest of them. Such a configuration is motivated by studies that point out a non-linear relationship of the Phillips curve; thus, the mixed model applies non-linear modeling with awareness of the relationships between the dependent variable and its predictors. The boosting models use 159 predictors and are tested between 2013 and 2019, the years when one of the biggest recessions in Brazilian history occurred. To test the validity of the model, we used the SARIMA model as a benchmark. The results indicated a superiority compared to the benchmark of the three models for the forecast horizons $h = 2$ through $h = 11$. In addition, the mixed model stands out with the best results in performance metrics, followed by the linear model and then by the completely non-linear model. The work concludes that mixed modeling tended to be superior, especially for intermediate forecast horizons ($h = 5$ trough $h = 11$), and also highlights the importance of using nonlinear modeling with awareness of the relationships between predictors and the dependent variable.

Keywords: Unemployment Forecast. High-Dimensional Forecast. Non-linear Forecast. Boosting, P-splines

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The use of machine learning techniques for macroeconomic time series forecasting is a growing trend in economic research. This project aims to expand the trend for the Brazilian macroeconomic environment by testing a particular machine learning technique, the component-wise boosting algorithm, for a particular time series, the unemployment rate of the Metropolitan Region of São Paulo (MRSP). 159 different predictors of the macroeconomic theme are used in a monthly forecasting exercise tested from 2013 through 2019[1].

Although considerable research has been done in macroeconomics using non-linear relationships between the variables, when applied to forecasting, linear models historically tended to have better performance compared to non-linear approaches (KAUPPI; VIRTANEN, 2021). Nevertheless, this thesis may be changing with the expansion of machine-learning algorithms in economic research. As given by Medeiros et al. (2021), new approaches using artificial intelligence consistently produce better performance results compared to traditional linear models. In this project, we test this statement with a boosting algorithm for the unemployment rate in Brazil.

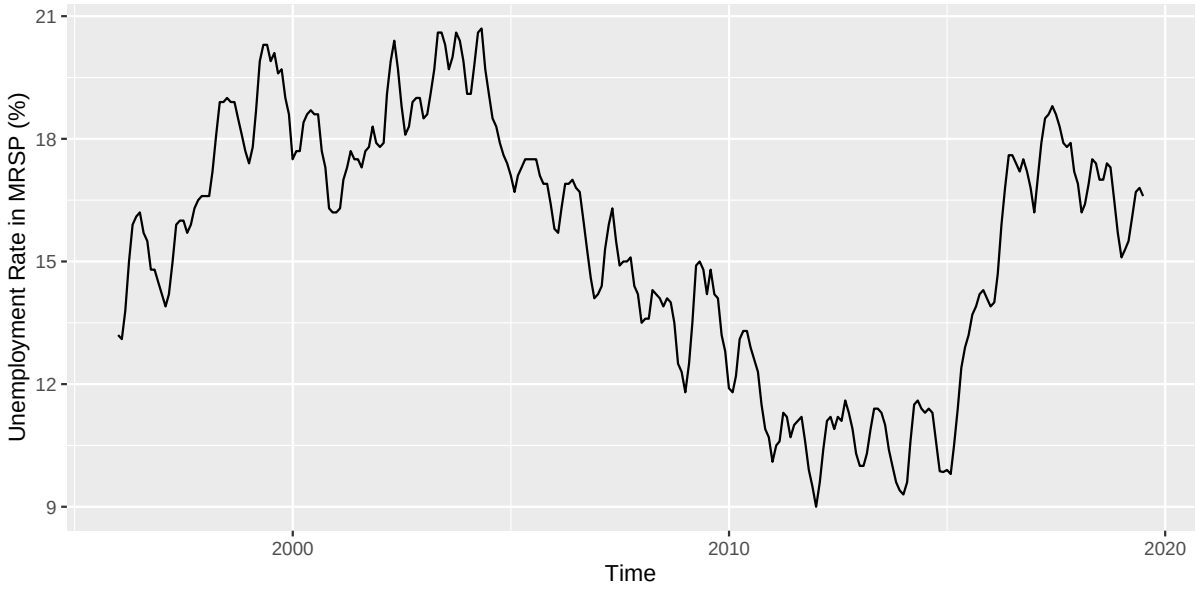
Between 2014 and 2016, Brazil faced one of the largest and deepest recession of its history. This has not only impacted the current conditions of the country's population but has also pushed society to be more concerned about the future of the country, increasing attention to information on economic status, such as forecasts of GDP, inflation, and interest rates. Between these macroeconomic numbers, unemployment rates are one of time series that brings most interest for all classes of society, since it can give insights about employment opportunities for workers or about product demand for companies. The precision of our knowledge of future unemployment creates a more stable path for the decision-making of individuals and companies.

Figure 1 shows the complete dependent time series. As can be seen, the data has non-ordinary movements, particularly following the beginning of Brazilian crisis in 2014. Modeling macroeconomic can became especially difficult when done in an environment of recession, thus our forecasting exercise represents a real challenge for any prediction strategy.

The methodology to test the validity of boosting algorithm and its different configurations will be based on a comparison of performance measures in a dedicated test segment of the time series. This test segment differs from another part of the series, the train segment, a time interval in which the models will fit their parameters with knowledge of the dependent variable

[1] The code repository for this project can be found at https://github.com/pedroskorin/boosting-thesis-ufrgs.

Figure 1 – Unemployment rate in the Metropolitan Region of São Paulo (MRSP)



Source: Elaborated by the author (2022)

true value. The benchmark selected to judge the general competitiveness of boosting is the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. This decision was based on the consolidation of the model in the unemployment rate forecasting literature, as can be seen in Chapter 2. Boosting will be considered a good instrument if it brings better performance measures than SARIMA. Also, the better configurations of boosting will be those that bring better performance measures in comparison to the other configurations.

In short, the boosting algorithm uses a set of learners based on predictors to create the desired forecast. The learners used in this project are divided into two types: linear learners, estimated by ordinary least-squared, and non-linear learners, modeled via p-splines. The different results given by both linear and non-linear learners are a crucial aspect of the project, since they insert the research into the larger discussion of non-linearity modeling in macroeconomic forecasts.

Three models are proposed and tested: a complete linear model, where all predictors enter in the algorithm with linear learners; a complete non-linear model, where all the predictors appear in the algorithm with p-splines learners; and finally, a mixed model, where only the predictors related to prices are modeled with non-linear learners, with the rest using linear ones. This last proposal is based on the discussion about the format of the Phillips curve, the trade-off relation between unemployment and inflation. Research indicates the non-linearity of this relation (DEBELLE; LAXTON, 1997) (XU et al., 2015) (BYRNE; ZEKAITE, 2020); therefore, the

incorporation of the third model asks if the conditional expectation of the unemployment rate in the MRSP could be better modeled respecting the most recent research on employment and price relation. Finally, we emphasize that the results presented in this project have a restricted nature. Here we use only one time series for forecasting, so the results of the project should be associated with possible specific characteristics of the time series such as the region or also the time interval of the exercise.

The rest of the project is organized as follows: First, in Chapter 2, we will present the literature review for both the boosting algorithm and unemployment forecasting. Secondly, in Chapter 3, we will explain how the boosting algorithm actually works, from the linear learners to the non-linear p-spline learners. In Chapter 4, it is time to explain the forecasting process, the benchmark model and the selected performance measures and in Chapter 5 we understand more deeply the forecasting exercise by exploring the created data set and especially the dependent variable. Chapter 6 brings the results and, finally, Chapter 7 concludes the project.

2 LITERATURE REVIEW

The review of the literature will be divided into two sections. First, we explore how boosting was developed, its history and its evolution to the model we use in this exercise. Macroeconomic applications and discussion of the linearity of the model are also reviewed. The second section concentrates on unemployment forecasting, independently of whether it is done using boosting or not. Reviewing other unemployment forecast models is important because we need to ensure that common practices are applied in this project based on cutting-edge research practices.

2.1 THE L2-BOOSTING ALGORITHM

The L2 boosting model is an iterative algorithm of the ensemble type, with a central objective of bias and variance reduction (ZHOU, 2019). We say that the boosting model is an iterative algorithm because it uses mathematical procedures that start with an initial value to generate a sequence of improved approximate solutions for the problem class, in which the i-th iteration is derived from past iterations. Furthermore, we say that the model is of ensemble type because it uses multiple learning algorithms to obtain better prediction performances.

The origin of boosting is based on the question proposed by Kearns and Valiant (1989): Can a set of weak learners create a single strong learner? In this question, weak learners mean methods with low predictive ability, that is, with accuracy only slightly better than random guessing. Strong learners, on the other hand, are methods with high accuracy. The first affirmative answer to the question comes with the 1990 article by Schapire (1990).

Schapire's paper describes a method of converting weak learning algorithms into one that achieves arbitrarily high accuracy. The article is built around a conjecture, dubbed the equivalence between strong and weak learning. This conjecture states that a conceptual class C has weak learning if and only if it has strong learning. It is clear that strong learning implies weak learning, but the non-trivial result consists in the return of the theorem: it is possible to construct a strong learning method from a weak method. Later on, the operation ends up being known as the *boosting* method, precisely because it improves weak methods, as "to boost" means "to improve".

One main obstacle in Schapire's paper is that an application was presented exclusively on

binary data. Later, still in the 1990s, we would observe two important steps in the development of the boosting algorithm. In Freund (1995), the author goes one step further than Schapire in substantiating the basic aspects of the method. First, by reducing the number of assumptions needed to apply the algorithm, and second, by applying the boosting method to non-binary data. The second important step is the work of (BREIMAN, 1996). Breinam manages to interpret the boosting algorithm as a gradient algorithm in a functional space, inspired by numerical optimizations and statistical estimation. In practice, this implies that boosting could work for methods beyond classification. The algorithm increases its generalization, and consequently creates more potential for applications.

Finally, a few years later, Friedman, Hastie, and Tibshirani (2000) and Friedman (2001) develop the boosting method in a more generic form, for regressions that are implemented as optimizers. They present a set of loss functions and, when using the quadratic error function as the loss function, since the loss function represents the Euclidean norm (with parameter equal to 2) we name the boosting algorithm L2 boosting. With the development of the new method, the next years of research on the algorithm focused on the theoretical foundation of the instrument, along with its main applications. The article also presents the two main tuning parameters of the algorithm for regression: M, the stopping iteration factor, and ν, the step-length factor. More details about their role in the algorithm will be presented in the future, but their presentation is important because part of the following discussion is about the best ways to choose these parameters.

The main authors responsible for the theoretical foundation of L2 boosting were Bühlmann and Yu (2003) and Bühlmann and Hothorn (2007). In the article Boosting with the L2-Loss: Regression and Classification, the authors verify that computationally L2 boosting is successful if learning is sufficiently weak. Attempting to apply the method to strong learners will lead to so-called overfitting, i.e., an extremely biased fit. The two types of weak learners that are explored in the article are the linear learners and the splines learners, the same as those used in this project. Following the history of the algorithm, in the article Boosting Algorithms: Regularization, Prediction and Model Fitting (BÜHLMANN; HOTHORN, 2007), the authors present explanations and illustrations of boosting concepts, along with new derivations of the model. The work is of a practical nature, focusing on presenting empirical examples of boosting applications, as well as the explicit algorithms. To support this presentation, the authors develop the dedicated mboost ("model-based boosting") package (HOTHORN et al., 2021) for the R language, which is currently being updated and serves as one of the main equipment for actual practice work with

the algorithm.

A key application soon realized for the model consisted of prediction exercises in high dimensionality (BUEHLMANN, 2006). One problem encountered in ordinary regression exercises is the limit on the number of covariates imposed by the number of observations. A system of equations must be solved to find the parameters; thus, if there are more variables than equations available, it becomes impossible to solve the system. However, the L2 boost does not have this constraint. Bühlmann's article, Boosting for High-Dimensional Linear Models, is the first to demonstrate the consistency of the exercise for the algorithm:

> *As the main result, we prove here that L2-Boosting for linear models yields consistent estimates in the very high-dimensional context, where the number of predictor variables is allowed to grow essentially as fast as O(exp(sample size)), assuming that the true underlying regression function is sparse in terms of the l1-norm of the regression coefficients (BUEHLMANN, 2006).*

The use of boosting as a tool in high-dimensional regressions has helped open the door to applications in several areas where such an exercise was previously impossible. I highlight here the area of macroeconomics. The paper Forecasting with many predictors: Is boosting a viable alternative? by Buchen and Wohlrabe (2011) uses component-wise boosting for US industrial production forecasting. Their exercise consisted on OLS base learners and a L2-loss function. It used 139 monthly time series that spread from 1959 through 2002, and both AIC and cross-validation were tested for the selection of the meta parameter M. By comparing with a set of other forecasting models, the authors concluded the boosting algorithm was a serious competitor for forecasting US industrial production.

The same authors in Wohlrabe and Buchen (2014) show the competitiveness of the boosting algorithm for a larger macroeconomic data set in a high-dimensionality exercise. The paper not only compared the algorithm with a consolidated benchmark, but also their objective was to analyze to what extent different configurations of the model could impact forecast performance. As a result, boosting mainly outperforms the autoregressive benchmark and, for the selection of the meta parameter M, the cross-validation technique outperformed the traditionally employed information criteria.

Forecasting regional macroeconomic aggregates has also been a subject of application of boosting. Robert Lehmann and Klaus Wohlrabe use component-wise boosting as a forecasting tool for the quarterly gross domestic product of three regions in Germany, the state of Saxony, the state of Baden-Wuerttemberg and East Germany from 1997 to 2013, in their paper Boosting and regional economic forecasting: the case of Germany (LEHMANN; WOHLRABE, 2017).

The purpose of the paper was to validate the application of boosting in the forecasting of regional variables and to observe the choices made by the algorithm. As a benchmark, they used the SARIMA model, and in the selection of M the authors used AICc. The authors conclude that boosting is quite successful in its goal, being superior to SARIMA for $h = 1$ and $h = 2$. Additionally, regional variables appear frequently in the algorithm's choice, indicating a preference for such indicators.

