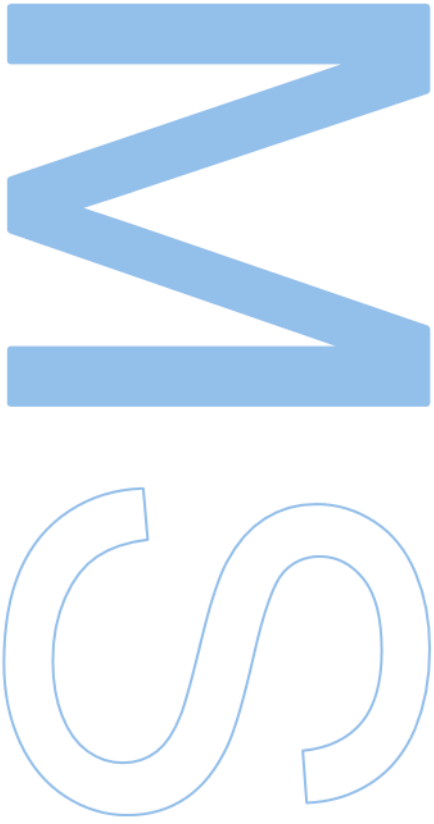
# Forecasting with Machine Learning methods: a case study with unemployment



Guilherme de Abreu Jeremias
Mestrado em Estatística Computacional e Análise de Dados
Departamento de Matemática
2023

**Orientador**
Maria Eduarda Silva, Professora Associada, Faculdade de Economia da Universidade do Porto




U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

# Agradecimentos

Findada mais uma etapa do meu percurso académico, marcada por muito trabalho e sacrifício, mas também por muito empenho e vontade de continuar a aprender e progredir, resta-me agradecer a todos aqueles que foram fundamentais para a sua conclusão:

Agradeço à Professora Maria Eduarda, por sempre me ter guiado de forma sábia, assertiva e próxima. Enalteço também a sua constante disponibilidade para me auxiliar em todas as questões e dúvidas, as sugestões sempre pertinentes e apoio infindável. Obrigado!

Agradeço-te também a ti Eduardo, por teres assumido o papel de “co-orientador”, e por todo o teu trabalho árduo, auxílio e comprometimento. Em particular, obrigado por me teres acompanhado de perto, incentivando-me na procura de fazer sempre melhor.

O meu obrigado também à Professora Margarida Brito e Professora Rita Gaio, pelo cuidado, compreensão e preocupação que sempre demonstraram para comigo no decorrer deste Mestrado.

À Andreia, que todos os dias me tranquiliza e incentiva a alcançar mais, e que durante todo este processo, tanto me deu e tão pouco pediu para si. Obrigado por fazeres com que as nossas vidas resultem: para mim, para ti, e, acima de tudo, para nós.

Aos meus pais, obrigado pelo vosso apoio incondicional, e por todos os dias me relembrarem que a vida é feita de um balanço muito ténue entre trabalho árduo e regozijo.

Por fim, agradeço à minha família e amigos, pelo vosso incentivo e apoio, e por estarem sempre presentes para mim. Em particular, obrigado a ti Simão, por juntos termos embarcado nesta “montanha-russa” que é fazer um Doutoramento e um Mestrado ao mesmo tempo.

# Resumo

A previsão de séries temporais tem como objetivo estimar os valores futuros de uma série temporal tendo por base os seus valores históricos. Diversos modelos foram desenvolvidos e utilizados para este propósito, contudo, em tempos mais recentes, os modelos baseados em aprendizagem de máquina têm ganho cada vez mais relevância no contexto de previsão de séries temporais. No entanto, persistem ainda inúmeros desafios sobre a sua aplicação, no qual se inclui a avaliação da sua performance em comparação com modelos clássicos.

O desemprego é um tema fulcral em Economia, existindo um objetivo generalizado de assegurar previsões precisas sobre a taxa de desemprego. Nos anos mais recentes, as pesquisas na internet têm sido cada vez mais utlizadas neste contexto, uma vez que foi já demonstrado que estes dados possuem um poder preditivo assinalável.

Tendo em vista o exposto, o objetivo principal desta dissertação era o de avaliar a utilidade das pesquisas no Google e algoritmos de aprendizagem de máquina para melhorar a previsão da taxa de desemprego de Portugal. Para dar reposta a este objetivo, diferentes abordagens de modelação foram seguidas ao incorporar séries temporais do “Google Trends”, nomeadamente através do desenvolvimento de um modelo de referência “ARIMA”, modelos de regressão com erros “ARIMA” e vários modelos de aprendizagem de máquina (todos baseados no algoritmo floresta aleatória).

De forma geral, os resultados obtidos demonstraram que um melhor desempenho em termos de previsão foi alcançado pela utilização de modelos de aprendizagem de máquina, por comparação com os modelos clássicos. Também importante, a melhor capacidade demonstrada por estes modelos pareceu ser largamente influenciada pelo pré-processamento e transformações das séries temporais consideradas na análise. Mais ainda, este trabalho permitiu confirmar que as pesquisas no Google possuem um poder preditivo elevado, demonstrando assim a importância destes dados para a previsão da taxa de desemprego. Respetivamente, a conclusão principal é a de que a combinação de algoritmos de aprendizagem de máquina e pesquisas no Google aparenta ser uma abordagem útil e promissora para a previsão de variáveis económicas.

Palavras-chave: séries temporais, previsão, taxa de desemprego, Google Trends, aprendizagem de máquina

# Abstract

Forecasting consists on predicting the future values of a time series by considering its historical data. Numerous models have been developed and utilized for this purpose but, more recently, machine learning-based models have been moving to the forefront of time series forecasting. Nevertheless, numerous challenges and missing gaps still exist regarding their application in such context, especially concerning the evaluation of their forecasting performance in comparison to traditional models.

Unemployment is a key topic in economics, and accurately forecasting the unemployment rate has always been a major goal. In recent years, internet searches have been increasingly considered in this context, since it has now been demonstrated that this data holds great predictive power.

Taking the above mentioned into account, the main objective of this thesis was to evaluate the utility of Google searches and machine learning algorithms for nowcasting the unemployment rate of Portugal. To tackle this goal, different modelling approaches were followed by incorporating several Google Trends time series, namely concerning the development of a benchmark ARIMA model, regression with ARIMA errors models and various machine learning models (all based on the random forest algorithm).

Overall, results demonstrated that a better nowcasting accuracy could be achieved by using machine learning-based models, over the more traditional models. Also important, the greater predictive capacity of such models seemed to be largely influenced by the pre-processing and feature engineering of the time series considered in the analysis. In addition, it was confirmed that Google searches hold predictive power, thus highlighting the relevance of such data source for unemployment rate forecasting. Respectively, the main conclusion of this work is that the combination of machine learning algorithms and Google searches is a suitable and promising framework for forecasting economic variables.

Keywords: time series, forecasting, unemployment rate, Google Trends, machine learning

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| ACF | Autocorrelation Function |
| AI | Artificial Intelligence |
| AIC | Akaike Information Criterion |
| AR | Autoregressive |
| ARIMA | Autoregressive Integrated Moving Average |
| ARMA | Autoregressive Moving Average |
| dGT | Deseasonalized Google Trends |
| GT | Google Trends |
| GVAR | Global Vector Autoregressive |
| MA | Moving Average |
| ML | Machine Learning |
| PACF | Partial Autocorrelation Function |
| regARIMA | Regression with ARIMA Errors |
| RF | Random Forest |
| RMSE | Root Mean Squared Error |
| SVI | Search Volume Index |

# 1. Introduction

Forecasting widely regards as one of the most important applications of time series analysis. This field has gained relevance in the most different scientific areas, such as economics, engineering, medicine and environmental sciences (Tiao and Box 1981; Enders 2015). Forecasting plays a crucial role in numerous industries by being an essential tool for solving practical business challenges and driving better decision-making (Kirchgässner et al. 2012). Specifically, the primary objective of forecasting is to accurately predict the future values of a time series considering historical data. However, this is not a simple task due to the unique and complex statistical properties of time series, which constrains the overall goal of providing reliable and efficient future estimates (Hyndman and Athanasopoulos 2018). Also, the forecasting field has drifted towards nowcasting, a concept related to predicting the present, very near future and recent past. Accordingly, nowcasting has been gaining popularity in economics, as timely predictions are often crucial for different applications within the field (Hyndman and Athanasopoulos 2018; Kapetanios and Papailias 2018).

Numerous mathematical models have been proposed for time series modelling and forecasting. Among these, a well-known and most commonly applied is the classical Autoregressive Integrated Moving Average (ARIMA) model (Hyndman and Athanasopoulos 2018). However, numerous academic studies reported the poor predictive performance of such ARIMA model in many real-world problems, mainly attributed to its lack of capacity to capture nonlinear patterns or relationships in the data (Makridakis et al. 2018). Several other mathematical models have been developed or adapted for time series analysis in recent years in order to tackle this issue. Among these, machine learning-based models have been moving to the forefront of time series analysis, mainly because they have shown solid empirical performance in forecasting exercises (Bojer 2022; Cerqueira et al. 2022).

In the Machine Learning (ML) context, Random Forest (RF) is one of the most popular algorithms, described as a robust and powerful method capable of providing accurate and reliable estimates (Breiman 2001; Cutler et al. 2012). This algorithm consistently showed good forecasting performance in many application fields, specifically economic sciences. Respectively, RF-based models showed utility in forecasting stock index movements, individual stock prices, unemployment and Gross Domestic Product, among other economic data (Gogas et al. 2021; Yoon 2021; Abraham et al. 2022). Nevertheless, numerous challenges and missing gaps exist regarding the application of ML algorithms in forecasting exercises. Namely, there is a lack of studies

comparing the forecasting performance of ML-based models with traditional ones, thus making it very difficult to understand when these advanced algorithms outperform classical time series models (Makridakis et al. 2018; Bojer 2022). This fact makes clear the need to further study the performance of ML algorithms under different forecasting scenarios and compare their predictive performance with classical models (Makridakis et al. 2018).

Unemployment is a central topic in economics and an issue of great interest for entire societies and nations. Due to its importance, monitoring the unemployment rate and forecasting its future values on time has always been a major goal for academics, businesses and policy-makers (Dritsakis and Klazoglou 2018). This critical task relies on the data gathered and stored by government agencies and central banks and annual reports of private and public companies (Jelena et al. 2017). However, announcements of such data typically refer to a delayed scenario, thus constraining the provision of accurate forecasts. On the other hand, the past decades have seen a rapid growth in the ability of computing systems to collect and transport vast amounts of data through a phenomenon often referred to as “Big Data” (Liu et al. 2016; Jun et al. 2018), which can provide quick predictions to unemployment figures. An example regards the data originating from internet searches, accessed at nearly zero cost and virtually instantaneously. Respectively, internet search engines are utilized by most of the world’s population daily, and Google search is the most popular among internet users (Haider and Sundin 2019).

Academic studies indicate that the data from the Google search engine is valuable for time series analysis as containing predictive power that can significantly improve the accuracy of forecasts (Artola et al. 2015; D’Amuri and Marcucci 2017; Mavragani and Gkillas 2020). Additionally, studies focusing on unemployment forecasting demonstrate that models incorporating Google searches outperform models that do not employ such data (Barreira et al. 2013; D’Amuri and Marcucci 2017; Naccarato et al. 2018). Some exceptions to these findings are also found, thus showing that more research is required (Barreira et al. 2013; D’Amuri and Marcucci 2017). In recent years, increased interest has been in combining Google search data with more advanced time series models, including ML-based algorithms. While there is already evidence that this combination can be a powerful and promising strategy to nowcast economic variables, this research field remains in its infancy (Kreiner and Duca 2020; Gogas et al. 2021).

This thesis aims to explore the contribution of Google search data and adoption of ML algorithms in the nowcasting of the Portuguese unemployment rate. In addition, it

is a specific objective of this work to compare the forecasting performance of RF-based models with the one of more classical ones, including univariate benchmark ARIMA and regression with ARIMA errors models. As previously mentioned, this theme of research is relatively recent, thus this work provides relevant information towards clarifying the importance of ML-based models in forecasting exercises. Also, this objective tackles one of the major data gaps of the field, i.e. the lack of comparison of forecasting performance between classical and ML-based models.

The remainder of this thesis is divided into five additional chapters. Chapter 2 exposes the fundamental concepts of time series, ML and forecasting; Chapter 3 reviews the literature regarding unemployment forecasting and the utilization of Google searches in this context; Chapter 4 presents in detail the objectives of this thesis and the methodology to accomplish them; Chapter 5 provides the main experimental results obtained; finally, Chapter 6 focuses on the discussion of the findings, the establishment of conclusions and recommendations for future research.

# 2. Fundamental concepts

In simple terms, a time series describes a sequence of observations arranged chronologically, which may assume different frequencies, such as daily, monthly and yearly. Time series can be labelled as discrete or continuous, respectively, when observations are recorded at specific times or continuously through time (Shumway and Stoffer 2017). The primary objective of time series analysis is to develop mathematical models that render plausible descriptions for the sampled data by exploring the nature and behaviour of the series under study (Shumway and Stoffer 2017). Various statistical methodologies and mathematical models are utilized for time series forecasting, ranging from classical and simpler methods, such as the naïve method that only uses the most recent observation as a forecast, to more recent and highly complex models, such as ML algorithms (Hyndman and Athanasopoulos 2018).

Over the last few years, Artificial Intelligence (AI) gained increasing attention from the public due to the development of high-profile innovations and applications, including self-driving vehicles, intelligent robots and toys, personalized medical care, image and speech recognition and automatic translations (Makridakis 2017). AI's success relies on algorithms that learn by trial and error, improving their performance over time (Makridakis 2017; Joiner 2018). In this regard, the sub-field of ML has emerged within AI as the method of choice for several of its core developments and best well-known applications. ML algorithms have been increasingly considered for time series analysis in scientific and business domains, especially for forecasting purposes (Krollner et al. 2010; Makridakis et al. 2018). Accordingly, some of the most relevant applications of ML algorithms within the sub-field of forecasting include their use to predict financial and macroeconomic time series, the direction of the stock market, and the supply and demand for products (Krollner et al. 2010; Parray et al. 2020). Above all, ML algorithms gained popularity in this context because they showed solid empirical performance, thus establishing themselves as serious contenders to classical time series statistical models (Bojer 2022; Cerqueira et al. 2022).

This chapter aims to present the key concepts related to the main issues addressed throughout this thesis, namely concerning time series, ML and forecasting. In specific, this chapter provides a description of the ARIMA and regression with ARIMA errors models, as well as the RF algorithm.

## 2.1. Time series

On a formal basis, a time series can be defined as a collection of random variables indexed according to the order they are obtained in time, i.e. considering $x_1$[1], $x_2$, $x_3$,…, where the random variable $x_1$ refers to the value taken by the series at the first time point, the variable $x_2$ denotes the value taken at the second time point, $x_3$ the value at the third time point, and so forward (Shumway and Stoffer 2017). Furthermore, such a collection of random variables is usually referred to as a stochastic process, with the observed values being alluded to as realizations of the stochastic process (Metcalfe and Cowpertwait 2009). In this regard, three statistical components are of interest: i) the mean; ii) the autocovariance, which measures the linear dependence between two points observed at different times on the same series; and iii) the autocorrelation, which measures the linear predictability of the series at time $t$, say $x_t$, using only the value $x_s$ (Metcalfe and Cowpertwait 2009; Shumway and Stoffer 2017). These components are described (Shumway and Stoffer 2017) as follows.

Mean function: $\mu_t = E(x_t)$

Autocovariance function: $\gamma(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$

Autocorrelation function: $\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$

Another major concept in time series analysis is stationarity. This concept can be interpreted as a form of statistical equilibrium in the sense that the statistical properties (mean and variance) of a stationary time series do not depend upon time. Furthermore, the autocovariance function is only dependent on the lag, i.e. the number of intervals between two measurements (Shumway and Stoffer 2017). In simpler terms, stationarity implies that it does not matter when a time series is observed, as its properties and behaviour should look the same at any time (Hyndman and Athanasopoulos 2018). Respectively, the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of a stationary time series is absolutely essential towards choosing the best statistical model to represent the underlying data, including the selection of appropriate orders for the classical Autoregressive Moving Average and Autoregressive Integrated Moving Average models (Shumway and Stoffer 2017; Hyndman and Athanasopoulos 2018). The ACF, denoted as $\rho(h)$, and PACF, denoted

[1] Here, following the notation of Shumway and Stoffer (2017), $x_t$ is used for the random variable and the observed value

as $\emptyset_{hh}$ for $h = 1,2, \ldots,$ of a stationary time series can be described, respectively (Shumway and Stoffer 2017):

$$\rho(h) = \frac{\gamma(t+h,t)}{\gamma(t+h,t+h)\gamma(t,t)} = \frac{\gamma(h)}{\gamma(0)}$$

and

$$\emptyset_{11} = \text{corr}(x_{t+1}, x_t) = \rho(1)$$

$$\emptyset_{hh} = \text{corr}(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t),\ h \geq 2.$$

### 2.1.1. Time series components

Identifying the components of a time series is generally perceived as a key step towards understanding its underlying process and behaviour (Hyndman and Athanasopoulos 2018). This enables the selection of suitable mathematical models to represent the time series, which is important for obtaining forecasts (Metcalfe and Cowpertwait 2009). Accordingly, the essential components of a time series are the trend, seasonal, cyclical and irregular variations, as represented in Figure 1 (Enders 2015; DataVedas 2018; Hyndman and Athanasopoulos 2018).

A trend corresponds to a long-term movement in the time series, either increasing, decreasing or even stagnating. However, trends do not always present a linear form, and the series can present more than one trend over different periods (Tsay 2005; Hyndman and Athanasopoulos 2018).

Seasonality refers to the possible existence of a seasonal factor influencing the behaviour of the series. Seasonality always occurs at a fixed and known period, such as a given month in the year or day of the week, and its presence can be determined by different factors, including customs, traditional habits, climate and weather conditions, among many others (Hyndman and Athanasopoulos 2018). Also, it is worth highlighting that trend and seasonality influence stationarity since both affect the values observed at different points in time (Hyndman and Athanasopoulos 2018).

