Surveillance of Tuberculosis by analysing Google Trends

Luana Gorgueira Santos

152021017

Dissertation written under the supervision of Professor Pedro Afonso Fernandes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at Católica-Lisbon School of Business & Economics

1st June 2023

Surveillance of Tuberculosis by analysing Google Trends

Luana Gorgueira Santos

Abstract

Tuberculosis remains a global health concern, having caused around 1.5 million deaths in 2020. The Portuguese medical authorities are facing challenges to meet the goals of the Word Health Organization in the area of Tuberculosis. Early detection of potential Tuberculosis outbreaks is crucial for effective intervention and control, but traditional surveillance systems often suffer from reporting lags and resource limitations, which were aggravated by the COVID-19 pandemic. This thesis explores the potential of using Google Trends to predict Tuberculosis incidence in Portugal. Past research have shown promising results in this area, suggesting that Google Trends search volume could complement existing surveillance methods. To improve Tuberculosis surveillance system, we developed a syndromic approach using 19 Tuberculosis-related terms extracted from Google Trends. Historical data on the incidence of Tuberculosis was extracted from the European Centre for Disease Prevention and Control. After joining both datasets, we applied different machine learning models to forecast the monthly Tuberculosis incidence. Nextly, four accuracy metrics, including the Akaike Information Criterion, were used to select the best predictive model. Our empirical analysis shows that the forecast matches the seasonal patterns of Tuberculosis incidence in Portugal. While there are possible limitations that need to be addressed in future research, the surveillance system developed in this study might be a valuable tool for public health authorities as it provides real-time information on potential new cases. In the long run, this system might help alleviate the burden of Tuberculosis and potentially mitigate future outbreaks.

Keywords: Tuberculosis surveillance; Google Trends; Search volume; Machine learning; Accuracy metrics; Tuberculosis incidence forecasting; Portugal.

Vigilância da Tuberculose por análise do Google Trends

Luana Gorgueira Santos

Resumo

A Tuberculose continua a ser um problema de saúde global, tendo causado cerca de 1,5 milhões de mortes em 2020. As autoridades médicas portuguesas enfrentam desafios para cumprir as metas da Organização Mundial de Saúde na área da Tuberculose. A deteção antecipada de potenciais surtos de tuberculose é crucial para uma intervenção eficaz. No entanto, os sistemas de vigilância tradicionais sofrem frequentemente de atrasos nos relatórios e limitações de recursos, agravados pela pandemia COVID-19. Esta tese explora o potencial uso do Google Trends para prever a incidência da Tuberculose em Portugal. Estudos anteriores mostraram resultados promissores nesta área, sugerindo que o volume de pesquisa do Google Trends pode complementar os métodos de vigilância existentes. Para melhorar o sistema de vigilância da Tuberculose, desenvolvemos uma abordagem sindrômica usando 19 termos relacionados à Tuberculose extraídos do Google Trends. Dados históricos sobre a incidência de Tuberculose foram extraídos do Centro Europeu de Prevenção e Controle das Doenças. Aplicámos diferentes modelos de Machine learning para prever a incidência mensal de Tuberculose. Em seguida, quatro métricas de precisão, inclusive o Critério de Informação de Akaike, foram usadas para selecionar o melhor modelo de previsão. A análise mostra que a previsão corresponde aos padrões sazonais de incidência da Tuberculose em Portugal. Embora existam possíveis limitações neste estudo, o método de vigilância desenvolvido poderá ser uma ferramenta indispensável para a saúde pública, pois fornece informações em tempo real sobre possíveis novos casos. A longo prazo, este sistema poderá ajudar a mitigar a Tuberculose em Portugal.

Palavras-chave: Vigilância da Tuberculose; Google Trends; Volume de pesquisa; Machine learning; Métricas de precisão; Previsão da incidência da Tuberculose; Portugal.

Acknowledgments

I would like to take this opportunity to thank everyone who supported me with my thesis and throughout this academic journey.

First and foremost, this project would not have been possible without my thesis supervisor Pedro Afonso Fernandes who constantly supported me throughout the process of writing my thesis.

A special thanks to my mother, Leonor Santos, and my father, Carlos Santos, for providing me the opportunity to study at Universidade Católica Portuguesa and giving me the best guidance on my academic, professional, and personal life.

I must express my deepest gratitude to my partner, Pedro Oliveira, for providing me with unfailing emotional support and continuous encouragement throughout the last 5 years of study and through the process of writing this thesis.

I would like to thank the first person I met at Universidade Católica Portuguesa, Mariana Felizardo, who is not only one of my best friends today but also someone who always supported me throughout my academic career since day one.

And last but not least, I would like to thank my dear friend Rui Pereira. Thank you for never making me forget my professional goals and for providing me with your medical experience and knowledge in this research.

Thank you.