Another application in macroeconomics that the boosting algorithm has already been used to is to forecast macroeconomic variables that can be disaggregated. For example, following the definition of the gross domestic product of some country, we could also forecast each component independently and then add these results. Furthermore, it is possible to think that some components, when selected, present a better forecasting performance than others. Within this discussion, Jing Zeng publishes her article Forecasting Aggregates with Disaggregate Variables: Does boosting help to select the informative predictors? (ZENG, 2017). Author Jing Zeng's work shows the effectiveness of using disaggregate variables combined with the boosting method to predict the original aggregate variable.

Related to the comparison of linear and non-linear boosting models, Boosting nonlinear predictability of macroeconomic time series (KAUPPI; VIRTANEN, 2021) brings both OLS and spline learners for a forecasting exercise with 129 macroeconomic time series. The authors use lags of the target variable in the exercise, and present three different models: (i) a model using only OLS linear learners; (ii) a two-staged model, where first a conventional linear auto-regressive model is applied and then a boosting with spline learners is used; and finally, (iii) a model using only spline learners. The paper indicates, by looking at the out-of-sample forecasting results and by applying the Giacomini & White test of Predictive Ability (GIACOMINI; WHITE, 2006), that the macroeconomic time series had a better than expected non-linear predictability, with the two-stage model bringing on average the best results. The decision regarding the three different models used in this project is based to some extend in the cited paper, with a completely linear model, a completely non-linear model, and a mixed one. The difference, however, is that here we will also use other predictors in the forecasting exercise.

A last non-unemployment related article of boosting to be considered, applied in the Brazilian context, is the article Using boosting for forecasting electric energy consumption during a recession: a case study for the Brazilian State Rio Grande do Sul by Lindenmeyer, Skorin, and Torrent (2021). The authors here also aimed to validate the use of the boosting algorithm in forecasting a time series. In this case, it consists of the monthly time series of

electricity consumption in Rio Grande do Sul from 2002 to 2017, the forecasting exercise being contemporaneous with the Brazilian political-economic crisis. The work used 822 predictors, with meteorological variables, regional economic variables, national economic variables, and international economic variables. As a benchmark, the authors used the SARIMA model. The paper's result is positive for the boosting side, which is superior to the SARIMA model for $h = 1$ and $h = 2$. Also, selecting M with AIC gives better results than selecting it by cross-validation. It is interesting to comment on the emphasis in the algorithm for moments of uncertainty, such as the recession in this case, because these are the moments where the forecast and a look for the future present the greatest demand.

2.2 FORECASTING UNEMPLOYMENT

In last section, the boosting algorithm, its history and its applications for macroeconomic time series were the main topics of the literature review. Now we move for the literature of forecasting unemployment, independently if it's done by the boosting method or not.

The first article reviewed was the only found that related boosting and specifically unemployment forecasting. The article Boosting nonlinear additive autoregressive time series (SHAFIK; TUTZ, 2009) presents a boosting algorithm with b-splines learners and tests the model in two exercises, comparing the results with a set of alternative competitive models. The first exercise is a univariate forecast for the Federal Reserve unemployment index. Fifteen lags of the series are used and prevision is done with and without seasonality control. The second exercise is done in a high-dimensionality environment, where the US unemployment rate is the target variable. The results in the first exercise are not very conclusive, with boosting performing closely with other benchmark models. However, boosting performs particularly well in the high-dimensionality forecast, where a more complex scenario is proposed. The conclusions supports the use of boosting in forecasts with multiple predictors, as done in this project.

An increasing trend in unemployment forecasting, especially for multivariate exercises, is the use of recent internet searches as predictors in the model. The articles The predictive power of Google searches in forecasting US unemployment (D'AMURI; MARCUCCI, 2017) and Short-term forecasting of the US unemployment rate (MAAS, 2020) use Google data to forecast unemployment in the US. The first article adopts a standard autoregressive model with explanatory variables to test the effectiveness of forecasting unemployment with Google data.

Comparing the root mean forecasting squared errors (RMFSE), the authors show the Google-based model outperforms a set of different benchmark models containing linear and non-linear approaches, particularly in longer time horizons. However, this superiority diminishes when larger samples and short horizons are used. The result goes in the opposite direction to Short-term forecasting of the US unemployment rate, which claims the effectiveness of Google-based models for short-horizon forecasting. This second article uses the Mixed-data sampling (MIDAS) regression proposed by Ghysels, Santa-Clara, and Valkanov (2006) and compares the results with an autoregressive (AR) model and the D'Amuri and Marcucci (2017) model. MIDAS performs well on shorter horizons compared to AR, with statistical significance using the Diebold-Mariano test (DIEBOLD; MARIANO, 2002). However, compared to the D'Amuri and Marcucci (2017) benchmark, the results are diffuse and no model has statistically significant better results.

Agents' expectations is also a set of predictors that was tested to see if it could help predict unemployment rates. In article Claveria (2019), the author studies the performance gain that exists in the addition of variables that reflect the future vision of consumers, particularly the degree of consensus between these consumers. The adopted methodology consists of first making a prediction via ARIMA as a benchmark and then adding the predictors to the ARIMA model, thus forming an ARIMAX model. The article uses monthly data from a set of European countries to test its hypothesis, and the main performance metric applied was MAPE. To ensure a test of statisical significance for the difference between the model's accuracies the author used Diebold-Mariano. In conclusion, the paper understands that in general adding the indicators of expectations implies an improvement in the performance of the models. In view of these results, we also added expectations indicators as predictors in this project.

The discussion about linear and non-linear methods has also emerged in the unemployment forecasting literature. The article Unemployment Rate Forecasting: A Hybrid Approach (CHAKRABORTY et al., 2021) combines linear and non-linear univariate techniques to predict unemployment rates in a series of countries: Canada, Germany, Japan, The Netherlands, New Zeeland, Sweden, and Switzerland. The study is motivated by the unusual behavior of unemployment time series in the last four decades, series with asymmetric cyclical movements and no consistent trend at all. Seven methods are proposed. Four single approaches: the Autoregressive Integrated Moving Average (ARIMA) model, the Autoregressive Neural Network (ANN) model, the Support Vector Machines (SVM) model, and the Artificial Neural Networks (ANN) model. Also, three two-step approaches are proposed: the hybrid ARIMA-ANN model, the hybrid ARIMA-SVM model, and the hybrid ARIMA-ARNN model. In the three cases, the first step

uses ARIMA and is meant to model the linear features of the time series and the second step uses the nonlinear technique. Performance metrics for one and three months ahead forecasting show the hybrid approaches, particularly ARIMA-ARNN, outperforms all linear and nonlinear single models consistently. This result helps motivate the exercise proposed in this project, especially the mixed model already cited.

Another article that uses non-linear techniques to forecast unemployment rates is Dumičić, Čeh Časni, and Žmuk (2015). The authors select five European countries where it is understood that the 2008 crisis had a significant impact on employment. They are Greece, Spain, Croatia, Italy and Portugal. In other words, here we also have a forecast focused on a time of recession, like what we ended up having in this project when forecasting unemployment rate in Brazil. The data are quarterly and the forecast outside the training sample occurs from 2008 to 2013. The objective of the article is to determine the most accurate of three smoothing methods for short-term unemployment forecasts: (i) Double exponential smoothing, (ii) Holt-Winters' multiplicative method, and (iii) Holt-Winters' additive method, where the choice of the smoothing method is justified by its ability to quickly adjust to changes in trends. To test the three models, the MAPE, MAE, and RMSFE metrics are compared. The analysis of the results does not provide an optimal model for all cases, since the choice of the best model ends up depending on the country under analysis and up to a certain limit on the metric used for the evaluation. It is important to highlight this variability because a good model does not only depend on its own characteristics, but it is also important to know when it is valid to use it. For example, different situations may require more specific models. In the case of the article, the Double exponential smoothing model is selected as the best candidate for the cases of Portugal and Spain. For Italy and Croatia, the model selected was the Holt-Winters' multiplicative method. Finally, for Greece, the optimal choice was the Holt-Winters' additive method.

Within the Machine-Learning prediction literature, article Katris (2020) compares several models to forecast time series of unemployment rates from more than 20 countries. There are five sets of models analyzed: the FARIMA, FARIMA/GARCH, Artificial Neural Networks (ANN), Support Vector Regression and Multivariate Adaptive Regression Splines models. As a benchmark, the ARIMA and Holt-Winters models are used and, as usual in the literature, the comparison metrics are MAE and RMSFE. The motivation for the diversity of models presented, according to the author, consists of the attempt to capture the different characteristics that may be present in the studied series, using models with respect to the characteristics of long-memory, heteroskedasticity (ie FARIMA and ARIMA/GARCH models) and non-linearity

(ie ANN, SVR and MARS models). In addition, an attempt is made to understand whether there is any dependence in choosing the most efficient model in relation to forecast horizons and geographic location of countries. For this, the author uses Friedman nonparametric statistical test and post hoc comparisons. Data are monthly, adjusted for seasonality, and models are univariate. The results again do not point to a model that is superior to all: we have a better performance of FARIMA models for small forecast horizons and a better performance of ANN-type models with larger horizons, such as h=12. The result again confirms the importance of understanding that different scenarios may require different tools.

Using the literature as a reference, we adopted some practices for the analysis of boosting in the Brazilian unemployment context. Since SARIMA appeared a number of times as a proposed model and as a benchmark in recent literature, it is the model selected for testing boosting in this project. Also, the metrics for measuring model performance follow what have been cited in recent papers on unemployment and boosting performance: MAPE, RMSFE, and the Giacomini & White test. How SARIMA is implemented as well as how the boosting functions will be explained in the next chapters.

3 THE BOOSTING METHOD

3.1 THE ALGORITHM

The boosting algorithm constructs iteratively a linear or non-linear model, depending on the model specifications. In this exercise, we will use both approaches. Given a vector of predictors $\vec{x}_t$ the boosting method is described as the following adjusted function:

$$\hat{f}(\vec{x}_t) = \hat{f}^{(0)} + v \sum_{m=1}^{M} b(x_t^m) \tag{1}$$

Where $\hat{f}^{(0)}$ is a constant value, x_t^m is the selected predictor in the mth iteration and the parameters v and M represent the metaperemeters of the model, with $v \in (0, 1)$ and $M \in \mathbb{N}$. The notation b represents univariate functions, and they consist of learners of boosting. In this project, b can take the form of a linear or a p-splines function. For example, the linear form of the function b is described as

$$b(x_t^m) = \beta \cdot x_t^m, \tag{2}$$

for some $x_t^m \in \vec{x}_t$ and β being a constant. The parameter v is known as the shrinkage parameter. It reduces the learner's variance and is used to improve the method's performance. It can be interpreted as the size of the step we are willing to make in each iteration. For example, if $v = 1$ we would embrace all the impact of the learner in the final model. On the other hand, if $v = 0$, we would not be embracing new learners at all. In the literature of boosting, the established value for v is 0.1, hence this is the value we use in the forecasting exercises.

The parameter M shows how many iterations the boosting algorithm is going to have. One needs to be careful when selecting the value of M, since too many iterations can produce over-fitting[2] in the model, and too few steps may not be enough to appropriately use information from predictors. The usual approach selecting M is using an information-criteria or a cross-validation process with a predetermined ceiling value. In this project we use k-fold to choose the number of iterations M, and we also set the maximum value M so that the choice of the ceiling of M does not interfere with the algorithm choice of M. For this, we select a maximum value

[2] Overfitting describes when a statistical model fits the previously observed data set very well, but proves ineffective at predicting new outcomes.

of M large enough for the choices to be smaller than the maximum, depending on the model structure.

In practice, the structure of the algorithm follows the following logic. First it sets the function of y as the average of y, such that $\hat{f}^{(0)} = \bar{y}$. After that, the algorithm regress the residuals created in this first guess $u^{(0)} = y - \hat{f}^{(0)}$ with each variable following it's respective model specifications. For example, it may regress $u^{(0)}$ with predictor x_i using a linear model, as in equation 3, or using a p-splines model, as in equation 4.

$$\hat{E}[u^{(0)}|x_i] = \hat{\beta}x_i \tag{3}$$

$$\hat{E}[u^{(0)}|x_i] = \hat{S}(x_i) \tag{4}$$

After regressing the residual with each predictor, the algorithm calculates the Sum of Squared Residuals (SSR) for each model and chooses the ith predictor with the lowest SSR. Note that this means the algorithm observes all possible contributions of the predictors and selects the combination of variable and model with the best fit for the error that still exists in the prediction. Bellow there is a summary of the step-by-step algorithm instructions.

1. Set $m = 0$ and begin with $f^{(0)} = \bar{y}$.

2. From each iteration step from $m = 1$ to M repeat:

   2.1. Calculate the error vector $u^{(m)} = y - f^{(m-1)}$.

   2.2. Regress $u^{(m)}$ with each predictor x_i, $i \in \{1, 2, ..., n\}$, using the specified function g_i (g_i being a linear or a p-splines univariate function of predictor x_i) and creating a prediction $\hat{u}_i^{(m)}$.

   2.3. Select $i^* \in \{1, 2, ..., n\}$ that minimizes the Sum of Squared Residuals $\sum_{k=1}^{t}(\hat{u}_i^{(m)} - u^{(m)})^2$.

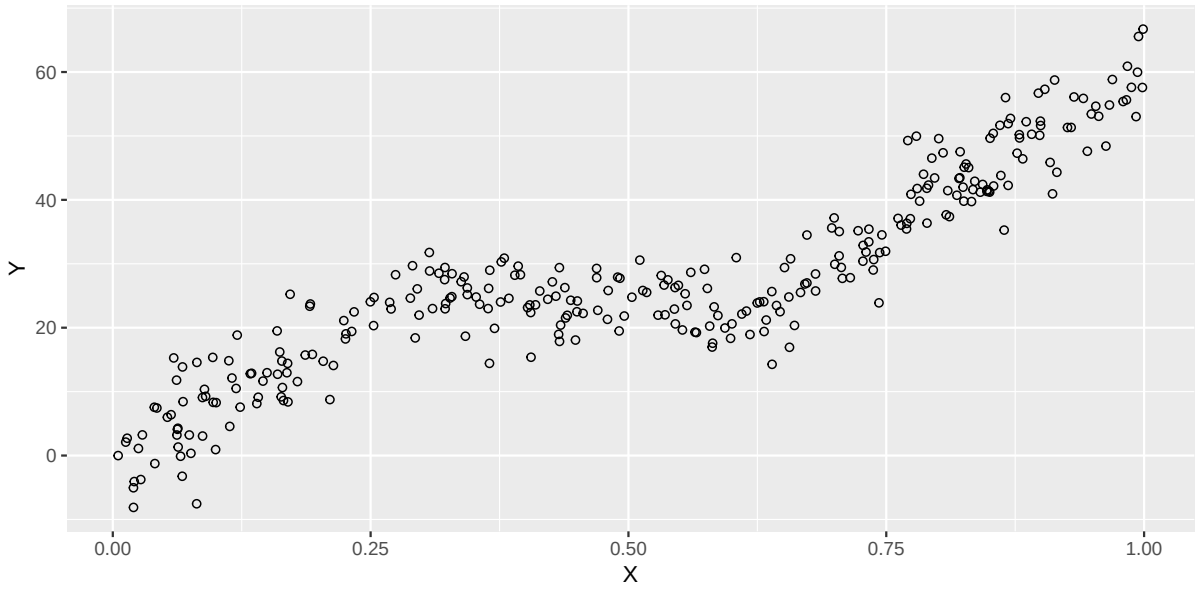
   2.4. Set $f^{(m)} = f^{(m-1)} + v \cdot g_{i^*}$.