The cyclical component relates to the possible existence of cycles in the data. These cycles correspond to rises and falls in values that are not of a fixed frequency, usually occurring in the form of medium-term fluctuations that repeat throughout time (Hyndman and Athanasopoulos 2018). Particular circumstances generally determine time series cyclicality. Paradigmatic examples include economic and finance series that frequently resent some cyclical fluctuations corresponding to shifts in macroeconomic conditions and business cycles (Tsay 2005).

Irregular variations are also commonly considered to be time series components. These correspond to non-random sources of variations that mainly occur due to unpredictable circumstances, such as wars, earthquakes, revolutions and pandemics (Shumway and Stoffer 2017; Hyndman and Athanasopoulos 2018). Such variations do not show a constant pattern of variation (Hyndman and Athanasopoulos 2018).

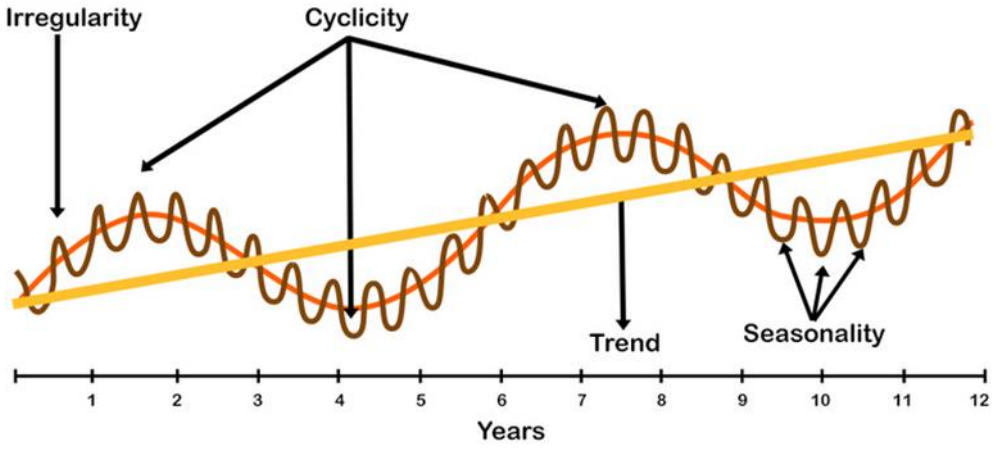



Figure 1. Representation of time series components. Source: DataVedas (2018).

### 2.1.2. ARIMA models

No other method has been more extensively applied for forecasting purposes than the ARIMA model. This method provides improved modelling capacity over the Autoregressive and Autoregressive Moving Average (ARMA) models, which were originally proposed to introduce correlation generated through lagged linear relations (Shumway and Stoffer 2017; Hyndman and Athanasopoulos 2018). Accordingly, a time series $\{x_t; t = 0, \pm1, \pm2, ...,\}$ is an ARMA$(p, q)$ model if it is stationary and satisfies:

$$x_t = \emptyset_1 x_{t-1} + \cdots + \emptyset_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \qquad (2.1)$$

with parameters $p$ and $q$ representing, respectively, the autoregressive and the moving average orders, and $w_t$ standing for white Gaussian noise, i.e. a sequence of independent and identically distributed random variables with zero mean and variance $\sigma_w^2 > 0$ (Shumway and Stoffer 2017). Also, when $q$ = 0, the model is named an Autoregressive (AR) model of order $p$ – AR$(p)$ – and when $p$ = 0, the model is called a Moving Average (MA) model of order $q$, thus MA$(q)$. The AR$(p)$ and MA$(q)$ processes satisfy the below equations, respectively:

$$x_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + w_t, \qquad (2.2)$$

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{2.3}$$

where the parameters $\emptyset_i$ and $\theta_i$ are such that the roots $z_1, \ldots, z_p$ of the polynomial $\emptyset(z) = 1 - a_1 z - \cdots - a_p z^p$ and $\theta(z) = 1 - b_1 z - \cdots - b_q z^q$ lie outside of the unit circle $|z_i > 1|$, in order to guarantee, respectively, stationary and invertibility (Shumway and Stoffer 2017). Accordingly, autoregressive models rely on the idea that the current value of the time series can be explained as a function of $p$ past values, so that $p$ determines the number of past steps required to forecast current values. On the other hand, moving average models assume that white noise $(w_t)$ is combined linearly to form the observed data (Shumway and Stoffer 2017). Additionally, ARMA, AR and MA models can be described with the use of the backshift operator, thus obtaining the concise representation, respectively (Shumway and Stoffer 2017):

$$\emptyset(B)x_t = \theta(B)w_t, \tag{2.4}$$

$$\emptyset(B)x_t = w_t, \tag{2.5}$$

$$x_t = \theta(B)w_t. \tag{2.6}$$

Many time series show a non-stationary behaviour, thus limiting the application of ARMA, AR and MA models. This prompted the development of non-stationary models, such as the ARIMA$(p, d, q)$ model, which accounts for stationarity by employing differencing to time series (Hyndman and Athanasopoulos 2018). As so, the order $d$ denotes the number of differences taken in the model, with a time series process $(x_t)$ being said to be an ARIMA$(p, d, q)$ if:

$$\nabla^d x_t = (1 - B)^d x_t \tag{2.7}$$

is an ARMA$(p, q)$ (Shumway and Stoffer 2017). The model can also be expressed as:

$$\emptyset(B)(1 - B)^d x_t = \theta(B)w_t. \tag{2.8}$$

Respectively, Shumway and Stoffer (2017) reviewed the most common framework for fitting ARIMA models to time series data, including plotting the data initially, transforming it when required, and inspecting the dependence orders, i.e. autoregressive, differencing and moving average orders. In particular, the number of differences required can be determined by successively plotting the differenced time series until the data becomes

stationary. Additionally, analysing the autocorrelation and partial autocorrelation plots can help to determine appropriate values for $p$ and $q$ (Shumway and Stoffer 2017; Hyndman and Athanasopoulos 2018). Then, the parameters of candidate models can be estimated. Besides, model selection should be based on information criteria, such as the Akaike Information Criterion (AIC), the corrected AIC for ARIMA models or Bayesian Information Criterion, with the best models being those that minimize such criteria (Shumway and Stoffer 2017; Hyndman and Athanasopoulos 2018).

#### 2.1.3. Regression with ARIMA errors model

While ARIMA models allow for the incorporation of relevant past information of a time series, they exclude the inclusion of other useful information, including external variables that may explain some of the historical variation and may lead to more accurate forecasts (Hyndman and Athanasopoulos 2018). One relevant model that tackles this issue is the regression with ARIMA errors. Respectively, Hyndman and Athanasopoulos (2018) comprehensively address the statistical properties and forecasting-related aspects of these models. Following their notation in this context, one can consider a regression model in the form of:

$$\mathcal{Y}_t = B_0 + B_1 x_{1,t} + \cdots + B_k x_{k,t} + \varepsilon_t, \tag{2.9}$$

where $y_t$ is a linear function of the $k$ predictor variables $(x_{1,t}, \dots, x_{k,t})$ and $\varepsilon_t$ is assumed to be an uncorrelated term, i.e. white noise. The errors from a regression can be allowed to contain autocorrelation, thus these can be described as $n_t$ rather than $\varepsilon_t$ to emphasize this change in perspective. The error ($n_t$) can then be assumed to follow an ARIMA model. As an example, an ARIMA(1,1,1) model can be described as:

$$\begin{aligned} \mathcal{Y}_t &= \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + n_t, \\ (1 - \emptyset_1 B)(1 - \beta) n_t &= (1 + \theta_1 B)\varepsilon_t, \end{aligned} \tag{2.10}$$

with $\varepsilon_t$ being a withe noise series. Accordingly, the model contains two error terms – from the regression model ($n_t$) and ARIMA model ($\varepsilon_t$) – but only the ARIMA errors are assumed to be white noise.

## 2.2. Machine learning

ML is a relatively new field of study that focuses on the use of algorithms that are able to improve automatically through experience and by the usage of data (Mitchell 1997;

Jordan and Mitchell 2015; Chinnamgari 2019). ML lies at the intersection of the fields of computer science and statistics. As Jordan and Mitchell (2015) highlighted, two main questions are at the forefront of ML: "*How can one construct computer systems that automatically improve through experience?*" and "*What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?*". Accordingly, ML is important for addressing fundamental scientific questions and business challenges, thus being relevant for the most different fields of knowledge and applications, including economics, finance and policing, marketing, biology and health care, and social and educational sciences (Jordan and Mitchell 2015; Vamathevan et al. 2019; Grimmer et al. 2021).

ML is based upon the notion that computers can learn without explicit programming and that logic can be developed based on the data provided to the ML algorithms (Figure 2). The most commonly applied ML algorithms are based on function approximation problems so that tasks are embodied into functions (Jordan and Mitchell 2015). For many applications, it can be far more powerful and easier to follow the frameworks of ML, e.g. training a system by providing examples of desired input-output behaviours, than programming and anticipating the desired responses for all possible inputs (Mitchell 1997; Jordan and Mitchell 2015). In such cases, the learning problem involves improving the accuracy of functions based on the experience of algorithms through a sample of known input-output pairs. Furthermore, ML functions can sometimes be represented in an explicit parameterized functional form. In other situations, they can be implicit or obtained via a search process, an optimization or a simulation-based procedure (Mitchell 1997; Jordan and Mitchell 2015).

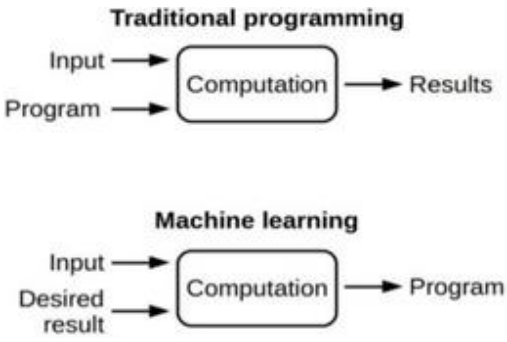



Figure 2. Differences between traditional programming and Machine Learning, i.e. the capacity of Machine Learning algorithms to automatically infer actions from the data without manual interventions. Source: adapted from Chinnamgari (2019).

There is an extensive range of ML algorithms and methods. Some well-known and widely applied ones are linear and logistic regression, linear discriminant analysis, support vector machines, naïve Bayes and RF. As Jordan and Mitchell (2015) summarized in their work, the core difference between ML algorithms is how they represent the candidate programs (e.g. decision trees, mathematical functions or general programming languages).

### 2.2.1. Machine learning approaches

ML methods are commonly and broadly classified into three main approaches, (i) supervised, (ii) unsupervised, and (iii) reinforcement learning systems, although these categories can sometimes be combined to accomplish appropriate results in particular applications (Jordan and Mitchell 2015; Chinnamgari 2019). Supervised learning involves teaching machines the relation between input and target variables, supported by classification and regression problems as the two core segments (Jordan and Mitchell 2015). Supervised systems (Figure 3) are the most widely used ML method, beneficial when one is very clear about the results of a problem but is unsure about the relationships in the data affecting the output (Chinnamgari 2019). Supervised learning systems generally form their predictions via a learned mapping $f(x)$ that produces an output $(y)$ for each input $(x)$, or a probability distribution over $y$ given $x$ (Jordan and Mitchell 2015).

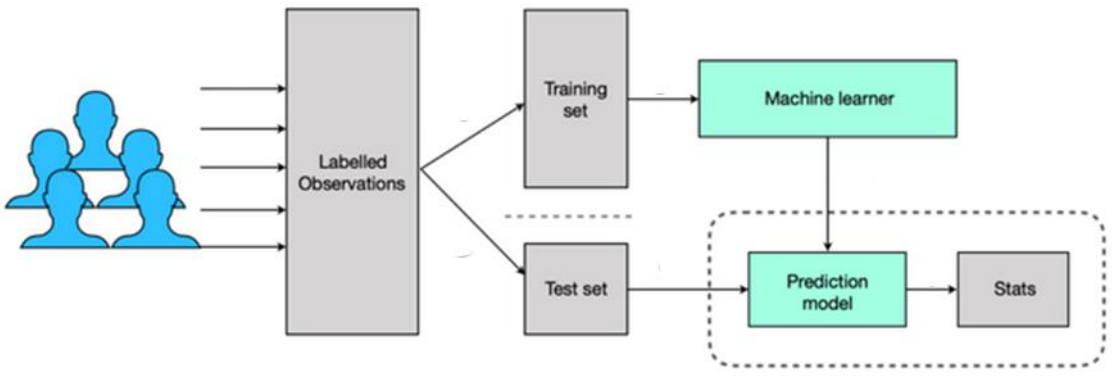



Figure 3. Summary of the supervised learning process within Machine Learning. Source: adapted from Chinnamgari (2019).

Unsupervised learning consists of allowing ML algorithms to learn independently, thus without supervision (Gareth et al. 2013; Jordan and Mitchell 2015). The two main categories of unsupervised learning are dimensionality reduction and clustering. Some of the most well-known algorithms within the field are the k-means and hierarchical clustering (Gareth et al. 2013). Unsupervised systems frequently aim to find relations

and hidden patterns in the data. Examples of applications include inspecting online sales data to identify different profiles of clients making purchases, grouping customers according to specific attributes or identifying the products purchased together (Gareth et al. 2013; Chinnamgari 2019). Also important, applying unsupervised systems is far more challenging than supervised ones since this exercise is much more subjective without expecting clear results (e.g. prediction of a target variable) (Gareth et al. 2013). In addition, such application is sometimes driven by the fact that labelled data is difficult and expensive to obtain, thus making unsupervised systems preferable over supervised ones in several business domains (Chinnamgari 2019).

Reinforcement learning counts on the notion that an "agent" can learn some behaviour by providing feedback from its environment (Morales and Zaragoza 2012). Typically, the agent can act in a given environment without supervision and receive either a positive or negative reward according to the action taken (Morales and Zaragoza 2012; Nian et al. 2020). Later, the ratio of rewards is evaluated, and the agent's path is re-evaluated. Some applications of reinforcement systems include learning-based robots, video games, natural language processing, autonomous driving and financial trading (Morales and Zaragoza 2012; Nian et al. 2020).

### 2.2.2. Random forest algorithm

RF is a supervised learning algorithm developed originally by Breiman (2001) concerning the concept of ensemble learning, i.e. the process of combining multiple "building blocks" in order to solve problems and improve models' performance (Gareth et al. 2013). Accordingly, the RF algorithm paves on constructing multiple decision trees. Thus, once combining their outputs, a single result with improved predictive accuracy is reached (Breiman 2001). This method can handle both classification and regression problems. Classification trees are appropriate when the target variable is categorical, and regression trees are suitable when the target variable is continuous (Cutler et al. 2012; Gareth et al. 2013). The explanatory variables can also be categorical or continuous.

Decision trees are generally grown from a "root" node that contains all the observations from the initial dataset, i.e. the entire predictor space (Cutler et al. 2012). Usually, through a binary partition ("split") on a single predictor variable, the instances of this node distribute to one of the two descendant nodes (Cutler et al. 2012). The split that a tree uses to partition a node into its two descendant nodes, one on the left and one on the right, is given by considering every possible split on every predictor variable. Different splitting criteria exist for classification and regression trees, such as entropy and residual sum of squares (Cutler et al. 2009). Some nodes do not split; these are the

"terminal" or "leaf" nodes since they form the final partition of the predictor space (Cutler et al. 2012).

Decision trees are deliberately grown larger than necessary until a stopping criterion is met, after which pruning back the tree is required to prevent over-fitting (Cutler et al. 2009). After pruning, the non-partitioned nodes, i.e. "terminal" nodes, are utilized to predict values. In regression and classification exercises, these predicted values are obtained, respectively, by averaging the responses of terminal nodes and by calculating the most frequent class or assessing the proportion of categories (Cutler et al. 2009, 2012). Overall, tree-based methods and their outputs present advantages over other methods, such as being easy to understand and capable of being displayed graphically, and having a facilitated capacity to handle qualitative predictors (Cutler et al. 2012; Gareth et al. 2013). On the other hand, sometimes these methods do not have the same level of accuracy of more recent regression and classification models, and sometimes are not robust, in the sense that even small changes in the data can determine big changes in the final estimated trees (Gareth et al. 2013).

RF relies on the idea of combining a large number of trees to provide improvements in prediction accuracy, as represented in Figure 4. In its original form, this method was an extension of the bagging framework, which was initially developed as a competitor to boosting algorithms (Cutler et al. 2012). In short, bagging involves the creation of multiple copies of the training data set using the bootstrap technique, where a separate decision tree is fitted to each one of these copies, and their combination allows for creating a single predictive model (Cutler et al. 2012; Gareth et al. 2013). On the other hand, trees produced by boosting are created sequentially. Thus, each tree is grown by utilizing the information provided by the previously grown tree (Gareth et al. 2013). Bagging is a strategy that controls the variance better and allows for improved accuracy in decision trees over boosting (Cutler et al. 2012; Gareth et al. 2013).

Similar to bagging, the decision trees of random forests are typically built on bootstrapped training samples, even though other strategies can be applied. However, each time a split is considered, a random sample of $m$ predictors is chosen as a candidate for that split from the complete set of $p$ predictors; yet, each split is only allowed to use one of the $m$ predictors (Gareth et al. 2013). Furthermore, a new sample of $m$ predictors is taken at each split, and typically this number corresponds to $m \approx \sqrt{p}$, i.e. the approximation to the square root of the total number of predictors (Gareth et al. 2013). The strategy utilized by RF forces each split to consider only a subset of the predictors, which is the core difference between such a methodology and other tree-based methods,

justifying an averaging output of resulting trees as less variable and more accurate and reliable (Breiman 2001; Cutler et al. 2012; Gareth et al. 2013).

RF shows unique advantages over other ML methods, including the capacity to handle large data sets, both classification and regression problems and missing values on the predictor variable; ability to be used for solving high-dimensional problems; depend on a few tuning parameters; provide measures of variable importance; and being relatively fast to train and predict (Cutler et al. 2012; Feng et al. 2018; Castilho 2020). The major disadvantage is that the algorithm does not offer simple summaries of relationships in the data, thus being generally perceived as a “black box” that is challenging to interpret (Kane et al. 2014; Varian 2014).