Luana Santos

Contents

Acronyms

AIC Akaike Information Criterion. 19, 20, 24

BCG Bacillus Calmette–Guérin. 3, 5

CDC US Centers for Disease Control and Prevention. 1, 6, 8, 9

CV Cross-Validation. 13, 14

DGS Direção-Geral da Saúde. 26–28

DNN Deep Neural Network. 14, 17, 24

GT Google Trends. 1, 2, 7–12, 14, 16, 26–28, 36

HPCs high-priority countries. 4, 5

LASSO Least Absolute Shrinkage and Selection Operator. 14–16, 24

MAE Mean Absolute Error. 18, 19, 24

ML Machine Learning. 11, 14, 17, 24, 26, 28

MLP Multilayer Perceptron. 17, 18

MSE Mean Squared Error. 19, 24

NTSS National Tuberculosis Surveillance System. 6

OLS Ordinary Least Squares. 14–16, 24

PLS Partial Least Squares. 14, 16, 24, 26

PTB Pulmonary Tuberculosis. 4, 12

RMSE Root Mean Squared Error. 19, 24

STL Seasonal-trend LOESS. 4

SVM Support Vector Machine. 14, 17, 24

TB Tuberculosis. iv, 1–14, 16, 20–28, 34–36

WHO World Health Organization. iv, 1, 3, 5

List of Figures

List of Tables

List of Equations

1 Introduction

1.1 Context and Motivation

Google Trends (GT) is a freely accessible tool that allows individuals to interact with search volume data, which may provide deep insights into health-related phenomena (Nuti et al., 2014). Over the past ten years, there has been a significant increase in the application of GT for research purposes. In the process, the focus of research has transitioned from analysing and diagnosing research trends to predicting and anticipating changes (Jun et al., 2018). In recent years, GT has gained traction as a novel surveillance tool for various diseases, including measles, dengue fever, and influenza.

Currently, Tuberculosis (TB) is still a major public health concern (Chapman et al., 2013). Further accelerated by COVID-19, the growth of cases of TB causes a major disruption to the traditional surveillance systems. Moreover, these systems present several limitations, such as time-consuming and resource-intensive, causing a gap between traditional and modern systems (Chapman et al., 2013). For instance, US Centers for Disease Control and Prevention (CDC) announce the TB morbidity surveillance results with months of reporting lag (Zhou et al., 2011). Even the Portuguese medical authorities are facing challenges to meet the goals of the World Health Organization (WHO) in the area of TB and consider that it is crucial to accelerate the decrease of new cases (Lusa, 2023). Therefore, more efficient surveillance methods are required to monitor the disease. From this perspective, the National Academy of Medicine, an American non-governmental organization, acknowledges the potential of using internet data in healthcare research, recognizing that it may serve as a valuable supplement to, and extension of, the current data foundations that are in place (Nuti et al., 2014). In this sense, GT may enable researchers to identify TB patterns and trends with terms related to the disease.

The use of digital data sources such as GT for disease surveillance is a new developing topic. Besides, applying GT offers a global perspective on TB trends, making it beneficial for monitoring TB incidence and prevalence in areas with limited surveillance infrastructure or resources. For instance, by analysing regions with higher search volumes related to TB through GT, public health departments can plan interventions to improve awareness, diagnosis, and treatment in those regions. Additionally, GT is a cost-effective tool since it does not require additional resources or data collection. Therefore, this topic might be interesting for researchers as they can gather valuable information on TB trends and patterns by analysing Google search

data without incurring extra costs.

In light of this, this study aims to verify whether GT can be effectively used to monitor TB trends and predict further outbreaks of the disease. In that case, the results of this research might improve and advance the current TB surveillance system.

1.2 Research Goal and Research Question

This study intends to overcome some of the previous limitations by testing a different approach on TB surveillance trough the analysis of GT. In this sense, this project aims explore whether GT is an accurate tool to be used alongside traditional surveillance methods to enhance TB monitoring in a monthly basis.

By analysing search patterns and comparing them to traditional surveillance methods, we aim to shed light on the strengths and limitations of this approach and provide insights into its potential for use in other settings. Within the body of literature that references GT, this study focuses particularly on the improvement of the early detection of TB incidence in Portugal. To investigate this topic, the following research question is raised: Can GT predict the incidence of TB in Portugal?

To answer the research question, the analysis will be conducted by assessing trends in search volume for TB-related terms and comparing them to real-time surveillance data. With this in mind, one hypothesis was developed as follows: There is a relation between existing TB cases and contemporary TB-related searches.

1.3 Dissertation Outline

The subsequent sections of this paper are organized as follows. Firstly, the follow-up section reviews the academic literature explained in sub-topics leading to an overall understanding of GT on analysing TB surveillance. Once the academic framework is settled, section 3 describes the data collection and presents the methodology and methods used for our analysis. Section 4 shows the empirical results of TB incidence forecasting in Portugal. Upon the presentation of the results, the subsequent section is devoted to the discussion of the findings. The final section of this study is composed by its conclusion and further developments.