3.2 P-SPLINES IN PRACTICE

As the name indicates, p-spline is a special type of spline function. A spline function is a piecewise polynomial function that maps values from a domain $[a, b]$ to the set of real numbers

$\mathbb{R}$. To build the spline function, first we partition the domain into n disjoint subintervals of $[a, b]$. Then, for each of the subintervals, we define polynomial functions P_n and the spline consists of the combination of the various polynomials created in all the subintervals. Two important parameters of the function are the knots and the degree of the spline. The knots are the points that define the partitions; that is, if we divide the domain $[a, b]$ into n subintervals, we will have $n - 1$ knots. The degree of a spline is related to its polynomials P_n that make up the function, where the degree of the spline is defined as the highest degree among the P_n polynomials. In this project, we use first-order splines.

Figure 2 – Scatter-plot example



Source: Elaborated by the author (2022)

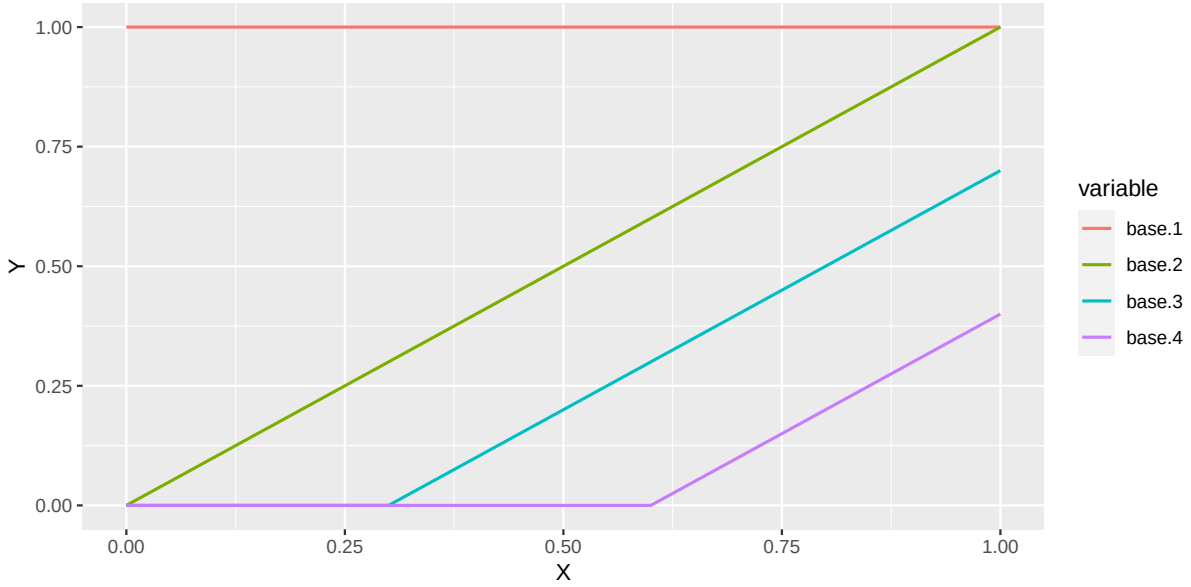
One way to represent the first-order spline model is to use basis functions of the form $(x - k)_+$, where k represents a knot in the model. The expression $(x - k)_+$, which could also be called the "positive part" of the function $x - k$, represents a function that is zero for values to the left of k and of the form $(x - k)$ for values to the right of k. The + sign is an indication to zero the function in the case of negative values. To illustrate how this helps to represent a spline model, take as an example data from Figure 2 and suppose that we select as knots the points from the x-axis 0.3 and 0.6. The basis for creating the first-order spline model would be given by the expressions[3]:

$$1, x, (x - 0.3)_+ \text{ and } (x - 0.6)_+$$

[3] The basis 1 represents a constant value.

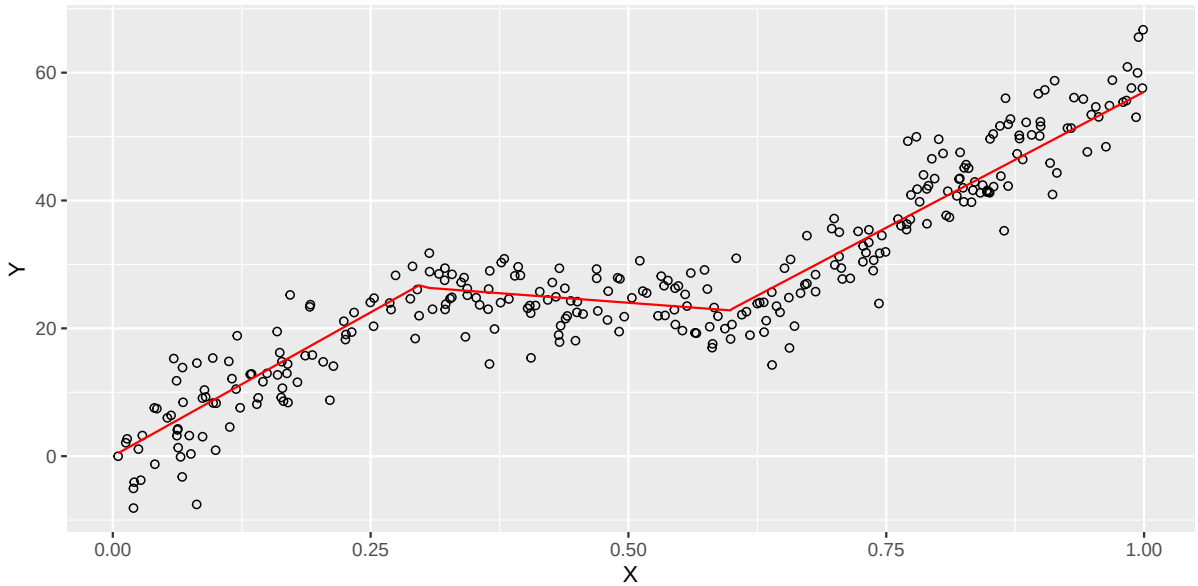
and graphically are represented by the Figure 3. To create the splines model, we find the parameters that fit the mentioned basis functions to the data we have. In doing so we find the results in Figure 4.

Figure 3 – Linear spline base learners



Source: Elaborated by the author (2022)

Figure 4 – Curve-fitting using splines 1



Source: Elaborated by the author (2022)

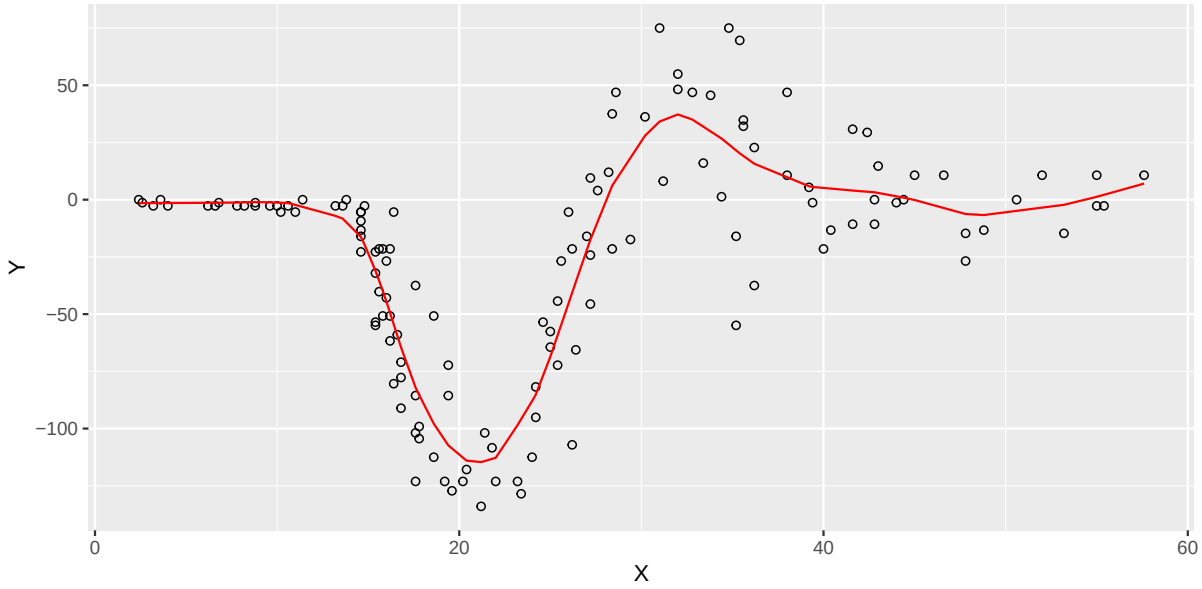
The generalization of the choice of knots leads us to the following way to represent the first degree spline model. Here, K represents the total number of knots in the model, and each $(x - \kappa_k)_+$ represents a *linear spline basis function*.

$$f(x) = \beta_0 + \beta_1 \cdot x + \sum_{k=1}^{K} b_k (x - \kappa_k)_+. \tag{5}$$

Remember, though, that we use p-splines in this project. The difference in p-splines from the one just explained is the penalty given on the coefficients b_k when fitting the model, thus given the name *penalized-splines* (p-splines). Using an adequate number of knots and a proper penalty format, the fit to the data will be flexible and smooth enough, as in the example in Figure 5. We follow the penalty format given in article Eilers and Marx (1996). To get into the details of splines and p-splines, we recommend the basic reference Ruppert, Wand, and Carroll (2003).

A final example that can help in the understanding of p-splines curve-fitting application is Figure 5, where using motorcycle crash helmet impact data from Silverman (1985) we illustrate a fitted p-spline function. The spline divides the domain [0,60] using 20 knots that are at equally spaced quantiles and a first degree polynomial. Note that in the example presented, we have an efficient solution to fit non-linear data.

Figure 5 – Curve-fitting using splines 2



Source: Elaborated by the author (2022)

In this project, the calculation of the boosting curves for both linear learners and p-spline learners is done using the mboost package implemented in R. For the p-splines we use 20 equidistant knots and polynomials of one degree as parameters of the function 'bbs'. The algorithm used in the package to calculate the p-splines functions is based on the article Eilers and Marx (1996).

4 FORECASTING EXERCISE SPECIFICATIONS

4.1 THE FORECAST ENGINE

The forecasting mechanism is an expanding window exercise that recalculates the parameters for each observation that is predicted. The exercise being an expanding window means that the size of the interval that the model uses to estimate its parameters increases over iterations; that is, at each iteration, we incorporate the most up-to-date information without removing older information. For example, if when forecasting the observation y_t we use the interval $\{y_1, y_2, ..., y_T\}$ to train the model, when forecasting y_{t+1} we would use the interval $\{y_1, y_2, ..., y_T, y_{T+1}\}$. Additionally, in each iteration, the algorithm calculates the parameters of the model again, incorporating potentially new information about the relationship between the predicted time series and its predictors.

To fit the different forecast horizons, we create a different dependent series y^{ph} to be predicted for each horizon h. We will show in next chapter the unemployment rate will need a transformation in order to reach stationarity. These transformations involve taking the log followed by a difference in the series. When we select the parameter h, we take the difference between the observation t and $t - h$ to build the new desired time series.

$$y_t^{ph} = ln(y_{t+h}) - ln(y_t) \tag{6}$$

By doing this shift, we calibrate the algorithm to consistently predict h steps ahead in the time series, so when applying the parameters to the most recent observations, the algorithm returns a prediction for the observation of the index $t + h$. To find the predicted value of y_t^{ph}, we estimate the following equation using the algorithm explained in last chapter:

$$\hat{y}_t{}^{ph} = \hat{f}_h^{(0)} + v \sum_{m=1}^{M_h} b^h(x_t^m) \tag{7}$$

The last step in the forecasting mechanism is to return the predicted value to its original form. We do this by summing the predicted value $\hat{y}_t{}^{ph}$ with the logarithm of y_t and then taking the exponential of the result:

$$\hat{y}_{t+h} = e^{(ln(y_t)+\hat{y}_t^{ph})} \tag{8}$$

For the prediction exercises performed in this paper, we set an index t^* to be the first

predicted observation, regardless of the h chosen. Note that in this format, for larger values of h, we have fewer observations being used in the training interval, since the initial observation is fixed at $t = 1$, thus shorting the interval of parameter estimation.

4.2 BENCHMARK SARIMA

To understand whether the proposed models are an efficient method of forecasting the desired time series, we need to compare their performance with what is called a benchmark. A benchmark model serves to verify the validity of the proposed model, one of the main objectives of this work. As we saw in the review of the literature, the SARIMA model was used regularly for this task; therefore, we embrace its use in this project.

The SARIMA model is an auto-regressive integrated moving average method. Unlike the models proposed in this paper, SARIMA is a univariate statistical method, which means that it uses only one time series as a predictor for the forecasting exercise. However, decisions have still to be made related to the model parameters. The main parameters that have to be selected are the ones of the composition of SARIMA: the number of AR (autoregressive) time lags, the order of MA (moving avarage) composition and also the number of times the data have had past values subtracted.