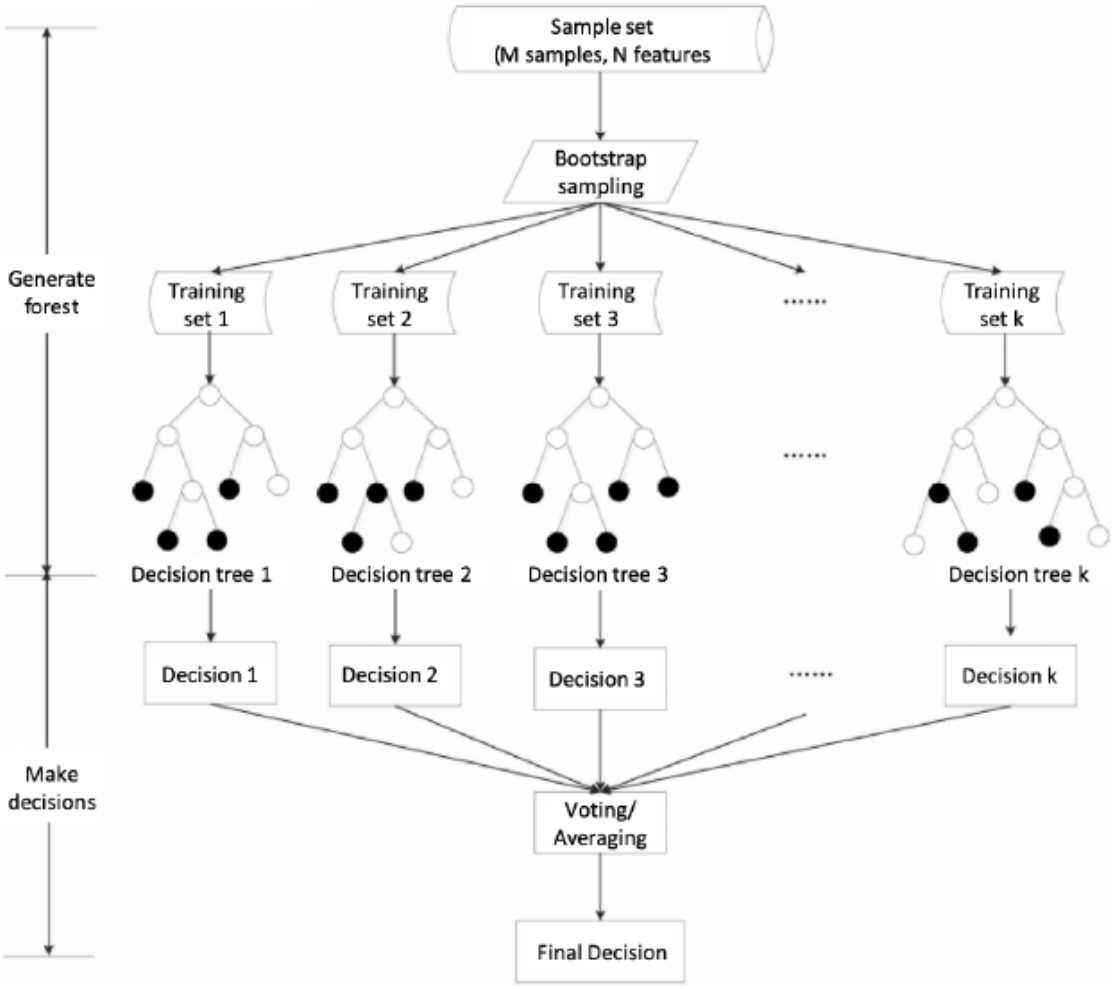



Figure 4. Representation of the traditional workflow of the Random Forest algorithm. Source: Castilho (2020), previously adapted from Feng et al. (2018).

## 2.3. Forecasting

The primary objective of forecasting is to predict the future values of a time series accurately (Shumway and Stoffer 2017). This is widely recognized as one of the most important aspects of time series analysis. It has become increasingly relevant in academia and business, especially by playing an important role in decision-making by management, government agencies and policymakers (Enders 2015; Mahalakshmi et al. 2016). Effective business planning and economic monitoring require forecasts at different time horizons, including the short and long terms (Montgomery et al. 2015; Hyndman and Athanasopoulos 2018). The success of these tasks primarily relies on the identification of the critical questions to be answered by the forecasting exercise, the ability to apply different methods and selection of the most appropriate ones for each specific problem, and the capacity to refine the forecasts over time (Metcalfe and Cowpertwait 2009; Montgomery et al. 2015). However, the achievement of these milestones depends upon the amount and quality of the data available, as well as on the capacity to identity the time series components and factors contributing to the behaviour of the series (Montgomery et al. 2015; Hyndman and Athanasopoulos 2018).

In more recent times, the field of forecasting has drifted towards the notion of nowcasting, a concept first utilized in the 1980s that originated from meteorological studies (Hopp 2022). At that time, the term described weather forecasting for the near future by utilizing information on the current meteorological conditions (WMO 2017). Also, the term has spread to many fields, including Economics, in which terminology appeared in related literature from the mid-2000s (Hopp 2022). Nowadays, nowcasting is a central element in many economic fields and business settings, being utilized for obtaining estimates of the macroeconomic situation and business cycles (Kapetanios and Papailias 2018; Berthomier et al. 2020; Caperna et al. 2020). Nevertheless, nowcasting remains a broad term whose definition may change according to the field of application; yet, in economics, it commonly refers to the methods of time series modelling that predict the present, the very near future and the very recent past (Bańbura et al. 2010). Accordingly, nowcasting is mostly required in situations requiring timely predictions, and nowcasting predictions can also serve as input for forecasting over longer time horizons (Bańbura et al. 2010; WMO 2017).

### 2.3.1. Forecasting with machine learning

A growing body of literature shows that ML-based models are particularly well-suited for forecasting and nowcasting economic data by presenting increased capacity in situations

of rapid changes in the external environment, such as changes in the macroeconomic environment, business cycle or even unexpected crisis and events (Chapman and Desai 2021; Dauphin et al. 2022). This increased superiority has been mostly attributed to the intrinsic capacity of ML algorithms to capture non-linear movements in macroeconomic and financial data (Katris 2020; Kreiner and Duca 2020; Gogas et al. 2021). As a paradigmatic example, Gogas et al. (2021) showed that a RF-based model reached the best accuracy for forecasting unemployment in the European zone, in comparison to two other ML-based and one benchmark model. Similarly, RF showed very good performance in other contexts, including forecasting virus outbreaks, monthly temperature variations, stock index movements, individual stock prices, and Gross Domestic Product variation (Kane et al. 2014; Naing and Htike 2015; Yoon 2021; Abraham et al. 2022). Furthermore, these and other studies provided important clues regarding the application of the RF algorithm in the time series context, including the notion that the parameter tuning stage should focus primarily on two hyperparameters, i.e. number of trees to grow in the forest and number of variables to consider at each split (Tyralis and Papacharalampous 2017; Kompella and Chilukuri 2019). Also, it was demonstrated that the application of time series cross-validation, such as the ones in the form of rolling or expanding windows, could lead to more accurate forecasts than with the cross-validation schemes commonly applied in ML studies, e.g. bootstrapping and k-fold cross-validation (Hyndman and Athanasopoulos 2018; Balogh et al. 2022). Respectively, in such cases the algorithm may rely on different data splits of the series for training rather than the traditional bootstrap resampling (Hyndman and Athanasopoulos 2018; Balogh et al. 2022).

Besides allowing for accuracy improvements, ML algorithms (including specifically RF) present the advantages of accelerating time series data processing, improve the capacity to analyse higher volumes of data and identify hidden patterns, increase the adaptability of models and automate forecast updates based on new data becoming available (Cerqueira et al. 2022; Haselbeck et al. 2022). Nevertheless, it is also fair to admit that several challenges and missing gaps still exist in this context, as demonstrated by academic studies, business insights and student competitions (Makridakis et al. 2018; Bojer 2022). Respectively, a main problem is that forecasting methods based on ML are much more complex to interpret than classical ones (Bojer 2022). In addition, there is still limited knowledge on why, when and how these methods outperform classical proven methods (Bojer 2022). In fact, there are few large scale studies that have comprehensively compared ML and traditional models, and sometimes ML-based models were not capable to outperform benchmark models (Makridakis et al.

2018). Thus, a key recommendation for future studies includes the need to focus more extensively on the comparison and validation of the forecasting performance of ML-based models, especially by utilizing benchmark models for this purpose (Makridakis et al. 2018; Cerqueira et al. 2022).

### 2.3.2. Forecasting accuracy

Determining forecasts' accuracy is essential for all applied fields, in the sense that more accurate forecasts support higher confidence in the models applied and outcomes of the forecasting exercises (Shcherbakov et al. 2013). Accordingly, the determination of forecasting errors allows us to capture the difference between an observed value and its forecast, mathematically described through the formula (Hyndman and Athanasopoulos 2018):

$$e_{t+h} = y_{t+h} - \hat{y}_{t+h|t},$$

with the observed and forecasted data being given by $y_{t+1}, y_{t+2}, \ldots$, and $\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots$, respectively. Table 1 depicts numerous metrics to measure forecast accuracy, grouped according to their statistical properties. Hyndman and Athanasopoulos (2018) classify such metrics as scale-dependent errors, percentage errors and scaled errors. The scale-dependent errors metrics (e.g. Root Mean Squared Error) concern those in which the forecast errors are given on the same scale as the data, thus being utilized for comparing forecasts with the same scale units (Hyndman and Athanasopoulos 2018). On the other hand, percentage errors metrics, such as the Mean Absolute Percentage Error, are calculated based on the value $P_t$, i.e. where $P_t = \frac{100e_t}{y_t}$, thus being unit-free and capable of comparing forecasting performances between different data sets (Shcherbakov et al. 2013; Hyndman and Athanasopoulos 2018). Finally, scaled errors metrics were originally proposed by Hyndman and Koehler (2006) as an alternative to using percentage errors when comparing accuracies across series with different units, and the authors defined a scaled error as:

$$q_i = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n} |y_i - y_{i-1}|}$$

with the numerator and denominator referring to values on the scale of the original data, and therefore $q_i$ is independent of the scale of the data (Hyndman and Koehler 2006; Hyndman and Athanasopoulos 2018).

Table 1. Most common measures for assessing forecast errors, their classification, formula and a brief comment. Sources: Shcherbakov et al. (2013), Hyndman and Koehler (2006) and Hyndman and Athanasopoulos (2018).

| Measure | Classification | Formula | Comment |
|---|---|---|---|
| Mean Absolute Error (MAE) | Scale-dependent errors | $MAE = mean(\|e_t\|)$ | One of the most utilized, since it is easy to understand and compute. It is based on the average of the absolute differences between prediction and actual observation |
| Root Mean Squared Error (RMSE) | Scale-dependent errors | $RMSE = \sqrt{mean(e_t^2)}$ | Commonly utilized and a popular measure when aiming to assess prediction errors in the same unit as the original series. It is based on averaging the square error value of each observed error |
| Mean Absolute Percentage Error (MAPE) | Percentage errors | $MAPE = mean(\|Pt\|)$ | One of the most popular measures of percentage errors. It is widely utilized for comparing forecasting performances between data sets |
| Mean Absolute Scaled Error (MASE) | Scaled errors | $MASE = mean(\|qi\|)$ | Frequently utilized to compare the accuracy of time series with different scale units. Besides, it has proven to be a reliable alternative to percentage errors |

# 3. Literature review

## 3.1. Unemployment forecasting

Unemployment has always been a critical topic concerning entire nations and societies. The unemployment phenomenon defines the condition of a country's economic equilibrium and is a social development barrier; thus, increases in unemployment are associated with rises in economic inequality and social instability. Also, the psychological burdens affecting individuals that remain out of the workforce and respective households have been widely described in the literature (Dritsakis and Klazoglou 2018; Chakraborty et al. 2021). In this regard, previous research has shown a direct relationship between job loss and physical and mental health deterioration, including a greater propensity for lifestyle risk behaviours, such as unhealthy eating habits, increasing alcohol use and smoking (Plessz et al. 2020; Macassa et al. 2021).

Unemployment generally acts as a measure of economic expansion and cycles at the macroeconomic level, thus being considered a vital measure of the overall economic health (Dritsakis and Klazoglou 2018; Chakraborty et al. 2021). Furthermore, increasing unemployment rates are typically correlated to reduced tax revenue, which originates from the loss of income and purchasing power by unemployed persons and increasing inflationary pressures (Gogas et al. 2021). In addition, higher unemployment accounts for increased government spending, mostly related to unemployment benefits provided to citizens and government retraining programs (Gogas et al. 2021). The unemployment rate is also a key metric looked at by central banks, such as the European Central Bank and the United States Federal Open Market Committee (Stockhammer and Sturn 2012; Gogas et al. 2021). Central banks aim to preserve monetary stability, including potentiating and reaching full employment (Stockhammer and Sturn 2012; Gogas et al. 2021). Accordingly, the unemployment rate is critical to determine the path of interest rates, which is the central banks' primary monetary policy tool (Stockhammer and Sturn 2012).

From a historical perspective, Gogas et al. (2021) highlight the significance of unemployment forecasting, which was perhaps first acknowledged in the 1970s, a decade marked by slow economic growth and high inflation, i.e. stagflation (Brunner et al. 1980). Before the stagflation of the 1970s, economists primarily relied on the traditional Phillips Curve, which assumes an inverse relationship between the inflation rate and unemployment (Gogas et al. 2021). This relation demonstrated to be a statistical stylized fact that was only valid in very short time frames, in which monetary policy is

unexpected (Muth 1961; Gogas et al. 2021). From this point in time onwards, accurately forecasting unemployment has become even more fundamental for fiscal and monetary policymakers to recognize the problem promptly in order to implement practices to mitigate it (Gogas et al. 2021). Later on, in the middle of the 1990s, the literature on unemployment forecasting bloomed due to significant advances in statistical methods, which allowed for improved accuracy of forecasts (Chakraborty et al. 2021). In more recent times, there has been renewed interest in forecasting and nowcasting unemployment due to the rise in more powerful statistical tools to perform this job, such as ML algorithms, as well as increasing availability of data through developments in data sources, collection devices and infrastructure (Simionescu and Zimmermann 2017; Bok et al. 2018; Gogas et al. 2021). This trend accelerated following the COVID-19 pandemic outbreak, where economies worldwide experienced a period of economic retraction marked by lockdowns, reduced demand for products and services and business closures, which forced many workers into temporary unemployment (Şahin et al. 2020; Hopp 2022). Such a fact led to a proliferation of investigations and government reports aiming to forecast how much the unemployment rate would increase and how persistent this increase would be over the subsequent months and years (Caperna et al. 2020; Hopp 2022).

Unemployment is also a central topic in Portugal. Research surveys conducted since the 1980s show that unemployment consistently appears as one of the main problems and concerns for Portuguese citizens (Marques 1999). Accordingly, the Portuguese population experienced a sense of loss of identity, stress and powerlessness due to unemployment rises, as well as a lack of purpose and structure in their daily lives (Macassa et al. 2021). In addition, reports account for in-hospital mortality increases in the country in periods of higher unemployment, thus indicating a link between health deterioration and poorer economic conditions (Pessoa et al. 2020). Moreover, there is evidence that unemployment is a more relevant topic in Portugal than in other European countries, justified by the fragile state of the Portuguese economy (Marques 1999; Baptista and Thurik 2007). Respectively, the effects of unemployment on the Portuguese labour market are very asymmetric due to discrepancies in geographical regions, sectors of activity, age groups and nature of labour ties (Almeida and Santos 2020). Portuguese regions whose economies are sustained by tourism or foreign investments are generally the most affected (Almeida and Santos 2020). At the same time, the population characterized as less educated, young and women with unstable employment relationships also seem particularly vulnerable (Pacifico and Thévenot 2016; Almeida and Santos 2020). Additionally, the Portuguese economy features a high proportion of

"micro-businesses" created for subsistence that have minimal impact on employment and economic growth; thus, macroeconomic fluctuations associated with the European business cycle and European Union cohesion funding determine the overall economic performance of the country (Marques 1999; Baptista and Thurik 2007). Therefore, the deterioration of macroeconomic conditions in Europe and inversions of the overall economic cycle reflect more deeply in Portugal than in other European countries (Baptista and Thurik 2007). As a paradigmatic example, Royo (2010) points out that although Portugal and Spain share a lot of traits, such as a similar location and political history, the Portuguese economy shows poorer performance. Additionally, Portugal has received intervention of the International Monetary Fund on three occasions (Macassa et al. 2021). The most remarkable was during the financial crisis of 2008, which substantially impacted the Portuguese unemployment rate (Pacifico and Thévenot 2016; Macassa et al. 2021).

## 3.2. Big data and Google Trends

Traditionally, forecasting economic and business dynamics leans on predictors gathered from governmental agencies, central banks, and companies' annual reports and financial statements (Wu and Brynjolfsson 2015). Such records are generally available with significant delay to the present and consist of information aggregated into a narrow scope of prespecified categories (Wu and Brynjolfsson 2015; Kapetanios and Papailias 2018). These data present several challenges for forecasting, especially when aiming to address time-sensitive issues and novel subjects/methods (Wu and Brynjolfsson 2015; Fornaro and Luomaranta 2020; Richardson et al. 2021). On the other hand, recent technological developments in devices and infrastructure promoted access to new data sources, namely data originating from internet searches, e-mails, smart cards, commercial websites, mobile phone apps and social media (Liu et al. 2016; Jun et al. 2018), derived from the Big Data advent. Big Data is a term that refers to information that presents five key characteristics (5 Vs), namely veracity, volume, velocity, variety and value (Ishwarappa and Anuradha 2015; Liu et al. 2016; Jun et al. 2018). Respectively, these relate to the credibility of such data, its large size, the swiftness at which it is generated and acquired, its multiplicity and heterogeneity, and its recognized power to transmute data into valuable information. In particular, Big Data can now be accessed at nearly zero cost, virtually instantaneously and at a fine-grained level of disaggregation; as a result, there has been an unprecedented amount of Big Data being collected, stored and communicated within organizations and companies (Wu and Brynjolfsson 2015; Jun et al. 2018). The interest in such data sources lies in the fact that users reveal information

about their needs, interests and concerns when using the internet. Thus, Big Data contains valuable clues on human behaviours, consumer patterns and sentiment, and therefore can be useful to predict economic sentiment, market conditions and business activities (Askitas and Zimmermann 2015; Jun et al. 2018).

Search engines are mechanisms to perform online searches. These are adopted worldwide and frequently used by a large percentage of the world's population, and Google search is the most popular among users (Haider and Sundin 2019). The records of Google searches are made available by Google Trends[2] (GT), a data exploration tool that was introduced in 2006 with the mission of facilitating the reporting of historical search indexes made to the Google search engine (Caperna et al. 2020; Mihaela 2020). More precisely, GT data is based on random representative samples of queries performed in the Google search engine. Instead of reporting an absolute volume of searches given a keyword, period and geographical area, GT returns the Search Volume Index (SVI), i.e. a normalized measure scaled by the maximum demand on the queried period, resulting in a 0 to 100 index (Choi and Varian 2012; Caperna et al. 2020). Accordingly, a value closer to 100 indicates the higher popularity of the search term in the selected time period and region (Choi and Varian 2012; Caperna et al. 2020). Also, different queries that can be assigned to a semantic domain are aggregated by GT into topics (Caperna et al. 2020). In addition, GT provides information on categories, and these originate from the classification of queries by a natural language engine into about 30 categories at the top level and 250 at the second level (Choi and Varian 2012).

GT indeed provides a vast amount of information. However, researchers point out some limitations related to this data. The main issues include the impossibility of accessing the raw data and the lack of general knowledge about the precise algorithms used to pre-treat and summarise such data (Kapetanios and Papailias 2018). There is also a sampling noise issue, originating from GT data relying on samplings of searches; thus, SVI varies according to the day of collection, which potentially introduces sample noise and influences forecasting exercises (Rovetta 2021; Cebrián and Domenech 2022). Also, the data repository may be prone to manipulation, e.g. search bots could generate irrelevant search queries, thereby changing the search indexes (Wu and Brynjolfsson 2015). Additionally, the profile of internet users is not random for socio-economic status and population age (Caperna et al. 2020). Such factors can introduce sample bias and influence studies utilizing GT as an informational source.

On the other hand, GT provides facilitated data access and collection, supporting data management and treatment. In fact, this data source is generally perceived as of good

[2] https://trends.google.com/trends/

quality, available continuously and without significant definitional changes (Kapetanios and Papailias 2018). In addition, compared to surveys, data from Google searches are less sensitive to small-sample bias (Caperna et al. 2020).

The potential of GT as a data source is to provide relevant information that enables the production of forecasts more timely and with higher accuracy (Kapetanios and Papailias 2018; Caperna et al. 2020). The usefulness of GT data has been widely demonstrated across diverse research and business contexts, including communications, medicine, environmental sciences, business and economics (Choi and Varian 2012; Jun et al. 2018; Kapetanios and Papailias 2018). Some examples include its usage to monitor the spread of epidemics and diseases, foreshadow housing prices, assess the current sentiment of stock market investors and predict stock returns, and forecast tourism inflows and macroeconomic variables (Artola et al. 2015; Wu and Brynjolfsson 2015; Bijl et al. 2016; D'Amuri and Marcucci 2017; Mavragani and Gkillas 2020).

## 3.3. Forecasting the unemployment: the role of Google Trends

Targeting unemployment forecasting, many time series models have extensively been employed, from traditional models to more recent ones. Some examples of classical methods include the utilization of ARIMA models for forecasting unemployment in several European, Asian and North American countries, among others (Jelena et al. 2017; Khan Jaffur et al. 2017; Lai et al. 2021). In many of these studies, classical models have proven useful and effective for forecasting unemployment; however, in several others, classical models could not allow for the effects of shocks in unemployment (Blanchard and Summers 1986; Galbraith and van Norden 2019). Such a divergence drives the exploration of more complex models for unemployment forecasting, including those based on ML and in data originating from internet searches (Feuerriegel and Gordon 2019; Galbraith and van Norden 2019; Gogas et al. 2021).

Askitas and Zimmermann (2009) report the connection between monthly unemployment rates in Germany and Google searches related to this topic for the first time after the introduction of GT, namely by using GT searches as predictors to forecast the evolution of unemployment under the uncertain context of the Great Financial Crisis. From this point onwards, researchers have increasingly utilized Google data, which is recognized as the current most powerful source to collect internet data to explain unemployment in real-time and improve future predictions (D'Amuri and Marcucci 2017; Jun et al. 2018; Mihaela 2020). In this regard, it is important to note that most studies on this matter take into account the national level rather than the regional level and that the

keywords collected from GT typically represent words and expressions in the language of the studied country (Simionescu and Zimmermann 2017; Simionescu et al. 2020). Accordingly, in the context of forecasting and nowcasting the unemployment rate with GT data, the keyword “jobs” has been the most frequently used for this purpose (Dilmaghani 2019; Kundu and Singhania 2020; Simionescu et al. 2020). In addition, other search terms have been utilized, including “unemployed”, “unemployment” and “unemployment benefit” (McLaren and Shanbhogue 2011; D’Amuri and Marcucci 2017). Interestingly, these studies showed that predictions constructed with models utilizing GT data were more accurate than those provided from the Survey of Professional Forecasters and models based on labour force flows (McLaren and Shanbhogue 2011; D’Amuri and Marcucci 2017). Moreover, there is potential for GT improving unemployment forecasting in developing countries as well. For instance, Lasso and Snijders (2016) showed that in Brazil there was a strong correlation between unemployment rates and Google search volumes of keywords related to job searches, while Chadwick and Sengül (2015) showed that the utilization of Google searches for words related to unemployment could improve unemployment nowcasting in Turkey.

Numerous studies targeted GT data to forecast and nowcast unemployment in South-Western European countries at a national level, mainly by collecting data from keywords in each countries’ language (Francesco 2009; Vicente et al. 2015; Simionescu and Cifuentes-Faura 2022). Several of these were carried out for Italy, being demonstrated that the most popular keyword for job searches was “offerte di lavoro” (job offers), and that unemployment forecasts based on models utilizing Google searches outperformed competitor models (Francesco 2009; Naccarato et al. 2015). Furthermore, data from GT was successfully utilized to improve the unemployment forecasting and nowcasting of specific population groups of Italy and France (Fondeur and Karamé 2013; Naccarato et al. 2018). Other studies showed promising results when aiming to forecast unemployment in Spain by utilizing Google searches, with the words “oferta de trabajo” and “desempleo” (i.e. job offers and unemployment) being the ones most frequently selected for this purpose (Vicente et al. 2015; González-Fernández and González-Velasco 2018; Simionescu and Cifuentes-Faura 2022).

Previous studies also utilized Google Trends data to forecast and nowcast unemployment variables in Portugal, as summarized in Table 2. Respectively, when Barreira et al. (2013) explored the potential of internet searches for nowcasting purposes, they found that GT data improved unemployment nowcasts in Portugal, France and Italy. The keywords selected for their analysis were “unemployment” and “benefits for unemployed”, and these were formulated in the language of each country. More recently,

Simionescu and Cifuentes-Faura (2022) tested whether unemployment rate forecasts based on GT data improved predictions at the national level for Spain and Portugal. By utilizing the keywords “unemployment” and “job offers” in each countries’ language, the authors showed that the predictions based on GT were significantly more accurate for both countries. Also, a study performed by Caperna et al. (2020) explored the economic consequences of coronavirus in the member states of the European Union. The authors nowcasted monthly unemployment rates to assess the relationship between Google searches and the underlying phenomenon, as well as to identify the keywords that improved predictive accuracy. Their results showed that nowcasting unemployment using the unemployment topic alone did not provided a statistically significant improvement over what a simple autoregressive model would predict for most countries. However, once adding all the keywords linked to the topic unemployment and performing variable selection with a RF-based model, the predictive accuracy increased significantly in almost all countries. Taking the above mentioned into account, it is brought into light that the production of economic indicators in real time is the focus of a growing body of literature, and that GT is a powerful tool towards accurately forecasting and nowcasting unemployment variables.

Table 2. Summary of studies that utilized Google Trends data to forecast and nowcast unemployment variables in Portugal. Other countries included in the analysis, the periods under study, searched terms and time series models employed were also highlighted.

| Authors | Other countries | Period | Google Searches | Model |
|---|---|---|---|---|
| Barreira et al. (2013) | France, Italy and Spain | January 2005 to August 2013 | Keywords: unemployment and unemployment benefit | ARMA |
| Simionescu and Cifuentes-Faura (2022) | Spain | January 2004 to December 2020 | Keywords: unemployment and jobs | ARMAX |
| Caperna et al. (2020) | More 26 countries (member states of European Union) | January 2015 to May 2020 | Unemployment Topic; top-25 search terms associated with the topic; top-10 search terms associated with the above ones; | AR; RF |

# 4. Experimental study

This chapter presents the practical analysis developed to tackle the main issues raised in the introduction section and explored in the following chapters, as well as detail the objectives for this experimental study, methodology for model creation and proposed models, and forecasting-related approaches.

## 4.1. Objectives

The main goal of the present study is to evaluate the utility of Google searches for nowcasting the Portuguese unemployment rate by comparing predictions resulting from classical methods and a machine learning algorithm. Accordingly, the following specific objectives concern the application of different modelling approaches and comparison of models' nowcasting performance to achieve the main goal:

i) Development of a benchmark model based on the classical ARIMA model;
ii) Development of regression with ARIMA errors models that incorporate Google searches as explanatory variables;
iii) Development of machine learning-based models, all based on the random forest algorithm, that incorporate Google searches as explanatory variables;
iv) Evaluation and comparison of the models' nowcast accuracy and discussion of results per the available literature.

## 4.2. Methodology

This empirical analysis comprises five main steps, illustrated in Figure 5: 1) collection of time series data regarding Google Trends search indexes and unemployment rate; 2) preliminary analysis of these series; 3) development of a benchmark ARIMA and regression with ARIMA errors (regARIMA) models, and in parallel, development and hyperparameter tuning of several random forest-based models; 4) production of nowcasts for the unemployment rate; and 5) comparison of forecasting accuracy among models.

The established methodology relies on the R software (R Core Team 2019), to collect and analyse the data.

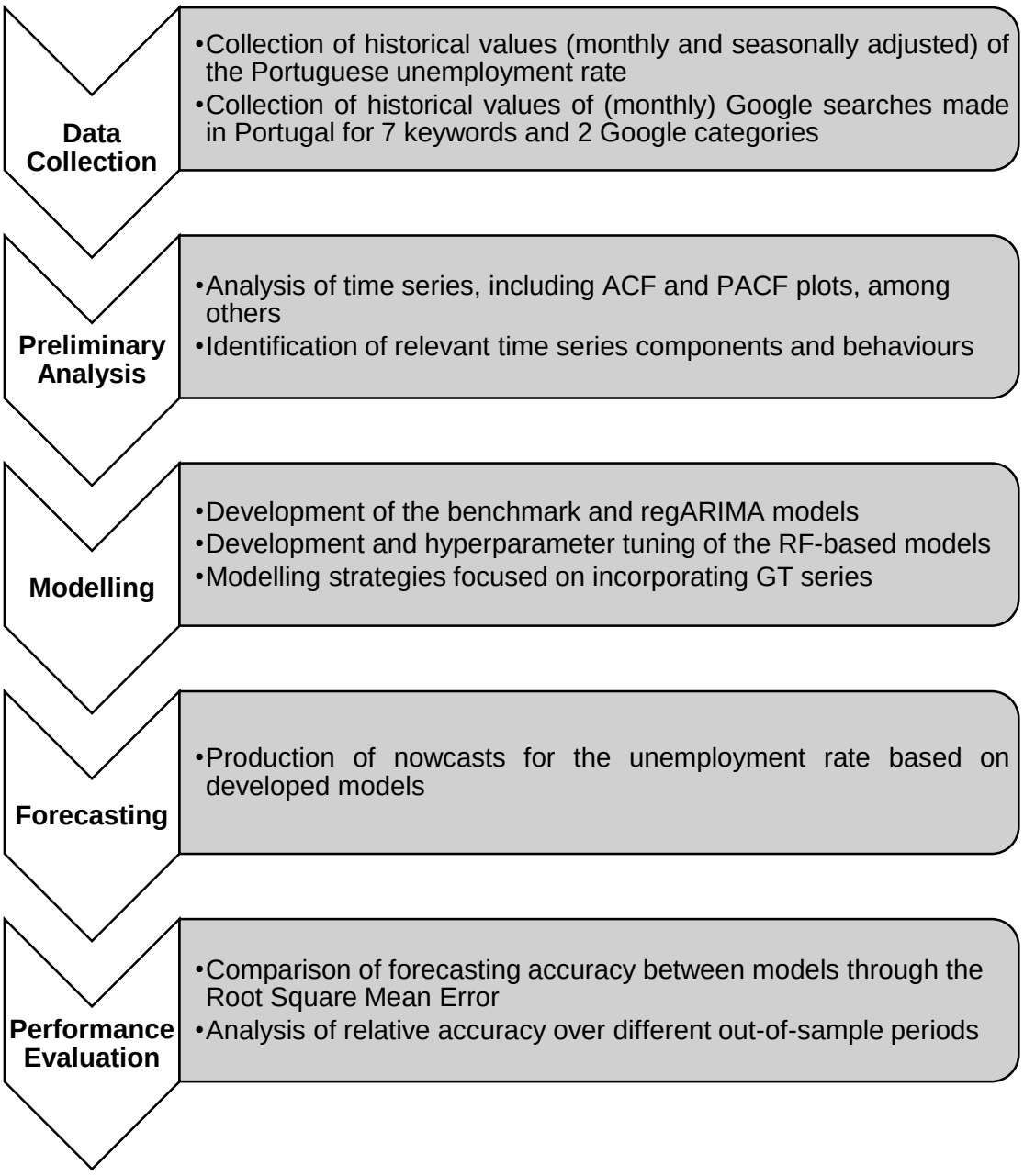



Figure 5. Schematic representation of the overall workflow and main steps of the analysis.

### 4.2.1. Data collection

This study focuses on time series related to the Portuguese unemployment rate and inquiries on the Google search engine. Data comprising this work span the period between January-2004 and March-2022. Respectively, time series are constrained to start in January-2004 to match the data availability of the GT series.

Google Trends[3] provides data from internet searches made on the Google search engine. Data collected concern the monthly Portuguese Search Volume Index values for seven keywords and two Google categories, as summarized in Table 3. The choice of keywords concerns individual words or expressions related to the unemployment phenomenon (unemployment, job, unemployment benefit, unemployment fund and job offers) and names of popular webpages for Portuguese job seekers – sapo jobs (website listing job vacancies) and iefp (an acronym for the Institute for Employment and Vocational Training in Portugal).

In order to control for the sampling noise issue, referred to in section 3.2, the data collection for each GT series is performed over 20 consecutive days and summarized into an average.

From the EUROSTAT[4] (statistical office of the European Union) was collected the Portuguese monthly unemployment rate seasonally adjusted data.

Table 3. Summary of the selected Google Trends parameters in the current experimental setting.

| Parameter | Values |
|---|---|
| Search keywords | *emprego* (*job*); *desemprego* (*unemployment*); *iefp* (*iefp*); *subsidio desemprego* (*unemployment benefit*)*; sapo empregos* (*sapo jobs*)*; fundo desemprego* (*unemployment fund*)*; ofertas de emprego* (*job offers*) |
| Search Google categories | *emprego* (*jobs;* ref. Google id60); *bem-estar e desemprego* (*welfare & unemployment*; ref. Google id706) |
| Language of searches | Portuguese |
| User location | Portugal |
| Time period | January-2004 to March-2022 |

### 4.2.2. Preliminary analysis

The preliminary analysis focuses on essential time series feature analysis, such as stationarity, seasonality, autocorrelation and partial autocorrelation functions, and behaviours over time.

### 4.2.3. Modelling strategies and proposed models

Three methods produce predictions for this experimental study, two involving classic time series methods (ARIMA and regARIMA) and the other employing a ML algorithm, namely the RF.

[3] https://trends.google.pt/
[4] https://ec.europa.eu/eurostat/web/main/home

The ARIMA, considered a benchmark model, utilizes only past values of the unemployment rate series as input, as described in equation 2.7. The selection of the *p*,*d*,*q* orders of this model is based on the "auto.arima" function ("forecast" package), namely by setting the Bayesian information criterion (Hyndman and Khandakar 2008). This function uses a variation of the Hyndman–Khandakar algorithm, thus combining unit root tests, minimization of information criteria and maximum likelihood estimates to determine the most suitable ARIMA model (Hyndman and Khandakar 2008).

The regARIMA model, described in equation 2.10, incorporates the GT series as explanatory series to the Portuguese unemployment rate and considers ARIMA errors (Table 4). The ARIMA orders are selected through the "auto.arima" function, namely by setting the Bayesian information criterion and the argument "xreg" to specify the predictor variables and their corresponding lags (Hyndman and Khandakar 2008).

The selection of the number of lags of each GT series allows for a maximum of 12 lags. The series and lag selection considers the coefficients' significance, i.e. GT series and lags showing coefficients not statistically significant are excluded from models while statistically significant ones are kept, considering a 5% significance level. The function "coeftest" ("lmtest" package) calculates the significance of coefficients (Zeileis and Hothorn 2002).

This procedure applies to the GT series and corresponding dataset without seasonality – designated as dGT in the following tables (Table 4). For this purpose, the GT series were decomposed with the Seasonal-Trend decomposition using Loess method.