To decide on the model parameters, we use the Hyndman and Khandakar (2008) automatic time series forecast algorithm. In each iteration, the algorithm selects the optimal parameters using an AIC criterion. Also, we open the possibility for seasonal components for the algorithm to choose when selecting the components of SARIMA, thus ensuring a possible seasonal dynamic. In short, the algorithm applies a sequence of different tests to select the parameters. If no seasonal component is found, the algorithm considers a model $ARIMA(p, d, q)$ where the selection of the parameter d is based on successive KPSS uniroot tests as (KWIATKOWSKI et al., 1992). On the other hand, if a seasonal component is found, the algorithm considers an $SARIMA(p, d, q)(P, D, Q)_m$ model where D is selected depending on an extended Canova-Hansen test (CANOVA; HANSEN, 1995)[4].

Following the explanation given by Hyndman and Khandakar (2008), in the case of model $ARIMA(p, d, q)$, it represents a process given by

[4] For more detail about the algorithm decision making process, see sections 3.1 and 3.2 of Hyndman and Khandakar (2008).

$$h_1(B)(1 - B^d)yt = c + h_2(B)\epsilon_t \tag{9}$$

where $\{\epsilon_t\}$ is a white noise process with mean zero and variance σ^2, h_1 is a polynomial of order p, h_2 is a polynomial of order q and B is the backshift operator. On the other hand, for the case of $SARIMA(p, d, q)(P, D, Q)_m$, it represents a process given by

$$h_3(B^m)h_1(B)(1 - B^m)^D(1 - B)^d yt = c + h_4(B^m)h_2(B)\epsilon_t \tag{10}$$

where h_3 and h_4 are polynomials of order P and Q respectively. With d and D known, the choice for p, q, P and Q is made using an information criterion.

4.3 PERFORMANCE MEASURES

To understand and compare the forecast quality of each proposed model, we introduce in this section performance measures. Performance measures aim to classify different models in terms of their ability to forecast a specific time series, and they generally do this by calculating the magnitude of error between the forecasting values and the real observations.

What is considered a good forecast depends on a set of different characteristics of what is forecasted. For example, in some scenarios, detecting a fall in a time series may be more important than detecting a rise, therefore, more weight should be given to a low capacity in predicting falls. In the case of this work, we choose performance measures that are already applied in the literature of unemployment rate forecast, so we are aligned with established measures.

The first performance measure selected is the Mean Absolute Error (MAE). As the name indicates, we get this indicator by taking the arithmetic mean on the Euclidean distance between the forecasted values and the real observations. To emphasize the performance measures use out-of-sample forecasting data, assume we are using the index $\{1, 2, ..., t\}$ to train the models and the index $\{t + h, ..., T\}$ to test them. With this notation, MAE is described as the equation below:

$$MAE = \frac{1}{(T - t + h - 1)} \sum_{i=t+h}^{T} |\hat{y}_i - y_i|. \tag{11}$$

The second performance measure aims at understanding the proportion of this absolute error with the real observations. It is the Mean Absolute Percentage Error (MAPE), and we

calculate it by dividing the absolute forecast error by the original value.

$$MAPE = \frac{1}{(T - t + h - 1)} \sum_{i=t+h}^{T} \frac{|\hat{y}_i - y_i|}{y_i}. \tag{12}$$

The third performance measure penalizes larger errors more proportionally to errors closer to the real observation. It does this by taking the average not on the simple errors, but on the square of them. It is the Root Mean Square Forecast Error (RMSFE):

$$RMSFE = \sqrt{\frac{1}{(T - t + h - 1)} \sum_{i=t+h}^{T} (\hat{y}_i - y_i)^2}. \tag{13}$$

The forth and fifth performance measures aim to bring information regarding the distribution of the forecast errors. P90 and P95 indicate the 90% and 95% percentiles of the calculated absolute errors. By doing that we can understand what is considered a large deviation from the real value of each model. If they are small values, we can assume that the model rarely makes large forecast mistakes.

Lastly, as seen in Chapter 2, the unemployment forecasting literature in general uses the Diebold-Mariano test more frequently to test for statistical significance of predictive ability. However, we avoided using the Diebold-Mariano test because it assumes non-nested models, which is not the case in this project. Thus, in a similar way to work Kauppi and Virtanen (2021), we used the unconditional version of the Giacomini & White test, which does not assume non-nested models, to test for statistical significance in model forecast accuracy. In practice, we apply the *gw.test* function available in the *afmtools* package from the R environment. The null hypotheses here is that both models being compared have the same forecast accuracy, so if we reject the null-hypothesis, we build evidence to validate one model over another. An important point is the fact that the validity of the Giacomini & White test lies in the hypothesis of "nonvanishing estimation errors", that is, using expanding-window as we do in this project would not be completely adequate for the test performed. However, again in the same way as Kauppi and Virtanen (2021), we understand that the application of a fixed size to the estimation window would not fully utilize the capabilities of the boosting models.

5 DATASET

5.1 UNEMPLOYMENT RATE ANALYSIS

After explaining the boosting model and specifications regarding how the forecast will be conducted, we turn to the dataset used in the exercises. The selected predicted time series y_t is the unemployment rate of the Metropolitan Region of São Paulo (MRSP). Its source is the Fundação Sistema Estadual de Análise de Dados, Pesquisa de Emprego e Desemprego (Seade/PED) and the time series contains hidden unemployment (precarious work[5] and discouragement unemployment) and open unemployment. This differentiation is important in emerging countries such as Brazil, where informal work represents a significant portion of jobs.

The unemployment rate in MRSP has monthly frequency and because of its definition, $y_t \in [0, 1]$. Our time series analysis starts in January 1996 and runs through July 2019, so we have 282 observations of more than 20 years of Brazilian employment history. In these twenty years, Brazil has experienced five technical recessions, defined as a drop in GDP for at least two consecutive quarters. As the algorithm will be tested to predict waves of unemployment, it should be able to capture information from past crises to understand the dynamics of upcoming ones. In table 1 we present the five recessions present in the mentioned period together with specific information.

Table 1 – Brazilian recessions between 1996 and 2019

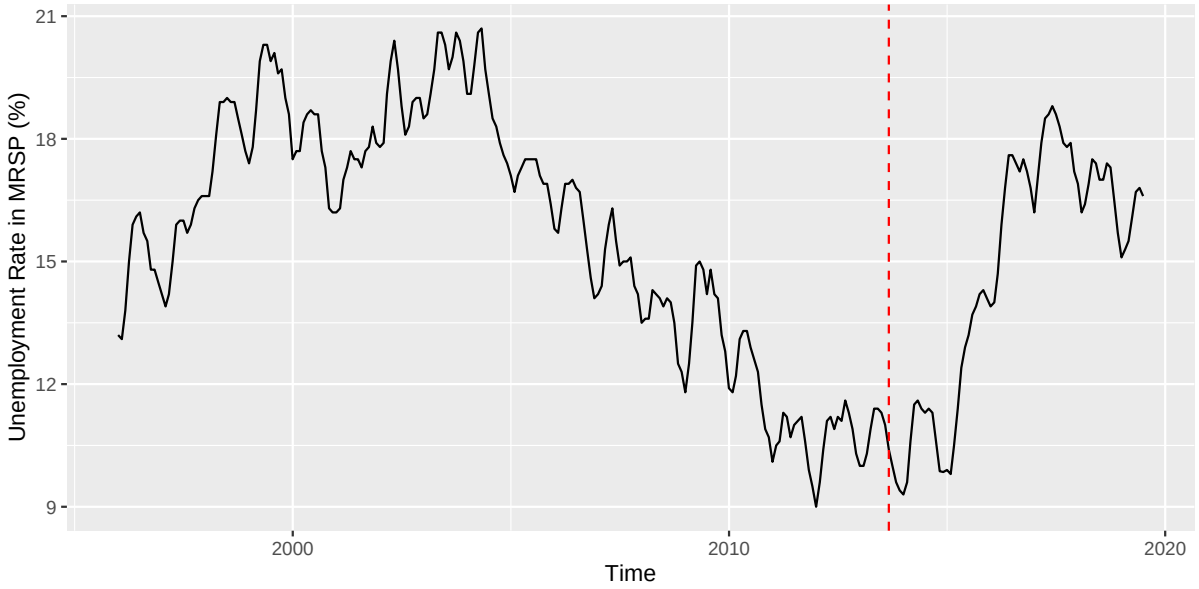
Recession Period	Duration	Cumulative Drop in GDP
1998-99	5 quarters	1,6%
2001	3 quarters	0,8%
2003	2 quarters	1,3%
2008-09	2 quarters	6,2%
2014-16	11 quarters	8,6%

Source: Elaborated by the author (2022)

As we use 12 lags of each predictor, the series that actually enters the algorithm starts in January 1996, resulting in a total of 271 complete observations. We divide the time series into two groups, the first being the training group, with 75% of the observations, and the second being the test group, with 25% of the observations. The 75% cut takes place in November 2013, that is, the boosting will have four previous recessions in its history and will be tested most decisively in the fifth and strongest recession.

[5] The concept of precarious work depicts a type of occupation in which people performed some work irregularly while looking for a more consistent occupation.

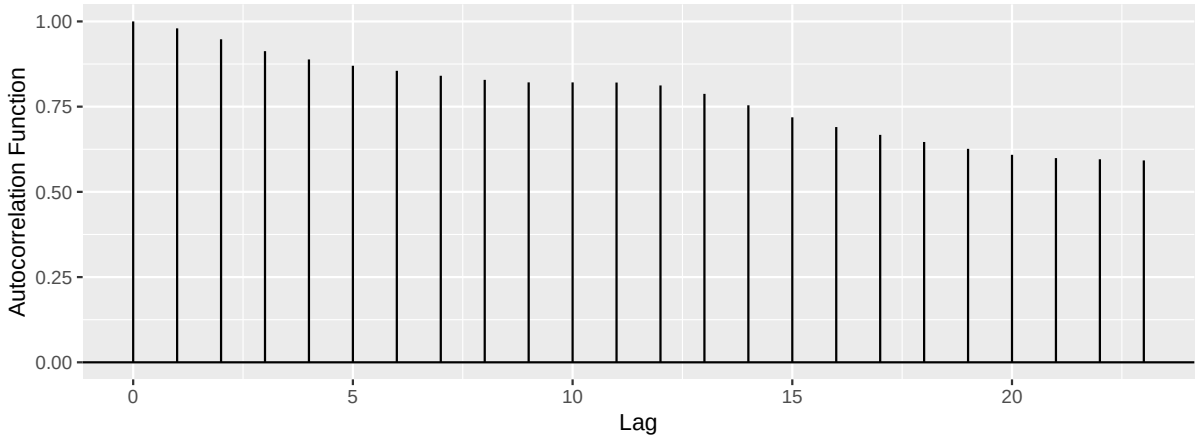
Figure 6 – Unemployment rate in MRSP with training and test cut



Source: Elaborated by the author (2022)

As explained in the forecast engine section,we will not apply the boosting algorithm exactly to the time series above. This happens because we ensure stationarity before engaging in the prediction exercise. Observing the auto-corelation plot for the train section of the time series, we have strong evidence of unit root presence.

Figure 7 – Autocorrelation function of the y_t train section



Source: Elaborated by the author (2022)

To build more evidence of non-stationarity from the original series, we performed the Dickey-Fuller test using three models. The first relates the difference of y_t only with its lag. The second model adds a constant α_0 to the first specification. Finally, the last model adds to the second a deterministic trend t. The Dickey-Fuller test corroborates the unit root thesis derived from the observation of the partial autocorrelation plot, which leads us to redo the test with

Table 2 – Augmented-Dickey-Fuller unit root test

Original Series

Model 1 $\Delta y_t = \tau y_{t-1} + u_t$

Value of test-statistic	τ		
	-0.5124		
Critical values	1%	5%	10%
τ	-2.58	-1.95	-1.62

Model 2 $\Delta y_t = \phi\alpha_0 + \tau y_{t-1} + u_t$

Value of test-statistic	τ	ϕ	
	-1.7315	1.5171	
Critical values	1%	5%	10%
τ	-3.46	-2.88	-2.57
ϕ	6.52	4.63	3.81

Model 3 $\Delta y_t = \phi_1\alpha_0 + \tau y_{t-1} + \phi_2 t + u_t$

Value of test-statistic	τ	ϕ_1	ϕ_2
	-3.5934	4.7518	7.1087
Critical values	1%	5%	10%
τ	-3.99	-3.43	-3.13
ϕ_1	6.22	4.75	4.07
ϕ_2	8.43	6.49	5.47

Transformed Series

Model 1 $\Delta y_t = \tau y_{t-1} + u_t$

Value of test-statistic	τ		
	-8.0038		
Critical values	1%	5%	10%
τ	-2.58	-1.95	-1.62

Model 2 $\Delta y_t = \phi\alpha_0 + \tau y_{t-1} + u_t$

Value of test-statistic	τ	ϕ	
	-7.9941	31.9691	
Critical values	1%	5%	10%
τ	-3.46	-2.88	-2.57
ϕ	6.52	4.63	3.81

Model 3 $\Delta y_t = \phi_1\alpha_0 + \tau y_{t-1} + \phi_2 t + u_t$

Value of test-statistic	τ	ϕ_1	ϕ_2
	-8.0531	21.6316	32.4307
Critical values	1%	5%	10%
τ	-3.99	-3.43	-3.13
ϕ_1	6.22	4.75	4.07
ϕ_2	8.43	6.49	5.47

Source: Elaborated by the author (2022)

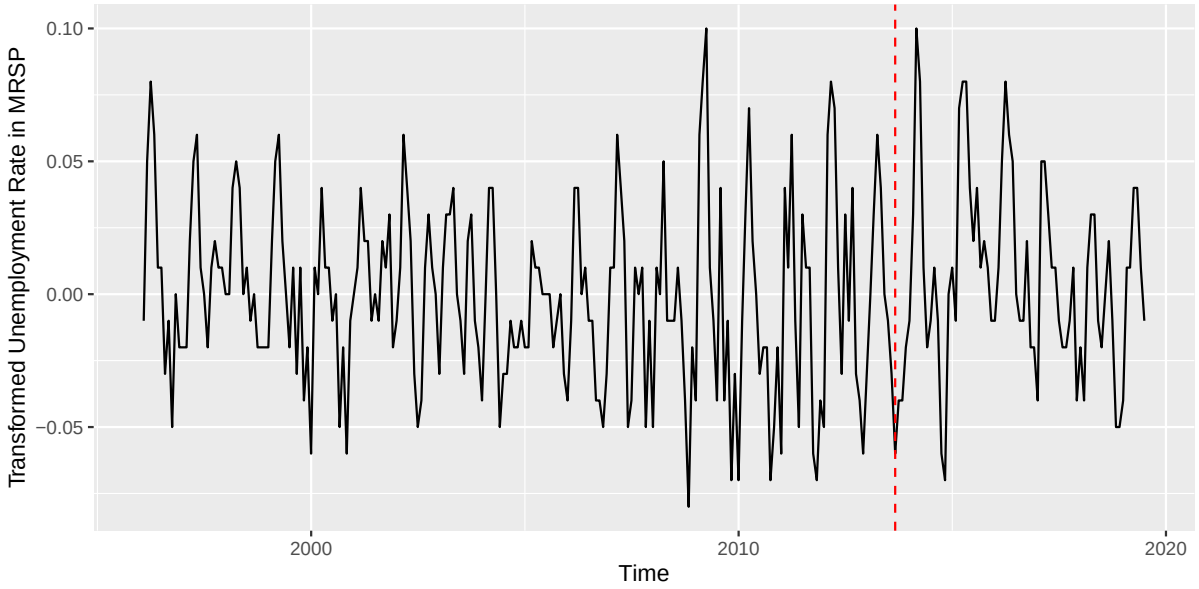
the difference of the log of the series. In this second test performed with the difference of the logarithmic value of the original series, we reject the null hypothesis of the presence of a unit root with significance of 1% for all model specifications [6]. We conclude applying the boost algorithm in this transformed series, already stationary according to the applied tests. The test results are shown in Table 2 and the plot for the transformed series appears in Figure 8.