Table 4. Modelling strategies applied to the benchmark and regression with ARIMA errors models.

| Method | Training Data | Final Model |
|---|---|---|
| ARIMA | Unemployment rate | Benchmark based on past unemployment rate values |
| RegARIMA | Unemployment rate + GT and lags | RegARIMA considering unemployment rate and relevant GT series and lags |
| RegARIMA | Unemployment rate + dGT and lags | RegARIMA considering unemployment rate and relevant dGT series and lags |

The RF models (see description of the algorithm in section 2.2.2) comprise the unemployment rate series and GT data as inputs for modelling. The tuning of these models focuses on two hyperparameters: i) the number of variables randomly selected from each model's global set when splitting a node – the total number of predictors varies from model to model according to the modelling strategy considered; and ii) the number of decision trees to grow (respectively, considering a prespecified number of trees, set as 400, 900 and 1500 trees). The optimal model is chosen based on the lowest RMSE.

The “caret” package supports the development of the ML models presented in this thesis (Kuhn 2008). It allows the implementation of the RF algorithm, while also enabling time-series cross-validation (Kuhn 2008; Fornaro and Luomaranta 2020; Balogh et al. 2022), as represented in Figure 6 (Hyndman and Athanasopoulos 2018). Also, the implemented RF algorithm provides measures of “variable importance”, thus giving indication on the variables with most predictive power. This is achieved by recording the prediction accuracy of each tree, and measuring it again after permuting each explanatory variable (Kuhn 2008).

A total of 12 RF-based models were developed on the basis of different modelling approaches, and by focusing on both GT and deseasonalized GT datasets.

One modelling strategy focuses on the inclusion of additional explanatory variables related to the unemployment rate series. Respectively, the feature engineering proposed by Krispin (2019) was considered, leading to the inclusion of the “external” explanatory variables: “trend”, a numeric index that captured the trend component of the unemployment rate series; “trend_sqr”, a second polynomial numeric index that captured the trend curvature of the series; “month”, a categorical variable for each month of the year that captured potential seasonal effects; and most relevant lags, capturing the correlation of the unemployment rate series with its lags. This strategy applies to both GT and deseasonalized GT datasets, as represented in Table 5.

Other modelling strategy focuses on the transformation of GT and deseasonalized GT datasets, namely by differencing those series presenting a unit root as determined by the Augmented Dickey-Fuller test.

Table 5. Modelling strategies applied to the development of random forest models with GT and dGT series.

| Algorithm | Training Data | Final Model |
|---|---|---|
| RF | Unemployment rate + GT | Tuned model considering unemployment rate and GT series |
| RF | Unemployment rate + its relevant lags, factor “month”, “trend” and “trend_sqr” + GT | Tuned model considering unemployment rate and its external variables and GT series |
| RF | Unemployment rate + GT (after differencing) | Tuned model considering stationary unemployment rate and stationary GT series |
| RF | Unemployment rate + dGT | Tuned model considering unemployment rate and dGT series |
| RF | Unemployment rate + its relevant lags, factor “month”, “trend”, “trend_sqr” + dGT | Tuned model considering unemployment rate and its external variables and dGT series |
| RF | Unemployment rate + GT (after differencing) | Tuned model considering stationary unemployment rate and stationary dGT series |

The other six RF models rely exclusively on the most relevant deseasonalized GT series and their 12 lags as explanatory variables, as represented in Table 6. Two modelling strategies apply for this purpose: i) visually analysing the cross-correlation functions of the unemployment rate and deseasonalized GT series (and their lagged versions), thus incorporating only the relevant predictors; ii) fitting RF models to the entire dataset (deseasonalized GT series and their lagged versions), and incorporating only the 25 variables yielding the highest "variable importance" into the final RF model.

Table 6. Modelling strategies applied to the development of random forest models with dGT series and lags. * Stands for models developed with the "variable importance" measure available through the random forest algorithm.

| Algorithm | Training Data | Final Model |
|---|---|---|
| RF | Unemployment rate + dGT and lags | Tuned model considering unemployment rate and relevant dGT series and lags |
| RF | Unemployment rate + its relevant lags, factor "month", "trend", "trend_sqr" + dGT and lags | Tuned model considering unemployment rate and its external variables and relevant dGT series and lags |
| RF | Unemployment + dGT and lags (after differencing) | Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags |
| RF* | Unemployment rate + dGT and lags | Tuned model considering unemployment rate and relevant dGT series and lags |
| RF* | Unemployment rate + its relevant lags, factor "month", "trend", "trend_sqr" + dGT and lags | Tuned model considering unemployment rate and its external variables and relevant dGT series and lags |
| RF* | Unemployment + dGT and lags (after differencing) | Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags |

### 4.2.4. Forecasting

All series are divided into an initial 80:20 ratio for training of ML models and nowcasting purposes using expanding samples, a time series cross-validation scheme commonly known as a rolling forecasting origin. In this strategy, the observations considered in the training and test sets change successively regarding each one-step ahead nowcast, as illustrated in Figure 6 (Hyndman and Athanasopoulos 2018). Benchmark ARIMA and regARIMA models are iteratively fitted to the train splits, while RF-based models are trained across the splits in order to reach a tuned model that is used for nowcasting. Accordingly, the prediction time frame comprehends the period between August-2018 to

March-2022, thus comprising a total of 44 months. The nowcasting accuracy is measured in terms of RMSE (described in section 2.3.2).

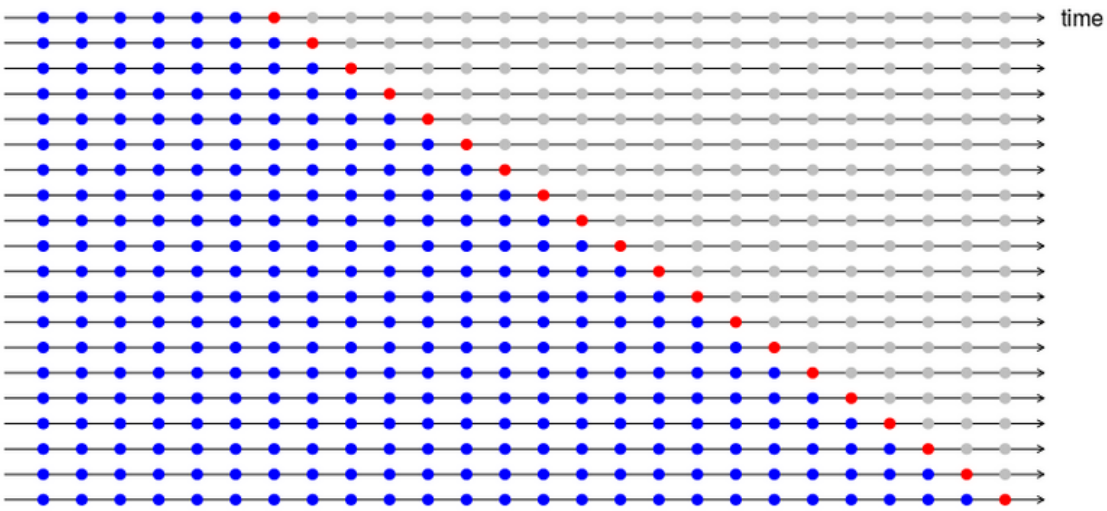



Figure 6. Representation of the cross-validation scheme adopted. The blue and red observations represent, respectively, the changing training and test sets. Source: adapted from Hyndman and Athanasopoulos (2018).

Additionally, to capture the COVID-19 influence on the forecasting performance, the whole prediction time frame is divided into a pre-pandemic period, from August-2018 to December-2019, and a pandemic period, from January 2020-March 2022.

# 5. Experimental results

This section details the major results obtained during the preliminary analysis and forecasting. Importantly, some minor results and details of the preliminary analysis are reported exclusively in the Appendix, including the analysis of ACF and PACF plots of the GT series.

## 5.1. Preliminary analysis

The GT series present different behaviours despite some common patterns being observed in Figure 7. Specifically, increasing trends (and maximum values) are observed in the years 2012 and 2013 for several of these series, after which a decrease in values occurs. Also, a sharp increase around the year 2020 is observed for several of them. This period is then followed by a rapid decline in values (Figure 7). These marked trends suggests that some GT series are not stationary.
Furthermore, it is somehow possible to identify the presence of seasonal components for several GT series through the ACF and PACF plot analysis (see Appendix).

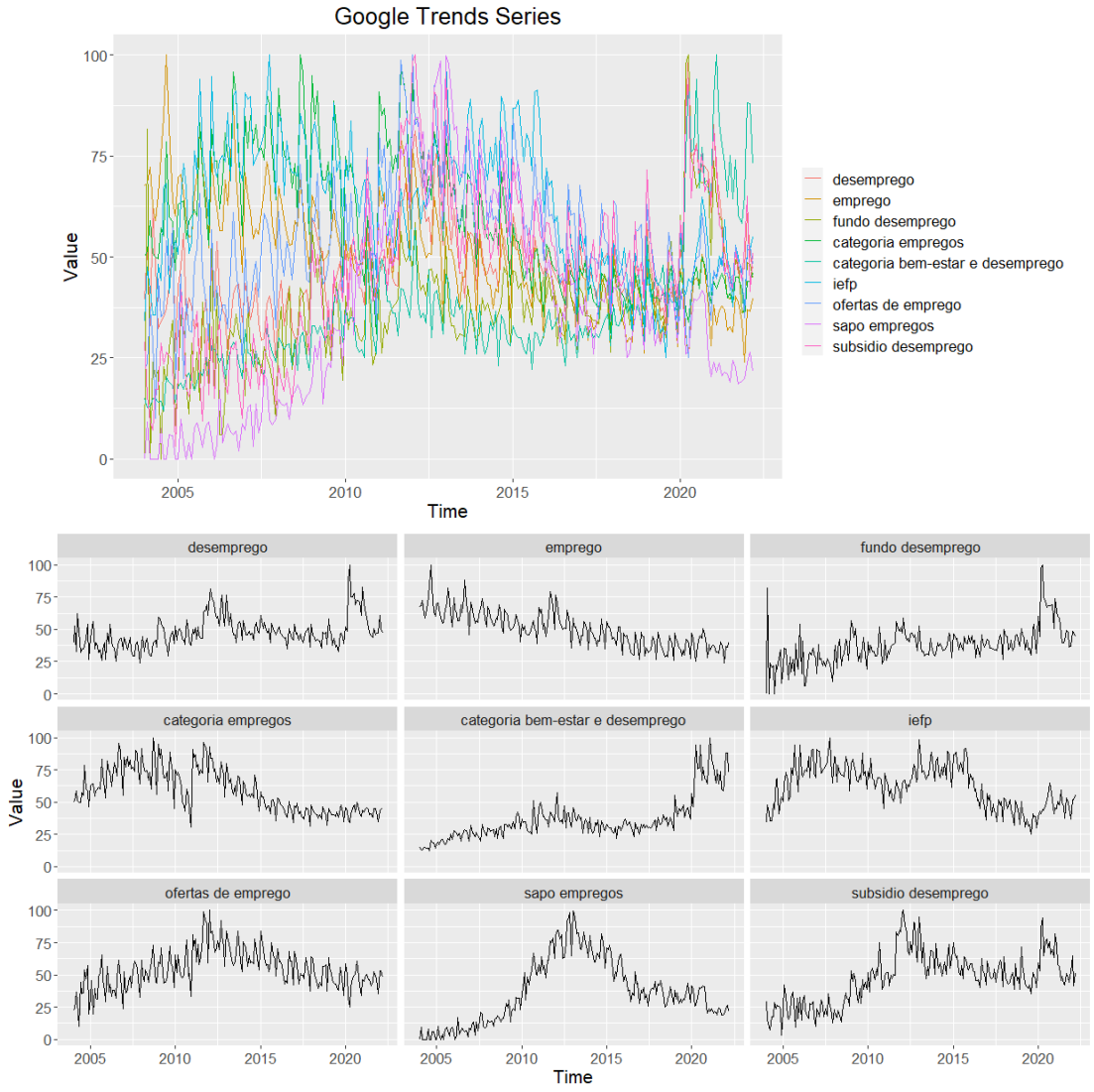



Figure 7. Time series plots of the Google Trends series.

The preliminary analysis of the unemployment rate series suggests the existence of changing trends, as illustrated in Figure 8. Interestingly, these resemble those observed for some GT series, including an overall increasing trend from 2004 to 2013. This increasing trend is accelerate in the years of 2012 and 2013, and after this period a linear decline in values is observed. Similarly to the GT series, there is a rapid increase in the unemployment rate values during the year 2020, after which a rapid decline follows. Besides, Figure 8 illustrates that the training set observations present a higher dispersion and are generally not matched by those in the test set at the 80:20 split ratio of the unemployment rate series.

Also, the presence of a unit root is detected for the unemployment rate and several GT series, thus showing the need of differencing in order to obtain stationarity (Table 7).


Figure 8. Train and test sets of the unemployment rate series considering the 80:20 split ratio, respectively.

Table 7. Estimated number of differences required to make stationary the unemployment rate and Google Trends series.

| Time Series | | Unit Root Test | |
|---|---|---|---|
| | | Raw data | Deseasonalized data |
| Google Trends | "desemprego" | 0 | 0 |
| | "emprego" | 0 | 1 |
| | "subsidio desemprego" | 0 | 0 |
| | "fundo desemprego" | 0 | 0 |
| | "ofertas emprego" | 0 | 0 |
| | "sapo empregos" | 1 | 1 |
| | "iefp" | 0 | 1 |
| | Category "empregos" | 0 | 1 |
| | Category "bem-estar & desemprego" | 1 | 1 |
| Unemployment rate | | 1 | - |

## 5.2. Forecasting

This sub-section details the forecasting performance of the benchmark and regARIMA models, as well as the various RF-based models proposed in Chapter 4. In addition, details on the accuracy results are reported and forecasting plots provided.

### 5.2.1. Benchmark and regression with ARIMA errors

The RMSE of nowcasts produced with the benchmark ARIMA and regARIMA models for the prediction time frame (August-2018 to March-2022) are displayed in Table 8. Specifically, the benchmark model shows a RMSE (0.3127) that is, approximately, half

of the regARIMA model (0.5236) that incorporates the most relevant deseasonalized GT series and their lags. Additionally, this regARIMA model shows a better RMSE (0.5236) than the corresponding model incorporating the most relevant GT series and lags (0.6191).

Besides, it is possible to identify that the benchmark is based on 44 ARIMA(1,1,1) models, while the orders of the two regARIMA models are relatively similar (Table 8).

Table 8. Model details, nowcasting results and accuracy of the benchmark and regression with ARIMA errors models.

| Final Model | Fitted Models | RMSE |
|---|---|---|
| Benchmark based on past unemployment rate values | 44 ARIMA (1,1,1) | 0.3127 |
| RegARIMA considering unemployment rate and relevant GT series and lags | 25 ARIMA (2,0,1)<br>14 ARIMA (0,0,4)<br>3 ARIMA (0,0,1)<br>1 ARIMA (2,0,2)<br>1 ARIMA (0,0,0) | 0.6191 |
| RegARIMA considering unemployment rate and relevant dGT series and lags | 27 ARIMA (2,0,1)<br>13 ARIMA (0,0,4)<br>1 ARIMA (2,0,2)<br>1 ARIMA (0,0,5)<br>1 ARIMA (0,0,3)<br>1 ARIMA (0,0,0) | 0.5236 |

### 5.2.2. Random forest

The RMSE of nowcasts produced with the RF models for the prediction time frame (August-2018 to March-2022) are displayed in Table 9 and 10. A lower RMSE is generally achieved with such ML-based models than with the benchmark and regARIMA models. Respectively, some RF models show a RMSE value that is lower than half of the benchmark model, which is the best performing one among classical models (benchmark and regARIMA models), as detailed previously in Table 8.

In accordance to the results of regARIMA models, more accurate nowcasts are obtained when RF models are built with deseasonalized GT series (Table 9). Still, for both GT and deseasonalized GT datasets, a better performance is achieved by employing feature engineering and transformations to the data, namely the incorporation of additional explanatory variables of the unemployment rate series or differencing the unemployment rate and GT series in order to obtain stationarity (Table 9). Respectively, the models that do not incorporate data transformations or feature engineering even show a worst RMSE (0.3285 and 0.3381) than the benchmark ARIMA model (0.3127).

Furthermore, among the ML-based models relying exclusively on the GT and deseasonalized GT series, the one incorporating deseasonalized GT series and

additional variables related to the unemployment rate series shows the best RMSE (0.1308) but this is closely matched (0.1317) by the model incorporating the stationary unemployment rate series and stationary deseasonalized GT series (Table 9).

Table 9. Model details and nowcasting accuracy of random forest-based models relying on GT and dGT series. Ntrees and mtry correspond to the tuned hyperparameter values of number of trees to grow and number of explanatory variables to consider at each split.

| Final Model | Tuned model | RMSE |
|---|---|---|
| Tuned model considering unemployment rate and GT series | ntree=900 mtry=9 | 0.3285 |
| Tuned model considering unemployment rate and its external variables and GT series | ntree=900 mtry=16 | 0.1388 |
| Tuned model considering stationary unemployment rate and stationary GT series | ntree=400 mtry=2 | 0.1486 |
| Tuned model considering unemployment rate and dGT series | ntree=900 mtry=9 | 0.3381 |
| Tuned model considering unemployment rate and its external variables and dGT series | ntree=400 mtry=16 | 0.1308 |
| Tuned model considering stationary unemployment rate and stationary dGT series | ntree=1500 mtry=2 | 0.1317 |

In comparison to the ML-based models relying exclusively on the GT and deseasonalized GT datasets, the RF models incorporating lagged versions of the deseasonalized GT series show a better RMSE, as displayed in Table 10. Interestingly, the accuracy of these better performing RF models is also increased when feature engineering and data transformations are applied to the unemployment rate and deseasonalized GT series (Table 10). Yet, such data transformations and feature engineering strategies do not lead to as much of a gain in RMSE as for the RF-based models relying exclusively on the GT and deseasonalized GT series (Table 9 and 10). Also important, the RF models utilizing measures of “variable importance” provided by the algorithm show a better performance than those incorporating lags by inspecting the cross-correlation functions, as represented in Table 10.