5.2 PREDICTORS

Inspired by the FRED-MD data set introduced by McCracken and Ng (2016), which contains variables from several different macroeconomic themes and aims to help studying US macroeconomic dynamics, in this project we built a Brazilian macroeconmics FRED-MD style data set. For that, we used the IPEADATA website combined with its Python-dedicated library *ipeadatapy* and extracted 159 variables from ten different themes. The themes were selected with the objective of providing a broad overview of the Brazilian economy, with both real and monetary predictors. All the time-series selected have monthly frequencies and also contain observations from January 1996 through July 2019. The themes and the number of variables can be seen in table 3.

[6] Because of the predictive nature of the project, we do not go into the interpretative details of the transformed variable, but focus on the stationary characteristic of the new time series.

Figure 8 – Difference of the log of unemployment rate in MRSP



Source: Elaborated by the author (2022)

Table 3 – Data breackdown

Theme	**Number of Predictors**
Consumption and Sales	28
Currency and Credit	17
Employment	5
Exchange Rates	7
Financial Accounts	5
Foreign Trade	21
National Accounts	7
Perception and Expectations	2
Prices	64
Wages and Income	3
Total	**159**

Source: Elaborated by the author (2022)

Transformations had to be done in some of the 159 variables in order to reach stationarity. To do this, we classified the predictors into three transformation groups, each of which specifies a different transformation requirement. The purpose of transformation groups is to enable the series to apply a difference function. For example, in the case of a strictly positive series, the first group of predictors, we can apply the logarithmic difference function to achieve this goal. The second group consists of series with positive or zero values. In this case, the difference function consists of summing a constant k over the entire series, thus making it strictly positive, and then applying the log difference. The third group consists of series with negative values. Here the log function cannot be applied, so we perform the relative difference, that is, the ratio between the

index value t with the index value $t-1$: $(x_t - x_{x-1})/x_{t-1}$.

With this grouping in mind, we applied an algorithm to find the stationarized version of all predictors. The algorithm starts by applying the Dickey-Fuller test to verify that the original series is stationary. If the test returns a null hypothesis rejection for 5% significance, we stop here. If not, we start the transformations iteratively. Depending on the group, the algorithm applies specific transformations and repeats the Dickey-Fuller test until it finds an iteration in which the test returns a rejection of the null hypothesis for significance of 5%. When we find such an iteration, the algorithm stops and returns the transformed series. This is the series that we use as a predictor in the boosting prediction exercises.

Table 4 – Transformations

Transformation Name	Tranformation	Number of Predictors
T0	No Transformation Applied	50
T1	First Difference	3
TL0	Log of the Series	1
TL1	Log and First Difference of the Series	93
TL2	Log and Second Difference of the Series	4
TLK1	Sum of a Constant K, Log and First Difference of the Series	2
TR1	First Relative Difference of the Series	5
TR2	Second Relative Difference of the Series	1

Source: Elaborated by the author (2022)

Table 5 – Transformation by theme

Theme	T0	T1	TL0	TL1	TL2	TLK1	TR1	TR2	Total
Consumption and Sales			1	27					28
Currency and Credit	4	2		7	2	2			17
Employment				5					5
Exchange Rates	1			6					7
Financial Accounts	4			1					5
Foreign Trade				21					21
National Accounts				6	1				7
Perception and Expectations				2					2
Prices	41	1		15	1		5	1	64
Wages and Income				3					3
Total	**50**	**3**	**1**	**93**	**4**	**2**	**5**	**1**	**159**

Source: Elaborated by the author (2022)

Detailed information on the transformation and sources for each of the predictors can be found in the Appendix A. Table 4 shows us that the most performed transformation was the time series log difference. After that we have the predictors that did not need to receive any transformation to achieve stationarity. Table 5 presents these results broken down by macroeconomic

theme. Note that the series that dominate the non-application of transformations consist of the price theme, with 41 of the 50 elements being from this theme.

To get an idea of what these categorizations really represent, figure 9 presents examples of the original series for each type of transformation performed. Those that underwent fewer transformations were those with formats closer to what we know of stationary series. Those that had to go through two differences had apparently non-linear trend elements, as in the case of the Brazilian GDP, with a trend visually closer to an exponential function.

Returning to the explanation of section 4.1, the algorithms are applied to forecast the transformed series of y_t and this transformation depends on the forecast horizon evaluated. The format of this transformation consists of Equation 6 from Chapter 4.

$$y_t^{ph} = ln(y_{t+h}) - ln(y_t) \tag{14}$$

We do not make transformations of the predictors depending on the forecast horizon. To return from the predicted series transformed to the original series, we do the inverse transformation path, according to Equation 8 from Chapter 4.

$$\hat{y}_{t+h} = e^{(ln(y_t)+\hat{y}_t^{ph})} \tag{15}$$

5.3 THE THREE PROPOSED MODELS

Taking into account both the model structure and the empirical evidence from the predicted time series, we propose three models to forecast the unemployment rate in Brazil. The three models are listed below, where $bols$ represents a linear function, and bbs a p-spline function. In addition, the vector $\vec{x}_t$ represents the set of predictors within the choice possibilities of the algorithm. Particularly in the Mixed Model we have two intermediate parameters M_1 and M_2 with $M_1 + M_2 = M$.

- Linear Model: a boosting algorithm in which all learners are linear.

$$\hat{f}(\vec{x}_t) = \hat{f}^{(0)} + v \sum_{m=1}^{M} bols(x_t^m) \tag{16}$$

- Non-linear Model: a boosting algorithm in which all learners are p-splines.

$$\hat{f}(\vec{x}_t) = \hat{f}^{(0)} + v \sum_{m=1}^{M} bbs(x_t^m) \tag{17}$$

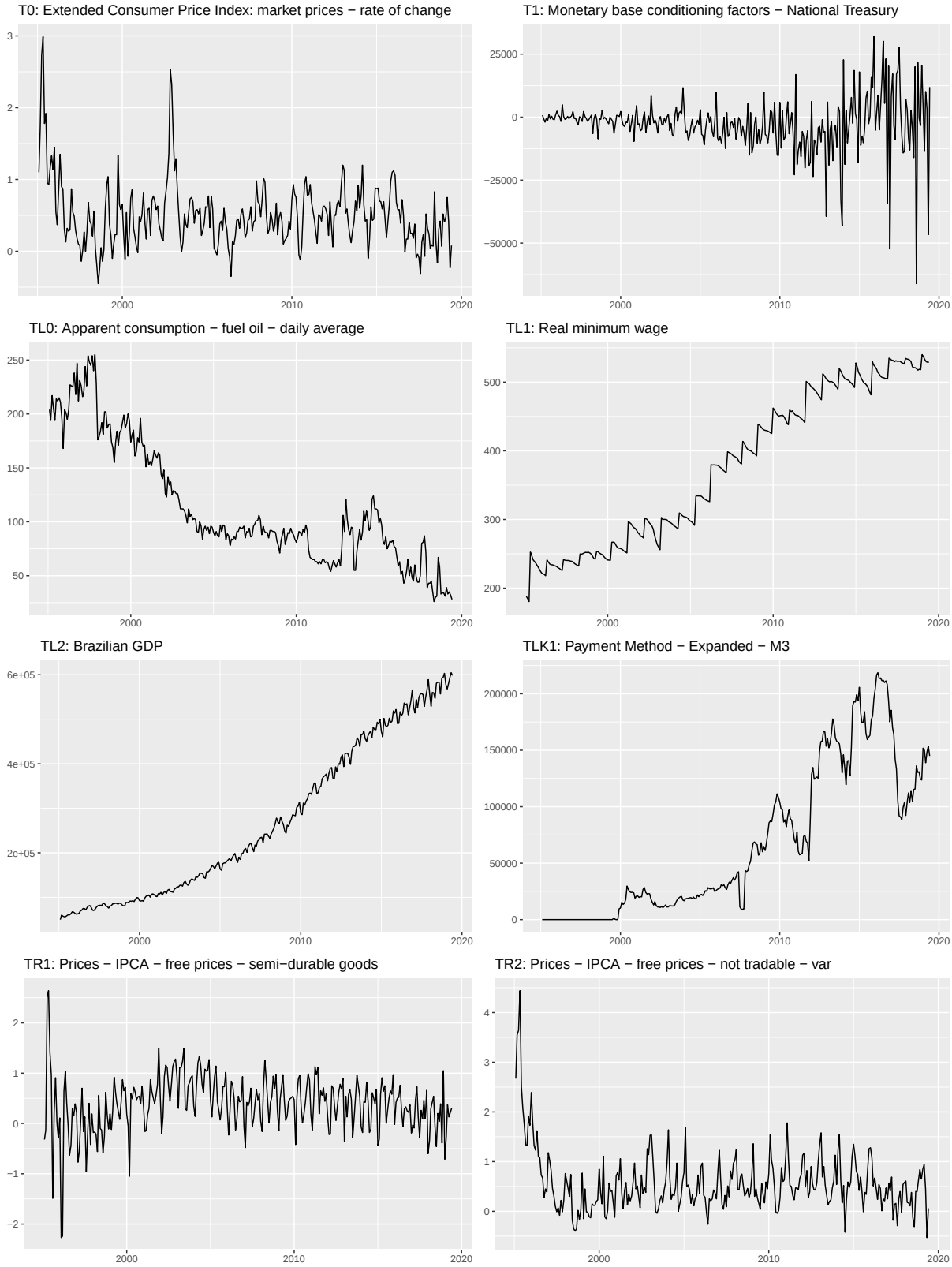
- Mixed Model: a boosting algorithm in which learners that use predictors of the Prices theme are p-splines and the rest of the learners are linear ones.

$$\hat{f}(\vec{x}_t) = \hat{f}^{(0)} + v \sum_{m_1=1}^{M_1} bols(x_t^{m_1}) + v \sum_{m_2=1}^{M_2} bbs(x_t^{m_2}) \tag{18}$$

The choice of model three, where we have the price variables in a non-linear format, takes into account established evidence of the non-linear relationship of the Phillips curve, the empirical curve that relates prices and unemployment. Articles such as Debelle and Laxton (1997), Xu et al. (2015) and Byrne and Zekaite (2020) show that non-linear models tend to fit better Phillips curves. Using these three models, we are able to isolate the non-linearity only in the price predictors to better understand whether it is more effective to present this relationship linearly or not. If we had only presented models one and three, for example, there could be doubts as to whether the performance would be explained by the non-linearity of prices or by the non-linearity of all predictors.

About the other parameters of the model, the value of the v is equal to 0.1, following most of the boosting literature. The M metaparameter is chosen via k-fold. In addition, we set a maximum value of M equal to 400 in the linear model, 1000 in the non-linear model and 700 in the mixed model. These choices happen to prevent the algorithm from hitting the maximum value in its selection process.

Figure 9 – Examples of time series by transformation type



Source: Elaborated by the author (2022)

6 RESULTS

6.1 MODELS PERFORMANCE MEASURES

In this chapter, we present the results of the forecasting exercise carried out using the three proposed models and the benchmark. The first analysis derives from the performance metrics described in chapter 3. Table 6 shows the value of each metric by model and forecast horizon of up to 12 months ahead, where the best results for a given model and forecast horizon are shown in bold. All values are presented in relation to the SARIMA benchmark; that is, values lower than 1 mean represent values lower than the benchmark.

First, comparing all models with the SARIMA benchmark, we can observe a superiority of the boosting models. This superiority appears mainly in the intermediate forecast horizons, between $h = 3$ and $h = 10$. However, in small horizons ($h = 1$ and $h = 2$) and also in larger horizons ($h = 11$ and $h = 12$), this large superiority loses space for an improvement of the SARIMA model. The trend that arises with the increase in the forecast horizon is a balance between the models when we see the performance measures, which may be related to the increase in uncertainty present in very distant forecasts. At $h = 12$, for example, the SARIMA benchmark becomes the most qualified model in metrics P95 and P90, the metrics that study the tails of the forecast error distribution. In the case of $h = 1$, the SARIMA model appears to be superior to the mixed model for P90 and P95 metrics and superior to the linear model for MAPE, RMSFE and P95 metrics. Despite this, the general analysis of the results shows a better predictive capacity of the boosting models.