Among all RF-based models, the best performing one relies on deseasonalized GT series and their lags, and incorporation of additional variables related to the unemployment rate series, with selection of the 25 most important variables through measures of “variable importance” (Table 10). Respectively, this models shows a RMSE (0.1205) that is less than half of the benchmark (0.3127) and the regARIMA model that relies on the most relevant deseasonalized GT series and lags (0.5236). Accordingly, this RMSE value (0.1205) is also lower than the one (0.1308) originating from the best RF model among those relying exclusively on GT and deseasonalized GT series, which

also incorporates deseasonalized GT series and follows the same feature engineering strategy, i.e. incorporation of additional variables related to the unemployment series. Moreover, it should be noticed that the RMSE of the overall best performing RF model (0.1205) is closely matched by the one focusing on the stationary unemployment rate series and stationary and deseasonalized GT series and lags (0.1227), and that follows the same measuring of "variable importance" approach (Table 10). In this context, Figure 9 presents the plots of "variable importance" provided by the RF algorithm, thereby illustrating that the best performing RF model attributes most importance to the previous lags of the unemployment rate series, while the second best model (focusing on stationary unemployment rate and stationary deseasonalized GT series and lags) presents more similar values of "variable importance" across the explanatory variables considered for modelling.

Table 10. Model details and nowcasting accuracy of random forest-based models relying on dGT series and lags. Ntrees and mtry correspond to the tuned hyperparameter values of number of trees to grow and number of explanatory variables to consider at each split * Stands for models developed with "variable importance" measures provided by the random forest algorithm.

| Final Model | Tuned model | RMSE |
|---|---|---|
| Tuned model considering unemployment rate and relevant dGT series and lags | ntree=400 mtry=17 | 0.2603 |
| Tuned model considering unemployment rate and its external variables and relevant dGT series and lags | ntree=400 mtry= 21 | 0.1230 |
| Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags | ntree=400 mtry=2 | 0.1229 |
| Tuned model considering unemployment rate and relevant dGT series and lags* | ntree=1500 mtry=13 | 0.2144 |
| Tuned model considering unemployment rate and its external variables and relevant dGT series and lags* | ntree=400 mtry= 13 | 0.1205 |
| Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags* | ntree=400 mtry=2 | 0.1227 |

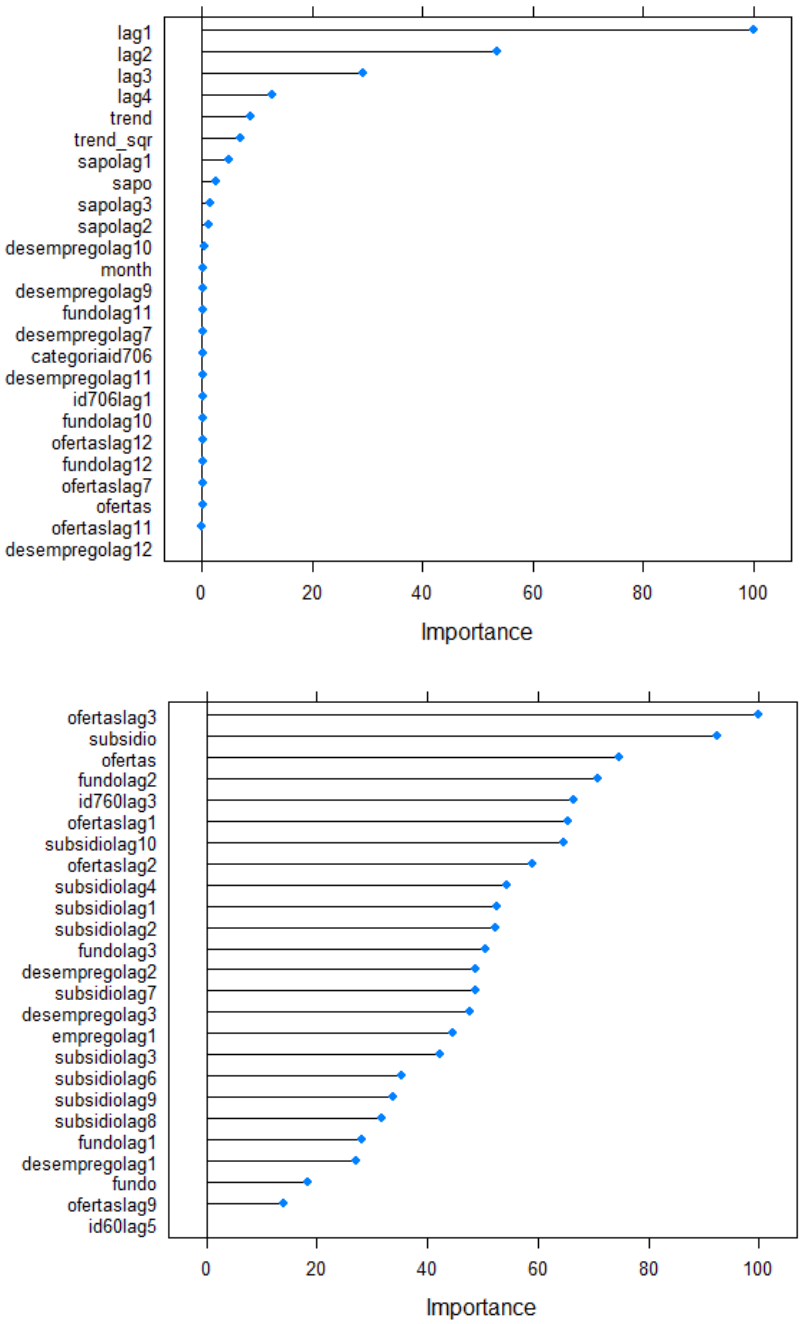



Figure 9. Variable importance measures provided by the random forest algorithm for the best model (above; focused on additional variables of the unemployment series and deseasonalized GT series and lags) and second best one (below; focused on stationary unemployment series and stationary deseasonalized GT series and lags).

### 5.2.3. Pandemic effects on accuracy

The effect of the COVID-19 pandemic on the accuracy of nowcasts produced with the benchmark and regARIMA models is displayed in Table 11, and in Table 12 and 13 for the RF-based models, through assessing models' accuracy across two prediction time

frames, i.e. the pre-pandemic (August-2018 to December-2019) and pandemic (January-2020 to March-2022) periods.

The benchmark and two regARIMA models perform better at pre-COVID times than in the period under pandemic influence (Table 11). In fact, the accuracy of classical models is very similar in the pre-COVID period, while both regARIMA models show decreasing accuracy during the COVID period in comparison to the benchmark.

Table 11. Pandemic effect on the accuracy of the benchmark and regression with ARIMA errors models. Green and red correspond to, respectively, the periods with best and worst model accuracy.

| Final Model | RMSE | Pre-COVID RMSE | COVID RMSE |
|---|---|---|---|
| Benchmark based on past unemployment rate values | 0.3127 | 0.1433 | 0.3827 |
| RegARIMA considering unemployment rate and relevant GT series and lags | 0.6191 | 0.1764 | 0.7779 |
| RegARIMA considering unemployment rate and relevant dGT series and lags | 0.5236 | 0.1679 | 0.6550 |

In agreement to classical models, almost all RF-based models present much better accuracies at pre-COVID times. Accordingly, their accuracies substantially decrease during the COVID period (Table 12 and 13).

Table 12. Pandemic effect on the accuracy of random forest-based models relying on GT and dGT series. Green and red correspond to, respectively, the periods with best and worst model accuracy.

| Final Model | RMSE | Pre-COVID RMSE | COVID RMSE |
|---|---|---|---|
| Tuned model considering unemployment rate and GT series | 0.3285 | 0.3628 | 0.3049 |
| Tuned model considering unemployment rate and its external variables and GT series | 0.1388 | 0.0634 | 0.1699 |
| Tuned model considering stationary unemployment rate and stationary GT series | 0.1486 | 0.0764 | 0.1798 |
| Tuned model considering unemployment rate and dGT series | 0.3381 | 0.3222 | 0.3478 |
| Tuned model considering unemployment rate and its external variables and dGT series | 0.1308 | 0.0694 | 0.1577 |
| Tuned model considering stationary unemployment rate and stationary dGT series | 0.1317 | 0.0630 | 0.1606 |

Besides, the overall best performing RF model (incorporates additional variables related to the unemployment rate series and deseasonalized GT series and lags) presents a slightly worst RMSE value (0.0607) in the pre-COVID period than the second best

performing RF model (0.0590), which incorporates the stationary unemployment rate and stationary and deseasonalized GT series and lags (Table 13). However, in the COVID period, the overall best performing RF model shows a lower RMSE (0.1460) than all the other ML-based models (Table 13).

Table 13. Pandemic effect on the accuracy of random forest-based models relying on dGT series and lags. Green and red correspond to, respectively, the periods with best and worst model accuracy. * Stands for models developed with "variable importance" measures provided by the random forest algorithm.

| Final Model | RMSE | Pre-COVID RMSE | COVID RMSE |
|---|---|---|---|
| Tuned model considering unemployment rate and relevant dGT series and lags | 0.2603 | 0.2200 | 0.2827 |
| Tuned model considering unemployment rate and its external variables and relevant dGT series and lags | 0.1230 | 0.0556 | 0.1507 |
| Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags | 0.1229 | 0.0689 | 0.1471 |
| Tuned model considering unemployment rate and relevant dGT series and lags* | 0.2144 | 0.1457 | 0.2482 |
| Tuned model considering unemployment rate and its external variables and relevant dGT series and lags* | 0.1205 | 0.0607 | 0.1460 |
| Tuned model considering stationary unemployment rate and relevant stationary dGT series and lags* | 0.1227 | 0.0590 | 0.1494 |

### 5.2.4. Forecasting plots

Forecasting plots represent the nowcasts produced with the benchmark model, regression with ARIMA models, and overall best performing RF model across the prediction time frame (August-2018 to March-2022), which includes the pre-pandemic (August-2018 to December-2019) and pandemic (January-2020 to March-2022) periods. Respectively, Figure 10 displays the unemployment rate values and corresponding nowcasts produced with the benchmark ARIMA and the two regARIMA models. It is suggested that the benchmark has a better nowcasting capability across the entire prediction time frame in comparison to both regARIMA models. Nevertheless, it is also clear that the two regARIMA models produced similar nowcasts to those of the benchmark across the pre-COVID period, while losing predictive power during the period affected by the pandemic.

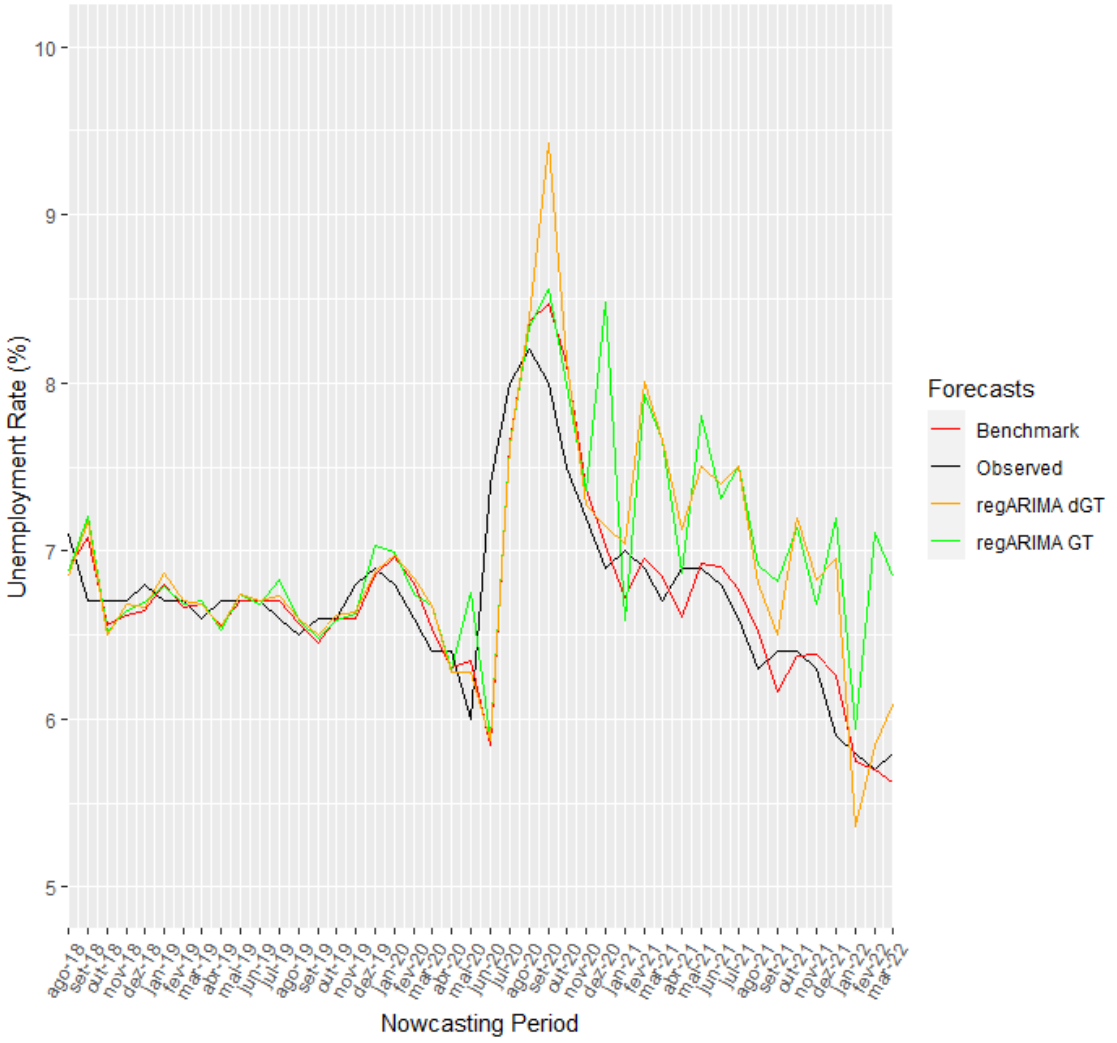



Figure 10. Plot of unemployment rate nowcasts produced with the benchmark and regression with ARIMA errors models across the prediction time frame.

Figure 11 displays the unemployment rate values and corresponding nowcasts produced with the benchmark model, and the regARIMA and RF models showing the best accuracy for the prediction time frame. Accordingly, this regARIMA model incorporates the most relevant deseasonalized GT series and lags, while the RF model relies on the 25 most relevant explanatory variables (deseasonalized GT series and lags, and additional variables related to the unemployment rate series) selected through measures of “variable importance”. The Figure illustrates the good nowcasting capabilty of the RF model, which produces more accurate nowcasts across the entire prediction time frame in comparison to the classical models. Still, the nowcasts produced with the ML and classical models are somewhat similar in the pre-COVID period, while differences in the models’ nowcasting capability become apparent during the COVID period. Respectively, Figure 11 illustrates the lower perfomance of all models in the year

2020, which is marked by a sharp increase in the unemployment rate values. The following nowcasting months show a steady decline in the unemployment rate values, with the regARIMA model losing predictive capability and the benchmark and RF model providing more accurate nowcasts.

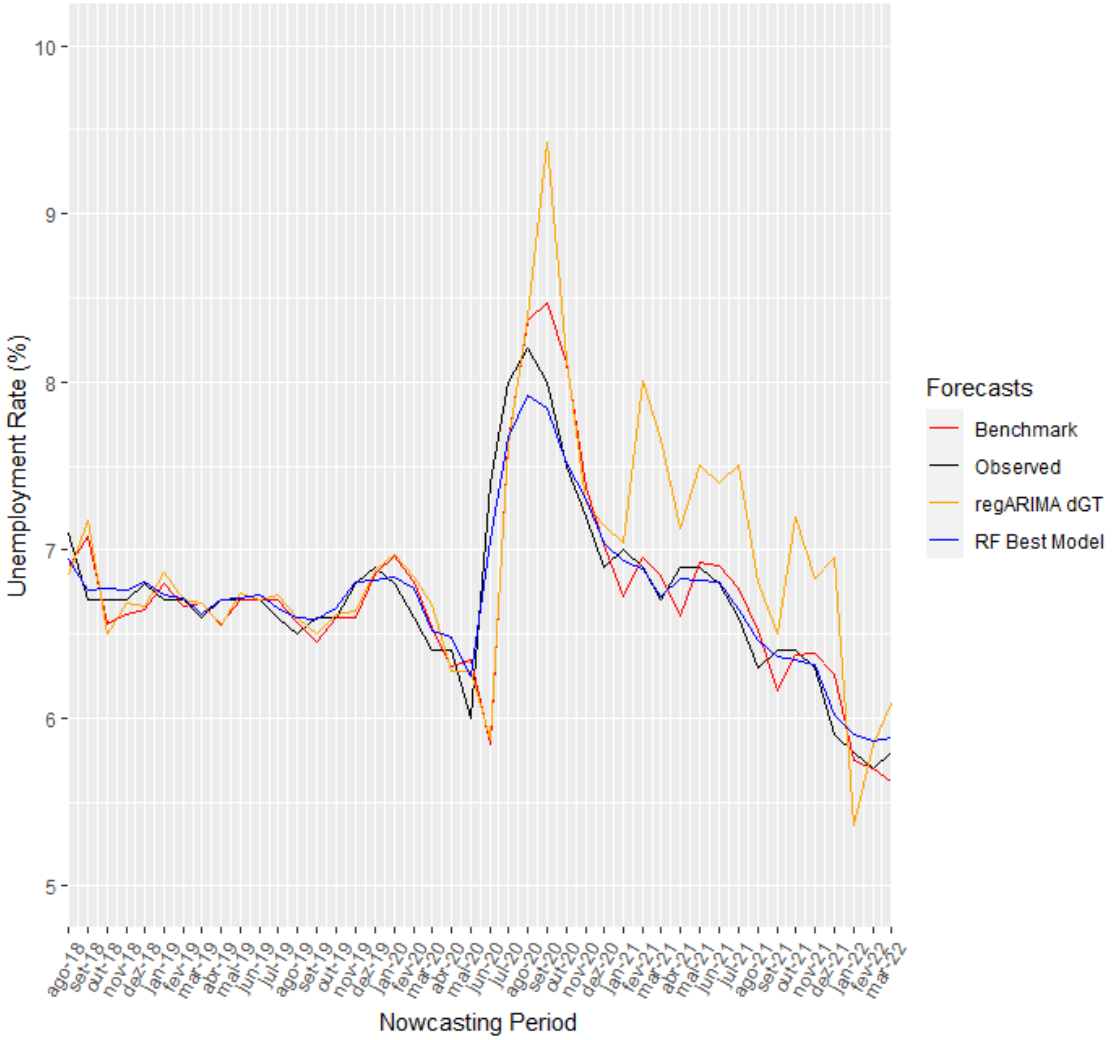



Figure 11. Plot of unemployment rate nowcasts produced with the benchmark and best regression with ARIMA errors and random forest models across the prediction time frame.