Table 6 – Performance measure results

Metric	Model	h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12
MAE	Linear	0.9824	**0.8253**	**0.7050**	**0.6579**	0.7695	0.8122	0.8285	0.8236	0.8515	0.9110	0.9703	1.0338
	Nonlinear	0.9618	0.8312	0.7337	0.6957	0.7452	0.7318	0.7487	0.8224	0.8673	0.9404	0.9857	1.0458
	Mixed	**0.9470**	0.8260	0.7127	0.6742	**0.7304**	**0.7096**	**0.7102**	**0.7232**	**0.7648**	**0.7840**	**0.8465**	**0.9303**
MAPE	Linear	1.0007	0.8262	**0.6895**	**0.6461**	0.7675	0.8068	0.8413	0.8393	0.8562	0.9051	0.9659	1.0106
	Nonlinear	0.9893	0.8357	0.7230	0.6729	0.7262	0.7125	0.7367	0.8133	0.8531	0.9284	0.9692	1.0233
	Mixed	**0.9735**	**0.8211**	0.6994	0.6519	**0.7109**	**0.6934**	**0.7060**	**0.7189**	**0.7533**	**0.7789**	**0.8390**	**0.9147**
RMSFE	Linear	1.0847	**0.6788**	**0.4800**	**0.4207**	0.6142	0.6581	0.7455	0.6800	0.7216	0.7925	0.8923	0.9693
	Nonlinear	0.9861	0.7237	0.5161	0.4857	0.5128	0.5455	0.6162	0.6863	0.7710	0.9222	0.9945	1.0481
	Mixed	**0.9761**	0.6867	0.4860	0.4345	**0.4503**	**0.4689**	**0.5028**	**0.5341**	**0.5981**	**0.6823**	**0.7661**	**0.8433**
P95	Linear	1.1768	**0.9071**	**0.8272**	0.8488	0.8450	0.8683	0.9241	**0.8944**	0.9589	**0.9509**	**0.9411**	1.0202
	Nonlinear	**0.9854**	1.0214	0.8836	0.9375	0.8341	0.8917	1.0087	1.0724	0.9956	1.0732	1.1186	1.0889
	Mixed	1.0556	0.9980	0.8628	**0.7899**	**0.7473**	**0.7669**	**0.8381**	0.8976	**0.8981**	0.9955	0.9997	**1.0086**
P90	Linear	0.9935	0.9218	0.8164	**0.6527**	**0.7315**	0.8009	0.9117	**0.7022**	**0.7380**	**0.8964**	0.9293	1.0559
	Nonlinear	**0.9604**	1.0044	0.7613	0.7938	0.8533	0.7769	0.9189	0.8655	0.9361	1.0219	1.0733	1.0933
	Mixed	1.1338	**0.8723**	**0.7466**	0.7080	0.7564	**0.7469**	**0.7802**	0.7283	0.8373	0.8976	**0.9212**	**1.0195**

Source: Elaborated by the author (2022)

The analysis between the three proposed models does not present a conclusion as strong

as what we see in the case of the benchmark. There is an interaction between the best models for each forecast horizon, mainly between the Linear and the Mixed models. A trend that we can observe is a slightly better performance of the Linear model in the short-term forecast, in $h = 2$, $h = 3$ and $h = 4$, followed by a dominance of the best results by the Mixed model for $h = 5$ until $h = 12$.

By adding the times that each model had the best performance, we find the results in table 7. From it we can claim a better general performance of the Mixed model. The Mixed model seems to be able to find an efficient way to reduce forecast errors by combining linear and non-linear characteristics. For example, focusing on the MAPE metric, the Mixed model is able to perform better in 10 of the 12 forecast horizons analyzed.

Table 7 – Number of forecast horizons for which the model had the best performance

Performance Measure	Linear	Nonlinear	Mixed	SARIMA
MAE	3	0	9	0
MAPE	2	0	10	0
RMSFE	3	0	9	0
P95	5	1	5	1
P90	5	1	5	1

Source: Elaborated by the author (2022)

When looking at the results between the performance metrics used, MAE and MAPE performance seem to describe reasonably similar information. Only in the forecast horizon $h = 2$ can we see a difference in the best model selected. Furthermore, in both metrics, the benchmark presents lower results from $h = 2$ to $h = 11$ in relation to all proposed models. Only in $h = 12$ both metrics for the SARIMA model manage to be superior to a proposed model, the Linear and Non-Linear models in this case.

The RMSFE metric follows a path similar to MAE and MAPE. RMSFE selects the same best models as MAE and selects a different model only once compared to MAPE. Consistency between models is preferable, as it indicates greater robustness in the conclusions we can draw from performance metrics. If many metrics point to the same model, we can be more confident that in fact this model had a greater predictive capacity.

The last metrics, P90 and P95, present structurally different information, as they do not verify concentration, but quantiles of the forecast errors distribution. In this case the results were a little different, but nothing out of the trend found in the other indicators. Here, as we advance the forecast horizons, the SARIMA model starts to gain prominence more quickly, ending up with a good performance at $h = 12$.

Table 8 – Giacomini & White test

h=1	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.5737	0.711	0.5023
Non-linear	0.4263	-	0.6717	0.4609
Mixed	0.289	0.3283	-	0.3958
SARIMA	0.4977	0.5391	0.6042	-

h=2	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.3988	0.557	**0.028***
Non-linear	0.6012	-	0.6468	**0.0372***
Mixed	0.443	0.3532	-	**0.0251***
SARIMA	0.972	0.9628	0.9749	-

h=3	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.1879	0.3655	**0.001****
Non-linear	0.8121	-	0.73	**5e-04*****
Mixed	0.6345	0.27	-	**0.0012****
SARIMA	0.999	0.9995	0.9988	-

h=4	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.231	0.4344	**2e-04*****
Non-linear	0.769	-	0.7062	**3e-04*****
Mixed	0.5656	0.2938	-	**7e-04*****
SARIMA	0.9998	0.9997	0.9993	-

h=5	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.7488	0.7925	**0.0011****
Non-linear	0.2512	-	0.652	**0.0081****
Mixed	0.2075	0.348	-	**0.0078****
SARIMA	0.9989	0.9919	0.9922	-

h=6	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.9818	0.9865	**0.0127***
Non-linear	**0.0182***	-	0.6768	**0.0061****
Mixed	**0.0135***	0.3232	-	**0.0076****
SARIMA	0.9873	0.9939	0.9924	-

h=7	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.9081	0.9659	0.0797
Non-linear	0.0919	-	0.7843	**0.0056****
Mixed	**0.0341***	0.2157	-	**0.0031****
SARIMA	0.9203	0.9944	0.9969	-

h=8	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.624	0.9533	**0.042***
Non-linear	0.376	-	0.9987	**0.0467***
Mixed	**0.0467***	**0.0013****	-	**0.0046****
SARIMA	0.958	0.9533	0.9954	-

h=9	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.5204	0.9717	0.0748
Non-linear	0.4796	-	0.9974	0.0663
Mixed	**0.0283***	**0.0026****	-	**0.0094****
SARIMA	0.9252	0.9337	0.9906	-

h=10	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.3562	0.9926	0.1635
Non-linear	0.6438	-	0.9998	0.2172
Mixed	**0.0074****	**2e-04*****	-	**0.0101***
SARIMA	0.8365	0.7828	0.9899	-

h=11	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.4842	0.9767	0.3423
Non-linear	0.5158	-	0.9989	0.3779
Mixed	**0.0233***	**0.0011***	-	0.0654
SARIMA	0.6577	0.6221	0.9346	-

h=12	Linear	Non-linear	Mixed	SARIMA
Linear	-	0.426	0.9681	0.5463
Non-linear	0.574	-	0.989	0.627
Mixed	**0.0319***	**0.011***	-	0.1887
SARIMA	0.4537	0.373	0.8113	-

Source: Elaborated by the author (2022)

So far, we have used only the results found by the performance metrics, without getting in touch with tests of statistical significance. The next analysis applies the Giacomini & White test to compare the prediction accuracy of the models presented. Recalling, Giacomini & White test is a test applied in two models. Its null hypothesis is the equality of the forecast accuracies of these two models, and the alternative would indicate a difference in this measure. Here we look for a tail alternative, where we check if a model has lower accuracy than another.

The result of this is the set of tables 8, where each matrix presents the p-value of a Giacomini & White test. The test is applied with the null hypothesis being the equality of the prediction accuracies of the row and column models and the alternative hypothesis a better accuracy of the model of the row compared to the column model.

$$\begin{aligned} H_0 &= \text{Row and column models have the same forecast accuracy.} \\ H_1 &= \text{Row model has a better forecast accuracy than column model.} \end{aligned} \tag{19}$$

From this table we can draw two conclusions. First, the superiority of the proposed

models in relation to the SARIMA benchmark has statistical significance in the intermediate forecast horizons, bringing more evidence to the thesis presented above. However, for smaller or larger values of h, no such significance is found. Second, the comparison between the boosting models showed a greater accuracy of the Mixed model in relation to the Non-linear model for $h = 8$ up to $h = 12$ and of the Mixed model in relation to the Linear model for $h = 6$ up to $h = 12$. Again we see results similar to those seen in the past analysis. We can also see that the small dominance of the Linear model in the small forecast horizons was not reflected in statistical significance.

6.2 SELECTED VARIABLES

After verifying the predictive capacity of the proposed boosting models, we proceeded to understand the choices made by each algorithm. We can understand the choices made by looking at the predictors selected in each iteration of the algorithm. Remember that boosting uses several weak learnenrs to compose a larger strong learner, where each learner consists of a function (linear or p-splines) that uses only one predictor, so the selection of predictors is an important step in the structure of the model.

One way to analyze the choices made would be to build a table with the frequencies in which the predictors appear in each model. However, with this, we would run the risk of not actually representing the importance of each predictor in the final result of the forecast, since it is possible for a variable to be chosen with a minor role in the algorithm. Thus, analysis of the relevance of the variables takes place using a technique that attempts to circumvent this problem. We use the *varimp* function of the *mboost* R library developed in Kuehn and Stoecker (2021):

> *This function extracts the in-bag risk reductions per booster step of a fitted mboost model and accumulates it individually for each base learner contained in the model. This quantifies the individual contribution to risk reduction of each base-learner and can thus be used to compare the importance of different base-learners or variables in the model. Starting from offset only, in each boosting step risk reduction is computed as the difference between in-bag risk of the current and the previous model and is accounted for the base-learner selected in the particular step (KUEHN; STOECKER, 2021).*

With the forecasting exercise taking place for the last 71 observations of the complete time series, the analysis of the importance of the variables was performed using an arithmetic mean of the importance of each variable in the 71 times that each algorithm was applied. This means that we find the average of the importance of the variables in each model for a given

forecast horizon. It is important to note that the choice of a variable does not indicate any causal inference *per se*, what we are doing is observing the algorithm's thinking logic to extract insights into its functioning and eventually find correlations between the dependent and predictor variables. The analysis is also valid to compare the choices of different models and different forecast horizons.

To make the analysis leaner, the tables presented have the ranking of the 5 most important variables for forecast horizons 1, 6, and 12. The choice of horizons was made to compare the selection of models in distant forecast horizons, to see if it is possible to find any relationship between the horizon and the selected variables.[7]

Table 9 – Top 5 variables selected for $h = 1$

Selected variables for the Linear model	**Importance**
L12 - Unemployment rate - MRSP	32.96%
L3 - Personnel employed - industry - index (2006 average = 100)	13.38%
L1 - Unemployment rate - MRSP	10.67%
L2 - Personnel employed - industry - index (2006 average = 100)	9.20%
L10 - Personnel employed - industry - index (2006 average = 100)	4.49%
Selected variables for the Mixed model	**Importance**
L12 - Unemployment rate - MRSP	34.34%
L3 - Personnel employed - industry - index (2006 average = 100)	14.05%
L2 - Personnel employed - industry - index (2006 average = 100)	9.90%
L1 - Unemployment rate - MRSP	9.85%
L10 - Personnel employed - industry - index (2006 average = 100)	4.46%
Selected variables for the Non-linear model	**Importance**
L12 - Unemployment rate - MRSP	23.92%
L3 - Personnel employed - industry - index (2006 average = 100)	15.05%
L1 - Monetary Base - Restricted (M0) - currency issued - average	9.59%
L2 - Personnel employed - industry - index (2006 average = 100)	7.84%
L1 - Unemployment rate - MRSP	6.96%

Source: Elaborated by the author (2022)

For the three models, in the forecast horizon equal to 1, the most important variable was lag 12 of the dependent variable, unemployment in the Metropolitan Region of São Paulo. We can interpret this as an adjustment to the seasonality of the series, since lag 12 means using the value in the same month but from last year and we did not perform any seasonality control for y_t.

Focusing on the analysis for the linear model, we highlight that the five most important series for $h = 1$ are related to employment in some way, basically a mixture of lags from the dependent series with a similar series of people employed in the industry. For $h = 6$, we have an

[7] The prefix L6 indicates that the selected variable is the 6th lag of the time series.