# 6. Discussion and conclusion

## 6.1. Discussion

The analysis of the GT series showed that they presented unique behaviours, despite some common patterns being sometimes identified. In specific, several Google series presented a seasonal pattern, as well as a strong trend component. Accordingly, previous studies explored the seasonality of time series collected from GT, and its effects on forecasting performance (Nagao et al. 2019; Woo and Owen 2019; Borup and Schütte 2022). In this regard, the experimental results showed that the seasonality removal from the GT series lead to more accurate unemployment nowcasts, regardless of the application of classical or ML-based models. These solid results reinforce the need for future studies to address the potential seasonality of GT series when aiming to nowcast macroeconomic variables.

Moreover, several GT and deseasonalized GT series showed unit root presence, thus demonstrating a non-stationary nature. In fact, stationarity is crucial for time series modelling by influencing how the data is perceived and predicted by both classical time series models and ML algorithms (Hyndman and Athanasopoulos 2018; Pavlyshenko 2019). In agreement, the experimental results showed that the RF models focusing on the stationary unemployment rate and stationary GT series always presented a good performance. Still, there is no consensus on the literature about the need to address the potential non-stationary nature of GT series when forecasting with ML, thus more research is urgently required on this matter (Nagao et al. 2019; Borup and Schütte 2022).

Overall, the above highlights the need to carefully address the characteristics of GT series when forecasting unemployment data, and that this can be particularly important in the context of nowcasting with ML algorithms.

Another major finding of the experimental study was that the benchmark model produced more accurate nowcasts than the two regARIMA models, which incorporated either the most relevant GT or deseasonalized GT series and their corresponding lags. First, it should be noticed that this work focused on a relative high number of GT series (9 in total); by contrast, past studies typically targeted a smaller number of keywords and categories related to the unemployment phenomenon (Barreira et al. 2013; D'Amuri and Marcucci 2017; Milani 2021). Respectively, even though the most relevant GT series (and lags) were selected for model input, the large number of GT series considered apparently failed to improve the predictive power of the regARIMA models. To address this issue in the future, it would be interesting to analyse the nowcasting performance of

regARIMA models when focusing on a smaller set of GT series, such as the ones most commonly considered in this context, e.g. “jobs” and “unemployment” (Barreira et al. 2013; D’Amuri and Marcucci 2017). Second, it was observed that the performance of the benchmark and regARIMA models was actually very similar during the pre-COVID period, with models diverging in accuracy at the time of rapid unemployment increase due to the pandemic (Almeida and Santos 2020; Caperna et al. 2020; Milani 2021). Similarly, the poorer performance of RF-based models was observed during the COVID period. Respectively, numerous studies have now demonstrated than even though the pandemic was a short period, it drastically changed the properties of time series data originating from different academic and business settings (Bobeica and Hartwig 2021; Zhang et al. 2021). As so, many models previously developed for economic forecasting were made inadequate, thus posing new challenges for the estimation of macroeconomic variables (Bobeica and Hartwig 2021; Ng 2021; Zamfir and Iordache 2022). To tackle this issue, it has been suggested that the data points around the outbreak period should be interpreted as outliers, and that forecasters should exclude them from the datasets when aiming to validate models or, in turn, interpret such values as an exogenous movement during model development (Bobeica and Hartwig 2021; Ng 2021). Accordingly, these strategies could be valuable add-ons to this dissertation by contributing to the further validation of the findings here reported.

The key results of the experimental study were that more accurate unemployment nowcasts were achieved with the RF-based models over the classical models. Similar findings have been previously reported, with some studies concluding that ML algorithms presented a greater predictive power in macroeconomic forecasting than traditional time series methods (Kreiner and Duca 2020; Gogas et al. 2021; Richardson et al. 2021). This apparent superiority of ML algorithms related to their better capacity to capture patterns and relationships in the data, thus making them particularly useful to forecast in situations of rapid changes in the economic environment (Makridakis et al. 2018) (Chapman and Desai 2021; Dauphin et al. 2022). In agreement to this, the best performing RF models showed increased nowcasting capability over the benchmark and regARIMA models across the period of rapidly increasing unemployment as a result of the COVID-19 pandemic. However, it should be noticed that the better performance of these RF-based models relied upon data transformations and feature engineering steps, including models’ incorporation of deseasonalized GT series and additional variables related to the unemployment rate series. Accordingly, it has been previously demonstrated that the predictive power of ML algorithms in different forecasting contexts was dependent upon the characteristics of the time series under study and

transformations applied to those datasets (Makridakis et al. 2018; Ensafi et al. 2022). Additionally, past studies demonstrated that feature engineering was a critical step during the development of more powerful ML-based models, by enabling the algorithms to retain critical information about the series characteristics (Feuerriegel and Gordon 2019; Krispin 2019; Bojer 2022). These factors have been considered major drawbacks in the application of ML algorithms in the time series context because they may impair the generalization and validation of findings (Makridakis et al. 2018; Bojer 2022). Adding to this, the experimental study revealed that the two best performing RF models relied on a very distinct set of explanatory GT series and lags (see the paragraph below), thus confirming the difficulty to interpret and generalize some findings reported here.

The final topic for discussion involves the comparison of the two RF models showing the most predictive power, regarding their similarities and differences. These models presented similar nowcasting accuracies, and both were developed by incorporating the 25 explanatory variables yielding the highest "variable importance" according to the RF algorithm. Accordingly, previous studies showed that variable selection for model building could be a suitable strategy when the aim is to forecast economic data with the RF algorithm and GT series as input data (Caperna et al. 2020; Borup and Schütte 2022).

In detail, the overall best RF model focused on the incorporation of deseasonalized GT series and lags, and additional variables related to the unemployment rate series, namely its most relevant past lags. Indeed, when plotting the measures of "variable importance", it was observed that the past unemployment rate lags were considered by the RF algorithm as the most relevant explanatory variables. These results further highlighted that feature engineering was critical to the good performance of this very best model (Feuerriegel and Gordon 2019; Krispin 2019). Also, this model attributed a relative high importance to the GT series "sapo empregos" and several of its lags, and this keyword is related to the Portuguese unemployment context since it is the name of a popular website among those seeking a job. This opens interesting perspectives for futures studies, especially because it has been argued that Google searches related to a specific region or country could hold the greatest predictive power (Pavlicek and Kristoufek 2015; Mihaela 2020).

On the other hand, the RF model showing the second best nowcasting accuracy followed a remarkably different modelling strategy, since it incorporated the stationary unemployment rate and stationary deseasonalized GT series and lags. Since this relative good performance was achieved by focusing exclusively on GT series and not on additional variables related to the unemployment rate series, these results then add to a

large body of literature suggesting that Google searches are an important source of time series data, by holding predictive power for forecasting economic data (Barreira et al. 2013; D'Amuri and Marcucci 2017; Borup and Schütte 2022).

## 6.2. Conclusion and future perspectives

The present dissertation aimed to clarify the contribution of Google searches, in combination with the adoption of ML algorithms, towards forecasting macroeconomic variables. To do so, an experimental study was developed with the goal of assessing the capacity of classical time series and ML-based models, both utilizing GT series, to nowcast the unemployment rate of Portugal.

Overall, the experimental results showed that the characteristics of the GT series (seasonality and stationarity) reflected in the accuracy of the proposed models. In particular, deseasonalized GT series lead to gains in accuracy, while stationarity influenced the performance of RF-based models. Accordingly, it is recommend for future studies to carefully address the properties of GT series prior to model development. Moreover, it was demonstrated that the benchmark model produced more accurate nowcasts than the two regARIMA models, which incorporated the most relevant GT and deseasonalized GT series and corresponding lags. Most likely, these results were determined by the rapid increase in the unemployment rate as a consequence of the COVID-19 pandemic, which affected the nowcasting capability of regARIMA models more profoundly. Still, both classical and RF-based models showed a poorer performance during times of COVID outbreak, thereby highlighting that future forecasting unemployment-related studies should carefully address these data points.

Finally, this dissertation demonstrated that a better nowcasting accuracy of the Portuguese unemployment rate could be achieved by employing the RF algorithm, over classical methods, i.e. benchmark ARIMA and regARIMA models. This greater predictive capacity was largely determined by the pre-processing (data transformations and feature engineering) of the target (unemployment rate) and explanatory (GT) series considered. Interestingly, the two best performing RF models followed drastically different modelling strategies, thus more research is required to understand the most promising strategies for modelling with ML in a time series context. Nevertheless, it was further confirmed that Google searches hold a great predictive power, thus being an important source of data for unemployment-related nowcasting studies. Overall, this dissertation supports the idea that combining ML with Google searches is a powerful and promising framework for nowcasting economic variables and macroeconomic outcomes.

# Appendix

This section details additional information concerning the preliminary analysis, both related to the GT and unemployment rate series.

The analysis of the ACF plot of the GT series supports the existence of seasonal or cyclic components in several of them (Figure 12). In addition, the trend component also reflects in the ACF plot for several GT series, as the ACF values decay very slowly for several of them. Moreover, the PACF plots of the GT series show more consistent patterns, namely rapid declining values followed by some significant spikes afterwards. Besides, the “seasonal” and “monthly” plots reveal that some GT series present a similar pattern of annual variation, as well as higher mean values in some months, mostly outside of the summer period (Figure 13).

Overall, these results reinforce the potential lack of stationarity of the GT series.

Moreover, the inspection of the unemployment series with its lagged values shows strong correlation for lags 1, 2, 3 and 4, and these incorporate the RF-based models, as detailed in Chapter 4.

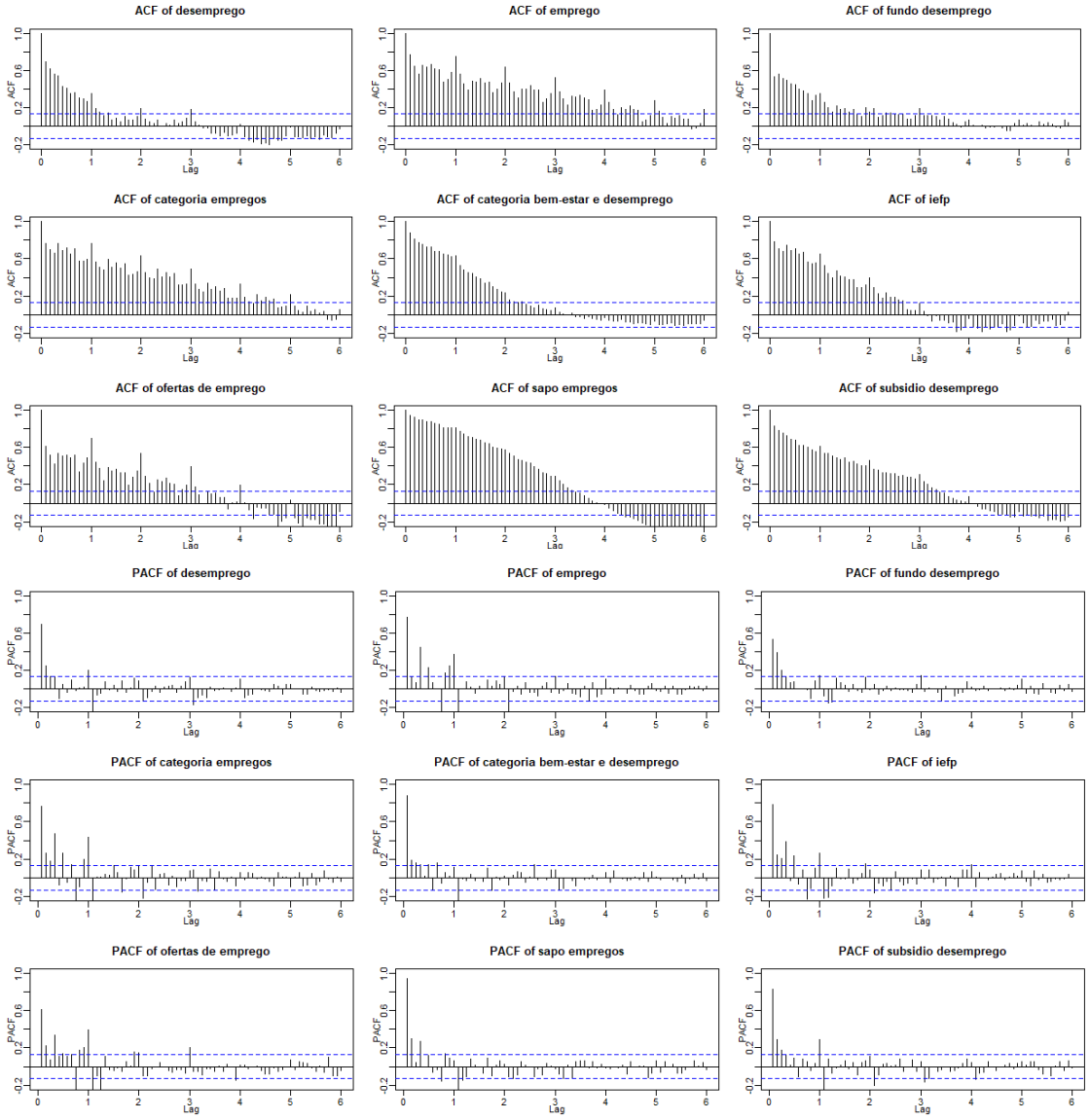



Figure 12. ACF (above) and PACF (below) plots for the GT series.

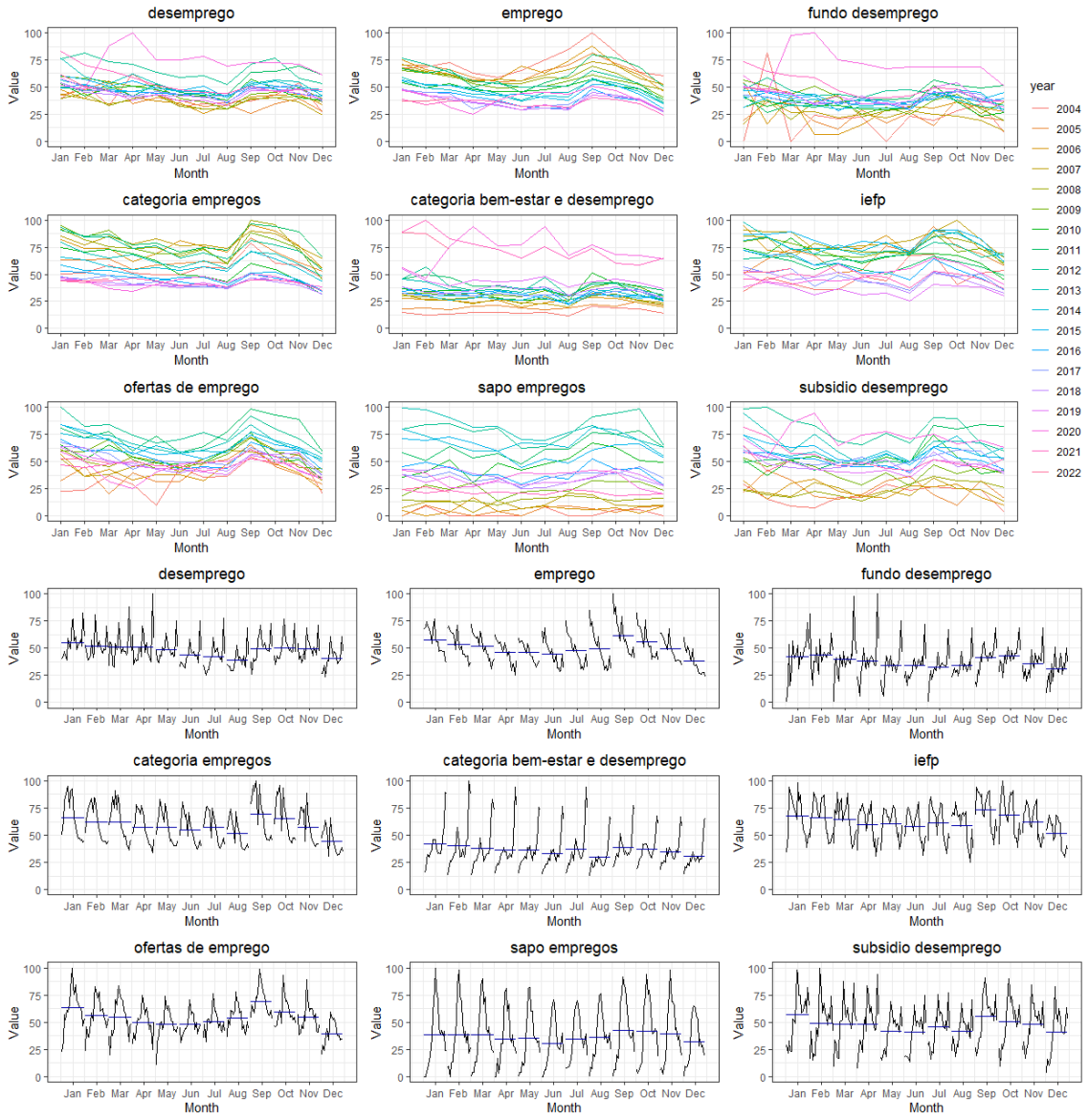



Figure 13. Seasonal (above) and monthly (below) plots for the GT series.