Table 10 – Top 5 variables selected for $h = 6$

Selected variables for the Linear model	**Importance**
L1 - Personnel employed - industry - index (2006 average = 100)	23.83%
L2 - Personnel employed - industry - index (2006 average = 100)	16.74%
L9 - Unemployment rate - MRSP	6.97%
L1 - Purchasing power parity (PPP) rate - household consumption	4.40%
L1 - Exports - prices - index (2006 average = 100)	4.23%
Selected variables for the Mixed model	**Importance**
L1 - Personnel employed - industry - index (2006 average = 100)	21.38%
L2 - Personnel employed - industry - index (2006 average = 100)	14.48%
L9 - Unemployment rate - MRSP	5.26%
L1 - Exports - prices - index (2006 average = 100)	4.51%
L1 - Purchasing power parity (PPP) rate - household consumption	3.18%
Selected variables for the Non-linear model	**Importance**
L1 - Personnel employed - industry - index (2006 average = 100)	21.20%
L5 - Monetary Base - Restricted (M0) - currency issued - average	12.41%
L2 - Personnel employed - industry - index (2006 average = 100)	9.49%
L2 - Consumption - electricity - trade - quantity	3.77%
L1 - Exports - prices - index (2006 average = 100)	3.51%

Source: Elaborated by the author (2022)

Table 11 – Top 5 variables selected for $h = 12$

Selected variables for the Linear model	**Importance**
L1 - Employed personnel - industry - deseasonalized index. (2006 average = 100)	20.48%
L12 - Apparent consumption - fuel oil - average - others - average - daily amount	8.54%
L10 - Contracted exchange - financial	4.13%
L2 - Employed personnel - industry - deseasonalized index. (2006 average = 100)	3.68%
L4 - Prices - IPCA - housing - var.	3.65%
Selected variables for the Mixed model	**Importance**
L1 - Employed personnel - industry - deseasonalized index. (2006 average = 100)	12.79%
L1 - Exports - prices - index (2006 average = 100)	6.77%
L9 - Prices - IPA-DI - origin - prod. industrial - index (Aug. 1994 = 100)	6.69%
L12 - Apparent consumption - fuel oil - average - others - average - daily amount	5.76%
L12 - Prices - IGP-DI - general - centered - end of period - index (Aug. 1994 = 100)	5.75%
Selected variables for the Non-linear model	**Importance**
L1 - Employed personnel - industry - deseasonalized index. (2006 average = 100)	13.05%
L6 - Apparent consumption - fuel oil - average - others - average - daily amount	5.28%
L2 - Employed personnel - industry - deseasonalized index. (2006 average = 100)	4.83%
L1 - Apparent consumption - fuel oil - average - others - average - daily amount	4.77%
L5 - Apparent consumption - fuel oil - average - others - average - daily amount	4.59%

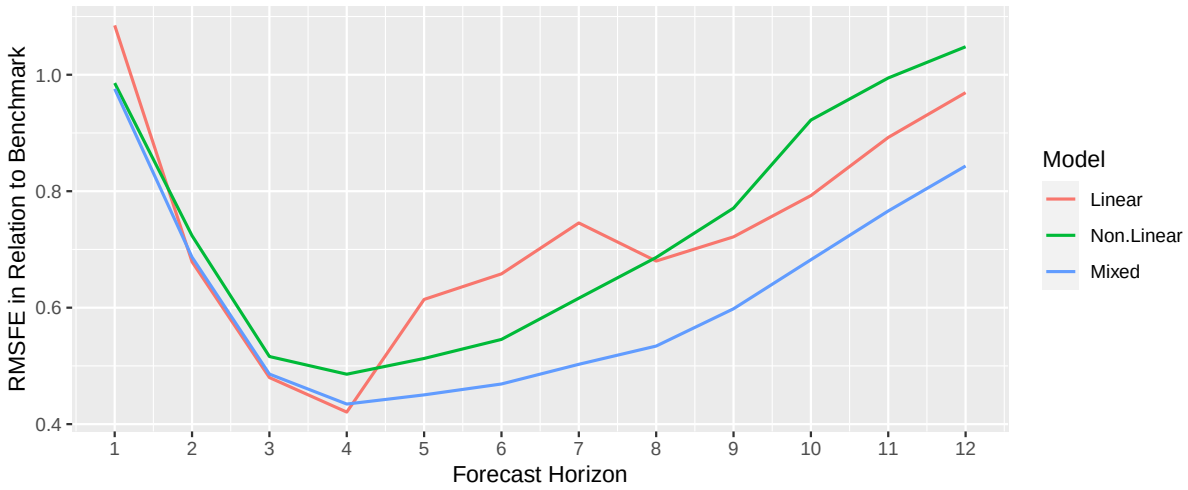
Source: Elaborated by the author (2022)

input of variables of other types, such as the purchasing power parity rate, a value for the real exchange rate. This signals a use of international economic factors to predict unemployment, which may be explained by Brazil's relationship with its economic partners. Finally, the analysis for $h = 12$ indicates a greater diversity of variables, with the entry of energy and price indicators, in addition to a predictor related to exchange rates. We see that as the forecast horizons increase,

the importance vector of the variables becomes less concentrated, and the most important predictor loses relative importance.

The numbers of the Mixed model for $h = 1$ and $h = 6$ is quite similar to what we saw in the table of the Linear model. Note that for the forecast horizon $h = 1$ and $h = 6$, the Mixed model has the same 5 most important variables as the linear one: variables mainly related to employment and for $h = 6$ with the relationship between Brazil and other countries. For $h = 12$, however, three of the five most variables in the ranking are related to prices, precisely those that have non-linear learners. It is important to emphasize this difference because, when we look at the result data of the performance metrics, for low values of h both the Mixed and Linear models have a similar behavior. However, for higher values of h the difference between the Linear and Mixed models is remarkable, as can be seen in Figure 10. Recall that the only difference between both models is the fact that the Mixed model has spline learners for price-related predictors. From this we can understand that, for longer-term forecasts, using prices in a non-linear format may be a way to improve the forecast performance.

Figure 10 – RMSFE comparison between models



Source: Elaborated by the author (2022)

The numbers of the most important variables of the Non-linear model deviates a little from the other models, especially for larger values of h. Here, we see more variables related to Brazil's money supply. Note that the monetary base plays an important role for $h = 1$ and also for $h = 6$. For $h = 12$, we see a strong presence of energy consumption, a predictor related to the real economy. However, this difference does not end up converting into better results.

In general, the types of predictors that were highlighted were: firstly, employment-related predictors, which is reasonable considering that the dependent series is an unemployment rate

series; then predictors related to the relationship between Brazil and abroad; and for greater values of h, predictors related to fuels, a theme of the Brazilian real economy. Furthermore, combining the analysis of the chosen variables and Figure 10, we can see that for small values of h the difference between the models is not large. Models tend to choose similar predictors and obtain similar results.

Looking at the existing literature on unemployment forecasting methods, mainly the part of the literature that compares linear and non-linear models, we can separate them into two groups. There are articles that point to a better performance of non-linear models for small horizons and a better performance of linear models for larger horizons, as in the case of Chakraborty et al. (2021); and there are also articles that show the opposite, a better performance of linear models for short horizons and a better performance of nonlinear models for longer horizons, such as Katris (2020). The exercise made in this project points more towards the second group of articles. Similar to Katris (2020), the Linear model had good results in short horizons (one to three months ahead), and for longer horizons, the best accuracy goes to the non-linear model, represented here by the Mixed model.

7 CONCLUDING REMARKS

In this work we proposed three different multivariate models to predict the unemployment rate series in the Metropolitan Region of São Paulo. We had two goals with these models: (i) to verify if they were good forecasting instruments compared to a benchmark already established in the literature and (ii) to understand how the different configurations between the models affected the forecasting performance.

As reported, the three models consisted of boosting applications with different configurations on the weak learners that make up the algorithm. The Linear model, with all linear learners; the Non-linear model, with all learners using the p-splines model; and the Mixed model, with the price-related learners being of the p-spline format and the rest of the predictors being of the linear format. The benchmark used to test the general validity of the models consisted of an SARIMA algorithm with automatic parameter selection. The motivation of the models aroused from the discussion on the non-linearity in the prediction of macroeconomic time series. The presence of the Linear and Non-linear models aimed to directly compare these two types of modeling, whereas the Mixed model is motivated by a more detailed analysis. In the Mixed model, by parsimony we started from linear learners, but with information on the literature on the non-linear relationship between prices and employment (the Phillips curve) we exclusively modified the price related learners for p-splines. The motivation of the Mixed model is linked to the use of non-linear models with attention to the structure of the phenomenon being predicted, thus avoiding large parameterization, but taking into account the studies of the subject (in this case, unemployment).

Regarding the first objective, the work concluded, using performance metrics, that the proposed boosting models were superior to the SARIMA benchmark for the great majority of forecast horizons tested: for two to eleven months ahead. The superiority of accuracy had statistical significance via the Giacomini & White test in the intermediate forecast horizons, for $h = 2$ to $h = 10$. The dominance of the proposed models reached its maximum at $h = 4$, when the RMSFE metric for the proposed models was less than half of the benchmark. Subsequently, for higher values of h the distance between SARIMA and the proposed models accuracy decreased.

Regarding the second objective, there was no model of absolute prominence among the three proposals, but the Mixed model overall had the best results when looking at the performance metrics. The forecast results for horizons 1 to 4 are similar between the three models, with a slightly better accuracy of the Linear model, but from $h = 5$ trough $h = 12$ it was possible to

notice the greater predictive capacity of the Mixed model. The Giacomini & White test showed better accuracy of the Mixed model in relation to the Linear model from $h = 6$ to $h = 12$ and in relation to the Non-linear model from $h = 8$ to $h = 12$. After the Mixed model, in general the Linear model had the best performance, and lastly the Non-Linear model.

One result that we can draw from the exercise performed is the importance of using non-linear methods with knowledge about the series being predicted. Although the most current literature points to a better performance of non-linear models in relation to linear ones, an application without knowledge of the data can point to the other side, with simpler models outperforming more complex instruments. This seems to have been the case in the project. When all learners were designated as p-splines, there was a decrease in accuracy compared to linear learners. However, when only part of these were changed using empirically found relationships, there was improvement. The Mixed model managed to combine well the linear and non-linear characteristics of the series in question and, therefore, presented better results.

It is important to remember that the conclusions here are preliminary, in the sense that a stronger argument demands additional exercises that call into question linear and non-linear methods in the forecast of macroeconomic series. Also, the project opens space for other questions to be asked when forecasting macroeconomic variables in Brazil. Here, we used only the unemployment rate for the Metropolitan Region of São Paulo. An analysis could be done for other variables in other regions of the country. It is possible that the Metropolitan Region of São Paulo has sufficiently different characteristics from other regions with a lower concentration of economic activity, motivating also an exercise for the less central regions of the country. In addition, in technical terms, we can also do exercises with other characteristics. Tree-based boosting algorithms have lately gained prominence for time series forecasting, so there is room for studies to test them as predictor models of macroeconomic time series in Brazil.

REFERENCES

BREIMAN, Leo. **Bias, variance, and arcing classifiers**. 1996.

BUCHEN, Teresa; WOHLRABE, Klaus. Forecasting with many predictors: Is boosting a viable alternative? **Economics Letters**, Elsevier, v. 113, n. 1, p. 16–18, 2011.

BUEHLMANN, Peter. Boosting for high-dimensional linear models. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 34, n. 2, p. 559–583, 2006.

BÜHLMANN, Peter; HOTHORN, Torsten. Boosting Algorithms: Regularization, Prediction and Model Fitting. **Statistical Science**, Institute of Mathematical Statistics, v. 22, n. 4, p. 477–505, 2007. DOI: 10.1214/07-STS242. Available from: https://doi.org/10.1214/07-STS242.

BÜHLMANN, Peter; YU, Bin. Boosting with the L 2 loss: regression and classification. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 462, p. 324–339, 2003.

BYRNE, David; ZEKAITE, Zivile. Non-linearity in the wage Phillips curve: Euro area analysis. **Economics Letters**, v. 186, p. 108521, 2020. ISSN 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2019.07.006. Available from: https://www.sciencedirect.com/science/article/pii/S0165176519302460.

CANOVA, Fabio; HANSEN, Bruce E. Are seasonal patterns constant over time? A test for seasonal stability. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 13, n. 3, p. 237–252, 1995.

CHAKRABORTY, Tanujit et al. Unemployment rate forecasting: A hybrid approach. **Computational Economics**, Springer, v. 57, n. 1, p. 183–201, 2021.

CLAVERIA, Oscar. Forecasting the unemployment rate using the degree of agreement in consumer unemployment expectations. **Journal for Labour Market Research**, Springer, v. 53, n. 1, p. 1–10, 2019.

D'AMURI, Francesco; MARCUCCI, Juri. The predictive power of Google searches in forecasting US unemployment. **International Journal of Forecasting**, Elsevier, v. 33, n. 4, p. 801–816, 2017.

DEBELLE, Guy; LAXTON, Douglas. Is the Phillips curve really a curve? Some evidence for Canada, the United Kingdom, and the United States. **Staff Papers**, Springer, v. 44, n. 2, p. 249–282, 1997.

DIEBOLD, Francis X; MARIANO, Robert S. Comparing predictive accuracy. **Journal of Business & economic statistics**, Taylor & Francis, v. 20, n. 1, p. 134–144, 2002.

DUMIČIĆ, Ksenija; ČEH ČASNI, A; ŽMUK, Berislav. Forecasting unemployment rate in selected European countries using smoothing methods. **World Academy of Science, Engineering and Technology: International Journal of Social, Education, Economics and Management Engineering**, v. 9, n. 4, p. 867–872, 2015.

EILERS, Paul HC; MARX, Brian D. Flexible smoothing with B-splines and penalties. **Statistical science**, Institute of Mathematical Statistics, v. 11, n. 2, p. 89–121, 1996.

FREUND, Yoav. Boosting a weak learning algorithm by majority. **Information and computation**, Elsevier, v. 121, n. 2, p. 256–285, 1995.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). **The annals of statistics**, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000.

FRIEDMAN, Jerome H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001. DOI: 10.1214/aos/1013203451. Available from: https://doi.org/10.1214/aos/1013203451.

GHYSELS, Eric; SANTA-CLARA, Pedro; VALKANOV, Rossen. Predicting volatility: getting the most out of return data sampled at different frequencies. **Journal of Econometrics**, Elsevier, v. 131, n. 1-2, p. 59–95, 2006.

GIACOMINI, Raffaella; WHITE, Halbert. Tests of conditional predictive ability. **Econometrica**, Wiley Online Library, v. 74, n. 6, p. 1545–1578, 2006.

HOTHORN, Torsten et al. **mboost: Model-Based Boosting**. 2021. R package version 2.9-5. Available from: https://CRAN.R-project.org/package=mboost.

HYNDMAN, Rob J; KHANDAKAR, Yeasmin. Automatic time series forecasting: the forecast package for R. **Journal of statistical software**, v. 27, p. 1–22, 2008.

KATRIS, Christos. Prediction of unemployment rates with time series and machine learning techniques. **Computational Economics**, Springer, v. 55, n. 2, p. 673–706, 2020.

KAUPPI, Heikki; VIRTANEN, Timo. Boosting nonlinear predictability of macroeconomic time series. **International Journal of Forecasting**, Elsevier, v. 37, n. 1, p. 151–170, 2021.

KEARNS, M.; VALIANT, L. G. Crytographic Limitations on Learning Boolean Formulae and Finite Automata. In: PROCEEDINGS of the Twenty-First Annual ACM Symposium on Theory of Computing. Seattle, Washington, USA: Association for Computing Machinery, 1989. (STOC '89), p. 433–444. ISBN 0897913078. DOI: 10.1145/73007.73049. Available from: https://doi.org/10.1145/73007.73049.