# References

Abraham R, El Samad M, Bakhach AM, et al (2022) Forecasting a Stock Trend Using Genetic Algorithm and Random Forest. J Risk Financ Manag 15:188. doi: 10.3390/jrfm15050188

Almeida F, Santos JD (2020) The effects of COVID-19 on job security and unemployment in Portugal. Int J Sociol Soc Policy 40:995–1003. doi: 10.1108/IJSSP-07-2020-0291

Artola C, Pinto F, Pedraza P De (2015) Can internet searches forecast tourism inflows? Int J Manpow 36:103–116. doi: 10.1108/IJM-12-2014-0259

Askitas N, Zimmermann KF (2015) The internet as a data source for advancement in social sciences. Int J Manpow 36:2–12. doi: 10.1108/IJM-02-2015-0029

Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting

Balogh A, Harman A, Kreuter F (2022) Real-Time Analysis of Predictors of COVID-19 Infection Spread in Countries in the European Union Through a New Tool. Int J Public Health 67:1604974. doi: 10.3389/ijph.2022.1604974

Bańbura M, Giannone D, Reichlin L (2010) Nowcasting. European Central Bank Working Paper

Baptista R, Thurik AR (2007) The relationship between entrepreneurship and unemployment: Is Portugal an outlier? Technol Forecast Soc Change 74:75–89. doi: 10.1016/j.techfore.2006.04.003

Barreira N, Godinho P, Melo P (2013) Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. NETNOMICS Econ Res Electron Netw 14:129–165. doi: 10.1007/s11066-013-9082-8

Berthomier L, Pradel B, Perez L (2020) Cloud cover nowcasting with deep learning. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, pp 1–6

Bijl L, Kringhaug G, Molnár P, Sandvik E (2016) Google searches and stock returns. Int Rev Financ Anal 45:150–156. doi: 10.1016/j.irfa.2016.03.015

Blanchard OJ, Summers LH (1986) Hysteresis in unemployment. National Bureau of Economic Research

Bobeica E, Hartwig B (2021) The COVID-19 shock and challenges for time series models. ECB working paper

Bojer CS (2022) Understanding machine learning-based forecasting methods : A decomposition framework and research opportunities. Int J Forecast 38:1555–1561. doi: 10.1016/j.ijforecast.2021.11.003

Bok B, Caratelli D, Giannone D, et al (2018) Macroeconomic nowcasting and forecasting with big data. Annu Rev Econom 10:615–643. doi: 10.1146/annurev-economics-080217-053214

Borup D, Schütte ECM (2022) In search of a job: Forecasting employment growth using Google Trends. J Bus Econ Stat 40:186–200. doi: 10.1080/07350015.2020.1791133

Breiman L (2001) Random forests. Mach Learn 45:5–32. doi: 10.1023/A:1010933404324

Brunner K, Cukierman A, Meltzer AH (1980) Stagflation, persistent unemployment and the permanence of economic shocks. J Monet Econ 6:467–492. doi: 10.1016/0304-3932(80)90002-1

Caperna G, Colagrossi M, Geraci A, Mazzarella G (2020) Googling unemployment during the pandemic: inference and Nowcast using search data. Joint Research Centre, European Commission

Castilho CM (2020) Time Series Forecasting with exogenous factors : Statistical vs. Machine Learning approaches. School of Economics and Management of the University of Porto

Cebrián E, Domenech J (2022) Is Google Trends a quality data source? Appl Econ Lett 30:811–815. doi: 10.1080/13504851.2021.2023088

Cerqueira V, Torgo L, Soares C (2022) A case study comparing machine learning with statistical methods for time series forecasting : size matters. J Intell Inf Syst 1–19. doi: 10.1007/s10844-022-00713-9

Chadwick MG, Sengül G (2015) Nowcasting the unemployment rate in Turkey: Let's ask Google. Research and Monetary Policy Department, Central Bank of the Republic of Turkey

Chakraborty T, Chakraborty AK, Biswas M, et al (2021) Unemployment Rate Forecasting: A Hybrid Approach. Comput Econ 57:183–201. doi: 10.1007/s10614-020-10040-2

Chapman JTE, Desai A (2021) Macroeconomic Predictions using Payments Data and Machine Learning. SSRN Electron J. doi: 10.2139/ssrn.3907281

Chinnamgari SK (2019) R Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5. Packt Publishing Ltd

Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec 88:2–9. doi: 10.1111/j.1475-4932.2012.00809.x

Cutler A, Cutler DR, Stevens JR (2009) Tree-based methods. In: High-Dimensional Data

Analysis in Cancer Research. Springer, pp 1–19

Cutler A, Cutler DR, Stevens JR (2012) Random Forests. In: Machine Learning. Springer, pp 157–175

D'Amuri F, Marcucci J (2017) The predictive power of Google searches in forecasting US unemployment. Int J Forecast 33:801–816. doi: 10.1016/j.ijforecast.2017.03.004

DataVedas (2018) Introduction To Time Series Data. https://www.datavedas.com/introduction-to-time-series-data

Dauphin MJ-F, Dybczak MK, Maneely M, et al (2022) Nowcasting GDP-A Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies. International Monetary Fund

Dilmaghani M (2019) Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate? J Econ Stud 46:422–445. doi: 10.1108/JES-11-2017-0346

Dritsakis N, Klazoglou P (2018) Forecasting Unemployment Rates in USA Using Box-Jenkins Methodology. Int J Econ Financ Issues 8:9–20

Enders W (2015) Applied econometric time series, 4th edn. Wiley

Ensafi Y, Amin SH, Zhang G, Shah B (2022) Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. Int J Inf Manag Data Insights 2:100058. doi: 10.1016/j.jjimei.2022.100058

Feng W, Sui H, Tu J, et al (2018) A novel change detection approach for multi-temporal high-resolution remote sensing images based on rotation forest and coarse-to-fine uncertainty analyses. Remote Sens 10:1015. doi: 10.3390/rs10071015

Feuerriegel S, Gordon J (2019) News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. Eur J Oper Res 272:162–175. doi: 10.1016/j.ejor.2018.05.068

Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment? Econ Model 30:117–125. doi: 10.1016/j.econmod.2012.07.017

Fornaro P, Luomaranta H (2020) Nowcasting Finnish real economic activity: a machine learning approach. Empir Econ 58:55–71. doi: 10.1007/s00181-019-01809-y

Francesco D (2009) Predicting unemployment in short samples with internet job search query data. MPRA Paper 18403, University Library of Munich, Germany

Galbraith JW, van Norden S (2019) Asymmetry in unemployment rate forecast errors. Int J Forecast 35:1613–1626. doi: 10.1016/j.ijforecast.2018.11.006

Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with applications in R. Spinger

Gogas P, Papadimitriou T, Sofianos E (2021) Forecasting unemployment in the euro

area with machine learning. J Forecast 41:551–566. doi: 10.1002/for.2824

González-Fernández M, González-Velasco C (2018) Can Google econometrics predict unemployment? Evidence from Spain. Econ Lett 170:42–45. doi: 10.1016/j.econlet.2018.05.031

Grimmer J, Roberts ME, Stewart BM (2021) Machine learning for social science: An agnostic approach. Annu Rev Polit Sci 24:395–419. doi: 10.1146/annurev-polisci-053119-015921

Haider J, Sundin O (2019) Invisible search and online search engines: The ubiquity of search in everyday life. Taylor & Francis

Haselbeck F, Killinger J, Menrad K, et al (2022) Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions. Mach Learn with Appl 7:100239. doi: 10.1016/j.mlwa.2021.100239

Hopp D (2022) Benchmarking Econometric and Machine Learning Methodologies in Nowcasting. In: United Nations Conference on Trade and Development

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice, 2nd edn. OTexts

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. J Stat Softw 27:1–22. doi: 10.18637/jss.v027.i03

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22:679–688. doi: 10.1016/j.ijforecast.2006.03.001

Ishwarappa, Anuradha J (2015) A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Comput Sci 48:319–324. doi: 10.1016/j.procs.2015.04.188

Jelena M, Ivana I, Zorana K (2017) Modeling The Unemployment Rate At The Eu Level By Using Box-Jenkins Methodology. In: EBEEC Conference Proceedings: The Economies of Balkan and Eastern Europe Countries in the Changed World. KnE Social Sciences, pp 1–13

Joiner IA (2018) Emerging library technologies: it's not just for geeks. Chandos Publishing

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Science (80- ) 349:255–260. doi: 10.1126/science.aaa8415

Jun SP, Yoo HS, Choi S (2018) Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technol Forecast Soc Change 130:69–87. doi: 10.1016/j.techfore.2017.11.009

Kane MJ, Price N, Scotch M, Rabinowitz P (2014) Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC

Bioinformatics 15:276. doi: 10.1186/1471-2105-15-276

Kapetanios G, Papailias F (2018) Big data & macroeconomic nowcasting: Methodological review. Econ Stat Cent Excell Natl Inst Econ Soc Res

Katris C (2020) Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. Comput Econ 55:673–706. doi: 10.1007/s10614-019-09908-9

Khan Jaffur ZR, Sookia NUH, Nunkoo Gonpot P, Seetanah B (2017) Out-of-sample forecasting of the Canadian unemployment rates using univariate models. Appl Econ Lett 24:1097–1101. doi: 10.1080/13504851.2016.1257208

Kirchgässner G, Wolters J, Hassler U (2012) Introduction to modern time series analysis. Springer Science & Business Media

Kompella S, Chilukuri KC (2019) Stock Market Prediction Using Regression. Int Res J Eng Technol 10:20–30

Kreiner A, Duca J (2020) Can machine learning on economic data better forecast the unemployment rate? Appl Econ Lett 27:1434–1437. doi: 10.1080/13504851.2019.1688237

Krispin R (2019) Hands-On Time Series Analysis with R: Perform Time Series Analysis and Forecasting Using R. Packt Publishing Ltd

Krollner B, Vanstone BJ, Finnie GR (2010) Financial time series forecasting with machine learning techniques: a survey. In: European Symposium on Artificial Neural Networks: Computational and Machine Learning. Bruges, Belgium

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26. doi: 10.18637/jss.v028.i05

Kundu S, Singhania R (2020) Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data. Int J Trade, Econ Financ 11:135–140. doi: 10.18178/ijtef.2020.11.6.679

Lai H, Khan YA, Thaljaoui A, et al (2021) COVID-19 pandemic and unemployment rate: A hybrid unemployment rate prediction approach for developed and developing countries of Asia. Soft Comput 6:615. doi: 10.1007/s00500-021-05871-6

Lasso F, Snijders S (2016) The power of Google search data; an alternative approach to the measurement of unemployment in Brazil. Student Undergrad Res E-journal

Liu J, Li J, Li W, Wu J (2016) Rethinking big data: A review on the data quality and usage issues. ISPRS J Photogramm Remote Sens 115:134–142. doi: 10.1016/j.isprsjprs.2015.11.006

Macassa G, Rodrigues C, Barros H, Marttila A (2021) Experiences of involuntary job loss and health during the economic crisis in Portugal. Porto Biomed J 6:e121. doi:

10.1097/j.pbj.0000000000000121

Mahalakshmi G, Sridevi S, Rajaram S (2016) A survey on forecasting of time series data. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). IEEE, pp 1–8

Makridakis S (2017) The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures 90:46–60. doi: 10.1016/j.futures.2017.03.006

Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One 13:e0194889. doi: 10.1371/journal.pone.0194889

Marques MM (1999) Attitudes and threat perception: unemployment and immigration in Portugal. South Eur Soc Polit 4:184–205. doi: 10.1080/13608740408539585

Mavragani A, Gkillas K (2020) COVID-19 predictability in the United States using Google Trends time series. Sci Rep 10:1–12. doi: 10.1038/s41598-020-77275-9

McLaren N, Shanbhogue R (2011) Using internet search data as economic indicators

Metcalfe A V, Cowpertwait PSP (2009) Introductory time series with R. Springer Science & Business Media

Mihaela S (2020) Improving unemployment rate forecasts at regional level in Romania using Google Trends. Technol Forecast Soc Change 155:120026. doi: 10.1016/j.techfore.2020.120026

Milani F (2021) COVID-19 outbreak, social response, and early economic effects: a global VAR analysis of cross-country interdependencies. J Popul Econ 34:223–252. doi: 10.1007/s00148-020-00792-4

Mitchell TM (1997) Does Machine Learning Really Work? AI Mag 18:11. doi: 10.1609/aimag.v18i3.1303

Montgomery DC, Jennings CL, Kulahci M (2015) Introduction to time series analysis and forecasting. John Wiley & Sons

Morales EF, Zaragoza JH (2012) An introduction to reinforcement learning. In: Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions. IGI Global, pp 63–80

Muth JF (1961) Rational expectations and the theory of price movements. Econom J Econom Soc 29:315–335. doi: 10.2307/1909635

Naccarato A, Falorsi S, Loriga S, Pierini A (2018) Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technol Forecast Soc Change 130:114–122. doi: 10.1016/j.techfore.2017.11.022

Naccarato A, Pierini A, Falorsi S (2015) Using Google Trend data to predict the Italian unemployment rate. Paper No. 203, Departmental Working Papers of Economics -

University “Roma Tre”

Nagao S, Takeda F, Tanaka R (2019) Nowcasting of the U.S. unemployment rate using Google Trends. Financ Res Lett 30:103–109. doi: 10.1016/j.frl.2019.04.005

Naing WYN, Htike ZZ (2015) Forecasting of monthly temperature variations using random forests. ARPN J Eng Appl Sci 10:10109–10112

Ng S (2021) Modeling macroeconomic variations after COVID-19. National Bureau of Economic Research

Nian R, Liu J, Huang B (2020) A review on reinforcement learning: Introduction and applications in industrial process control. Comput Chem Eng 139:106886. doi: 10.1016/j.compchemeng.2020.106886

Pacifico D, Thévenot C (2016) Faces of joblessness in Portugal: anatomy of employment barriers. OECD

Parray IR, Khurana SS, Kumar M, Altalbe AA (2020) Time series data analysis of stock price movement using machine learning techniques. Soft Comput 24:16509–16517. doi: 10.1007/s00500-020-04957-x

Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with google searches: Evidence from the Visegrad Group countries. PLoS One 10:e0127084. doi: 10.1371/journal.pone.0127084

Pavlyshenko BM (2019) Machine-learning models for sales time series forecasting. Data 4:15. doi: 10.3390/data4010015

Pessoa E, Bárbara C, Viegas L, et al (2020) Factors associated with in-hospital mortality from community-acquired pneumonia in Portugal: 2000-2014. BMC Pulm Med 20:18. doi: 10.1186/s12890-019-1045-x

Plessz M, Ezdi S, Airagnes G, et al (2020) Association between unemployment and the co-occurrence and clustering of common risky health behaviors: Findings from the Constances cohort. PLoS One 15:e0232262. doi: 10.1371/journal.pone.0232262

Richardson A, van Florenstein Mulder T, Vehbi T (2021) Nowcasting GDP using machine-learning algorithms: A real-time assessment. Int J Forecast 37:941–948. doi: 10.1016/j.ijforecast.2020.10.005

Rovetta A (2021) Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research. Front Res Metrics Anal 6:670226. doi: 10.3389/frma.2021.670226

Royo S (2010) Portugal and Spain in the EU: paths of economic divergence (2000-2007). In: Análise Social. JSTOR, pp 209–254

Şahin A, Tasci M, Yan J (2020) The Unemployment Cost of COVID-19: How High and How Long? Econ Comment (Federal Reserv Bank Cleveland). doi: 10.26509/frbc-

ec-202009

Shcherbakov MV, Brebels A, Shcherbakova NL, et al (2013) A survey of forecast error measures. World Appl Sci J 24:171–176. doi: 10.5829/idosi.wasj.2013.24.itmies.80032

Shumway RH, Stoffer DS (2017) Time Series Analysis and Its Applications: with R Examples, 4th edn. Springer

Simionescu M, Cifuentes-Faura J (2022) Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. J Policy Model 44:1–21. doi: 10.1016/j.jpolmod.2021.09.011

Simionescu M, Streimikiene D, Strielkowski W (2020) What does google trends tell us about the impact of brexit on the unemployment rate in the UK? Sustainability 12:1–10. doi: 10.3390/su12031011

Simionescu M, Zimmermann KF (2017) Big data and unemployment analysis. Discussion Paper Series 81, Global Labor Organization (GLO)

Stockhammer E, Sturn S (2012) The impact of monetary policy on unemployment hysteresis. Appl Econ 44:2743–2756. doi: 10.1080/00036846.2011.566199

Team RC (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org/

Tiao GC, Box GEP (1981) Modeling multiple time series with applications. J Am Stat Assoc 76:802–816. doi: 10.1080/01621459.1981.10477728

Tsay RS (2005) Analysis of Financial Time Series, 2nd edn. John wiley & sons

Tyralis H, Papacharalampous G (2017) Variable selection in time series forecasting using random forests. Algorithms 10:114. doi: 10.3390/a10040114

Vamathevan J, Clark D, Czodrowski P, et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477. doi: 10.1038/s41573-019-0024-5

Varian HR (2014) Big data: New tricks for econometrics. J Econ Perspect 28:3–28. doi: 10.1257/jep.28.2.3

Vicente MR, López-Menéndez AJ, Pérez R (2015) Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technol Forecast Soc Change 92:132–139. doi: 10.1016/j.techfore.2014.12.005

WMO (2017) Guidelines for nowcasting techniques. World Meteorological Organization, Geneva, Switzerland

Woo J, Owen AL (2019) Forecasting private consumption with Google Trends data. J Forecast 38:81–91. doi: 10.1002/for.2559

Wu L, Brynjolfsson E (2015) The future of prediction: How Google searches foreshadow housing prices and sales. In: Economic analysis of the digital economy. University of Chicago Press, pp 89–118

Yoon J (2021) Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. Comput Econ 57:247–265. doi: 10.1007/s10614-020-10054-w

Zamfir IC, Iordache AMM (2022) The influences of covid-19 pandemic on macroeconomic indexes for European countries. Appl Econ 54:4519–4531. doi: 10.1080/00036846.2022.2031858

Zeileis A, Hothorn T (2002) Diagnostic checking in regression relationships. R News 2:7–10

Zhang H, Song H, Wen L, Liu C (2021) Forecasting tourism recovery amid COVID-19. Ann Tour Res 87:103149. doi: 10.1016/j.annals.2021.103149