KUEHN, Tobias; STOECKER, Almond. **varimp: Model-Based Boosting**. 2021. R package version 2.9-5. Available from: https://search.r-project.org/CRAN/refmans/mboost/html/varimp.html.

KWIATKOWSKI, Denis et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? **Journal of econometrics**, Elsevier, v. 54, n. 1-3, p. 159–178, 1992.

LEHMANN, Robert; WOHLRABE, Klaus. Boosting and regional economic forecasting: the case of Germany. **Letters in Spatial and Resource Sciences**, Springer, v. 10, n. 2, p. 161–175, 2017.

LINDENMEYER, Guilherme; SKORIN, Pedro Pablo; TORRENT, Hudson da Silva. Using boosting for forecasting electric energy consumption during a recession: a case study for the Brazilian State Rio Grande do Sul. **Letters in Spatial and Resource Sciences**, Springer, p. 1–18, 2021.

MAAS, Benedikt. Short-term forecasting of the US unemployment rate. **Journal of Forecasting**, Wiley Online Library, v. 39, n. 3, p. 394–411, 2020.

MCCRACKEN, Michael W; NG, Serena. FRED-MD: A monthly database for macroeconomic research. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 34, n. 4, p. 574–589, 2016.

MEDEIROS, Marcelo C et al. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 39, n. 1, p. 98–119, 2021.

RUPPERT, David; WAND, Matt P; CARROLL, Raymond J. **Semiparametric regression**. Cambridge university press, 2003.

SCHAPIRE, Robert E. The strength of weak learnability. **Machine learning**, Springer, v. 5, n. 2, p. 197–227, 1990.

SHAFIK, Nivien; TUTZ, Gerhard. Boosting nonlinear additive autoregressive time series. **Computational Statistics & Data Analysis**, Elsevier, v. 53, n. 7, p. 2453–2464, 2009.

SILVERMAN, Bernhard W. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 47, n. 1, p. 1–21, 1985.

WOHLRABE, Klaus; BUCHEN, Teresa. Assessing the macroeconomic forecasting performance of boosting: evidence for the United States, the Euro area and Germany. **Journal of Forecasting**, Wiley Online Library, v. 33, n. 4, p. 231–242, 2014.

XU, Qifa et al. The Phillips curve in the US: A nonlinear quantile regression approach. **Economic Modelling**, Elsevier, v. 49, p. 186–197, 2015.

ZENG, Jing. Forecasting aggregates with disaggregate variables: Does boosting help to select the Most relevant predictors? **Journal of Forecasting**, Wiley Online Library, v. 36, n. 1, p. 74–90, 2017.

ZHOU, Zhi-Hua. **Ensemble methods: foundations and algorithms**. Chapman and Hall/CRC, 2019.

APPENDIX A – DATASET SOURCES AND PREDICTORS TRANSFORMATIONS

Table 12 – Dataset sources and predictors transformations

Name	Transformation	Source
Consumption and Sales Theme		
SPC - number of queries	TL1	ACSP/IEGV
Usecheque - number of queries	TL1	ACSP/IEGV
Apparent consumption - fuel alcohol - average - daily amount	TL1	ANP
Apparent consumption - petroleum derivatives - average - daily amount	TL1	ANP
Apparent consumption - gasoline - average - daily amount	TL1	ANP
Apparent consumption - LPG gas - average - daily amount	TL1	ANP
Apparent consumption - petroleum derivatives - others - average - daily amount	TL1	ANP
Apparent consumption - fuel oil - average - others - average - daily amount	TL0	ANP
Apparent consumption - diesel oil - average - daily amount	TL1	ANP
Real revenue - industry - index (2006 average = 100)	TL1	CNI
Real revenue - industry - deseasonalized index (2006 average = 100)	TL1	CNI
Consumption - electricity - Midwest Region (CO) - quantity	TL1	Eletrobras
Consumption - electricity - trade - quantity	TL1	Eletrobras
Consumption - electricity - industry - quantity	TL1	Eletrobras
Consumption - electricity - Northeast Region (NE) - quantity	TL1	Eletrobras
Consumption - electricity - North Region (N) - quantity	TL1	Eletrobras
Consumption - electricity - other sectors - quantity	TL1	Eletrobras
Consumption - electricity - residential - quantity	TL1	Eletrobras
Consumption - electricity - Southeast Region (SE) - quantity	TL1	Eletrobras
Consumption - electricity - South Region (S) - quantity	TL1	Eletrobras
Consumption - electricity - quantity	TL1	Eletrobras
Consumption - electricity - trade - average tariff per MWh	TL1	Eletrobras
Consumption - electricity - industry - average tariff per MWh	TL1	Eletrobras
Consumption - electricity - residential - average tariff per MWh	TL1	Eletrobras
Consumption - electricity - average tariff per MWh	TL1	Eletrobras
Current Economic Conditions Index (ICEA)	TL1	Fecomercio SP
Vehicle sales by dealerships - automobiles	TL1	Fenabrave
Vehicle sales by dealerships - total	TL1	Fenabrave
Currency and Credit		
Credit operations - balance of the credit portfolio - Total	TL2	Bacen
Money Aggregate - Expanded - M2 - Savings Deposits - End of Period	TL2	Bacen
Money Aggregate - Restricted (M1) - demand deposits - average	TL1	Bacen
Monetary base conditioning factors - BC rediscount operations	T0	Bacen
Monetary base conditioning factors - deposits from financial institutions	T0	Bacen
Monetary base conditioning factors - external sector operations	T0	Bacen
Monetary base conditioning factors - operations with federal public securities	T0	Bacen
Monetary base conditioning factors - National Treasury	T1	Bacen
Monetary base conditioning factors - monetary base variation	T1	Bacen
Monetary Base - Restricted (M0) - average	TL1	Bacen
Restricted Monetary Base (M0) - % of GDP - end of period	TL1	Bacen
Money Aggregate - Restricted (M1) - average	TL1	Bacen
Money Aggregate - Expanded - M3 - oper. committed with federal bonds - end of period	TLK1	Bacen
Money Aggregate - Expanded - M3 - oper. pledged with title. fed. - % of GDP - end of period	TLK1	Bacen

Table 12 continued from previous page

Name	Transformation	Source
Monetary Base - Restricted (M0) - currency issued - average	TL1	Bacen
Money Aggregate - Restricted (M1) - paper money held by the public - average	TL1	Bacen
Money Aggregate - Restricted (M0) - bank reserves - average	TL1	Bacen
Foreign Trade		
Imports - prices - index (2006 average = 100)	TL1	Funcex
Imports - quantum - index (2006 average = 100)	TL1	Funcex
Terms of trade - index (2006 average = 100)	TL1	Funcex
Exports - basic products - prices - index (2006 average = 100)	TL1	Funcex
Exports - manufactured products - prices - index (2006 average = 100)	TL1	Funcex
Exports - semi-manufactured products - prices - index (2006 average = 100)	TL1	Funcex
Exports - prices - index (2006 average = 100)	TL1	Funcex
Exports - basic products - quantum - index (2006 average = 100)	TL1	Funcex
Exports - manufactured products - quantum - index (2006 average = 100)	TL1	Funcex
Exports - semi-manufactured products - quantum - index (2006 average = 100)	TL1	Funcex
Exports - quantum - index (2006 average = 100)	TL1	Funcex
Exports - profitability - index (Dec. 2003 = 100)	TL1	Funcex
Imports - (FOB)	TL1	MDIC
Exports - Gasoline	TL1	MDIC
Exports - aggregate factor - basic products - (FOB)	TL1	MDIC
Exports - aggregate factor - industrialized products - (FOB)	TL1	MDIC
Exports - aggregate factor - manufactured products - (FOB)	TL1	MDIC
Exports - aggregate factor - semi-manufactured products - (FOB)	TL1	MDIC
Exports - (FOB)	TL1	MDIC
Exports - aggregate factor - special transactions - (FOB)	TL1	MDIC
Exports (Kg) - Gasoline	TL1	MDIC
Exchange Rates		
Contracted exchange - financial - purchase	TL1	Bacen
Exchange rate - R/US - commercial - purchase - average	TL1	Bacen
Exchange rate - R/US - commercial - sale - average	TL1	Bacen
Contracted exchange - commercial - import	TL1	Bacen
Contracted exchange - financial	T0	Bacen
Contracted exchange - financial - sale	TL1	Bacen
Purchasing power parity (PPP) rate - household consumption	TL1	IPEA
National Accounts		
GDP	TL2	Bacen
GDP - accumulated 12 months	TL1	Bacen
Apparent consumption - consumer goods - index (2012 average = 100)	TL1	IPEA
Apparent Consumption - Consumer Goods - desase Index. (2012 average = 100)	TL1	IPEA
Apparent consumption - semi-durable and non-durable consumer goods - desase Index. (average 2002 = 100)	TL1	IPEA
Apparent consumption - intermediate goods - index (2012 average = 100)	TL1	IPEA
Apparent consumption - capital goods - index (2012 average = 100)	TL1	IPEA
Prices		
Prices - IPCA - free prices - tradable - var.	TR1	Bacen
Prices - IPCA - core by exclusion - ex1 - var	T0	Bacen
Prices - IPCA - free prices - not tradable - var.	TR2	Bacen
Prices - IPCA - free prices - var.	T0	Bacen
Prices - IPCA - free prices - durable goods - var	T0	Bacen

Table 12 continued from previous page

Name	Transformation	Source
Prices - IPCA - free prices - non-durable goods - var	T0	Bacen
Prices - IPCA - free prices - semi durable goods - var	TR1	Bacen
Prices - IPCA - free prices - services - var	TR1	Bacen
Prices - IPCA - monitored prices - var.	T0	Bacen
Double-weighted core - IPCA	T0	Bacen
Prices - IPC (FIPE)	T0	Fipe
Prices - IPC - 1st quadsweek (FIPE)	T0	Fipe
Prices - IPC -2nd quadsweek (FIPE)	T0	Fipe
Prices - IPC -3rd quadsweek (FIPE)	T0	Fipe
Prices - CPI - general - index (June 1994 = 100) - MRSP	TL1	Fipe
Prices - IGP-10 - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IGP-DI - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IGP-DI	T0	FGV
Prices - IGP-DI - general - centered - end of period - index (Aug. 1994 = 100)	TL1	FGV
Prices - IGP-M - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IGP-M	T0	FGV
Prices - IGP-M - 1st ten-day period	T0	FGV
Prices - IGP-M - 2nd ten-day period	T0	FGV
Prices - IGP-OG - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IGP-OG	T0	FGV
Prices - INCC-10 - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - INCC-DI - general - index (Aug. 1994 = 100)	TL2	FGV
Prices - INCC-DI	T1	FGV
Prices - INCC-M	T0	FGV
Prices - INCC-M - 1st ten-day period	T0	FGV
Prices - INCC-M - 2nd ten-day period	T0	FGV
Prices - IPA-10 - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPA-DI - origin - prod. livestock - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPA-DI - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPA-DI	T0	FGV
Prices - IPA-DI - origin - prod. industrial - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPA-M	T0	FGV
Prices - IPA-M - 1st ten-day period	T0	FGV
Prices - IPA-M - 2nd ten-year period	T0	FGV
Prices - IPC-10 - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPC-DI - general - index (Aug. 1994 = 100)	TL1	FGV
Prices - IPC-DI (FGV)	T0	FGV
Prices - IPC-M	T0	FGV
Prices - IPC-M - 1st ten-day period	T0	FGV
Prices - IPC-M - 2nd ten-day period	T0	FGV
Prices - INPC - general - index (Dec. 1993 = 100)	TL1	IBGE/SNIPC
Prices - INPC - food and beverages - var.	T0	IBGE/SNIPC
Prices - INPC -residence articles - var.	T0	IBGE/SNIPC
Prices -INPC - general	T0	IBGE/SNIPC
Prices - INPC - personal expenses - var.	T0	IBGE/SNIPC
Prices - INPC - housing - var.	T0	IBGE/SNIPC
Prices - INPC - health and personal care - var.	T0	IBGE/SNIPC

Table 12 continued from previous page

Name	Transformation	Source
Prices - INPC -transportation - var.	T0	IBGE/SNIPC
Prices - INPC - clothing - var.	TR1	IBGE/SNIPC
Prices - IPCA - general - index (Dec. 1993 = 100)	TL1	IBGE/SNIPC
Prices - IPCA - food and beverages - var.	T0	IBGE/SNIPC
Prices - IPCA - articles of residence - var.	T0	IBGE/SNIPC
Prices - IPCA - personal expenses - var.	T0	IBGE/SNIPC
Prices - IPCA - general	T0	IBGE/SNIPC
Prices - IPCA - housing - var.	T0	IBGE/SNIPC
Prices - IPCA - health and personal care - var.	T0	IBGE/SNIPC
Prices - IPCA - MRSP - var.	T0	IBGE/SNIPC
Prices - IPCA - transport - var.	T0	IBGE/SNIPC
Prices - IPCA - clothing - var.	TR1	IBGE/SNIPC
Employment		
Hours worked - industry - index (2006 average = 100)	TL1	CNI
Hours worked - industry - deseasonalized index. (2006 average = 100)	TL1	CNI
Personnel employed - industry - index (2006 average = 100)	TL1	CNI
Employed personnel - industry - deseasonalized index. (2006 average = 100)	TL1	CNI
Unemployment rate - MRSP	TL1	SEADE
Wages and Income		
Minimum wage - purchasing power parity (PPP)	TL1	IPEA
Real minimum wage	TL1	IPEA
Minimum wage	TL1	MTE
Perception and Expectations		
Consumer Confidence Index (ICC)	TL1	Fecomercio SP
Expectations Index (IEC)	TL1	Fecomercio SP
Financial Accounts		
Stock index - Ibovespa - closing	T0	Anbima
Savings - nominal income - 1st business day (until 05.03.2012)	T0	Anbima
Interest rate - CDI / Over - accumulated in the month	T0	Bacen
Interest rate - Over / Selic - accumulated in the month	T0	Bacen
Savings account (total) - Balances	TL1	Bacen

Source: Elaborated by the author (2022)