2 Literature Review

2.1 Evolutionary History of Tuberculosis

To introduce the main concept of this research, we will start by explaining some minor aspects of the history behind TB. The latter is caused by the bacterium *Mycobacterium tuberculosis* that spreads through the air and most often affect the lungs (Mohajan, 2015). Accordingly, there have been several major TB outbreaks throughout history, but one of the most significant was the epidemic that occurred in North America and Europe in the late 19th century (Daniel, 2006).

Several studies refer to TB as a highly infectious respiratory disease that is still prevalent in many parts of the world, especially among the developing countries, with the greatest burden in Sub-Saharan Africa and Southeast Asia (Mohajan, 2015). Over the past decades, significant progress has been made in the global fight against the disease due to the expanded access to TB diagnosis and treatment, and the development of TB new vaccines.

Nevertheless, the COVID-19 pandemic has reversed years of this global progress and for the first time in over a decade, TB deaths have increased in certain countries, according to the WHO's 2021 Global TB report. The WHO (2021a) also stated that there was about 1.5 million deaths from TB in 2020. Most people who develop the disease are adults, men accounted for 56.5%, women accounted for 32.5% and children for 11% in 2021 (WHO, 2022). This situation became worse in the last years because many countries have reallocated their human, financial, and other resources from TB prevention in response to COVID-19. As a result, this limited the availability of many essential services (WHO, 2021a).

In the last decade, Portugal has seen a decrease of about 40% in the incidence rate of TB, with incidence values below 20 per 100 000 population (DGS, 2018). As a result, the vaccination strategy with Bacillus Calmette–Guérin (BCG) changed from universal to a selective strategy in 2016, that is, taking BCG is no longer mandatory for all children (Euronews, 2021), as long as they comply with the eligibility criteria for vaccination. Nonetheless, due to the weight of COVID-19 pandemic, there was an annual percentage increase from 3.9% in the period 2015-2019 to 8.6% in the period 2016-2020, even with the significant drop in the TB incidence rate in 2020 (DGS, 2022), see Figure 1.

Regarding the areas with the highest incidence in Portugal, Lisbon, Porto and Península de

Setúbal stand out, with the high rates being strongly linked to socioeconomic factors, such as increased unemployment and immigration rates (Areias et al., 2015). It was shown that Lisbon and Porto are the cities that most contribute to seasonal peaks of TB due to their demographic density (Bras et al., 2014). The previous study also predicted the monthly incidence rates for Pulmonary Tuberculosis (PTB) in Portugal using Seasonal-trend LOESS (STL) decomposition to model trend and seasonality. The time series showed a downward trend in and seasonality of PTB diagnosis, with a peak in March and a trough in December, which can be expected to persist if past conditions continue. Bras et al. (2014) also concluded that the mean seasonal amplitude was consistently higher in males and in adults aged between 25 and 54 years.

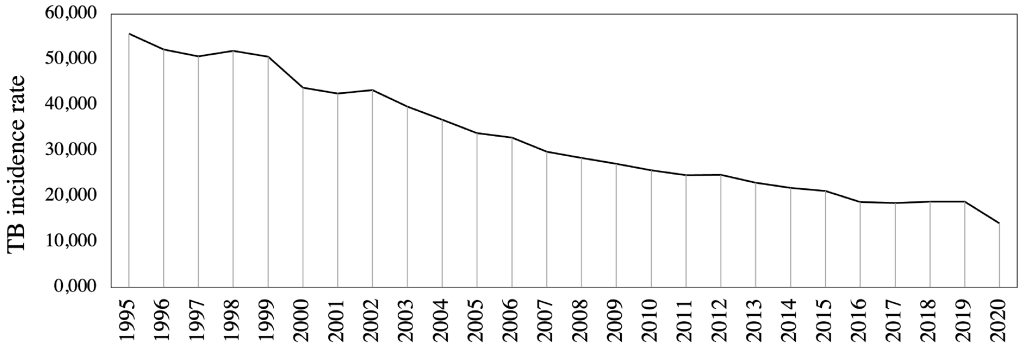

Figure 1: Evolution of TB incidence rate in Portugal, 1995-2020. Source: ECDC Atlas.

Until December 2017, DGS (2018) also showed that 64% of cases were men and the average age of the population that contracted the disease was 50.2 years. Although in Portugal most cases of TB occur in the native population, the proportion of cases in immigrants continues to increase, corresponding to 27.4% of the total in 2020. In that same year, the incidence rate in immigrants was 60.5 per 100 000 population, that is, about four times higher than the general population. Angola is the main contributor to the number of cases of TB in immigrants, followed by Guinea-Bissau, Brazil and Cape Verde (DGS, 2022).

Taking 2021 as an example, the incidence rate [1] in the 18 high-priority countries (HPCs) [2] is almost twice as high as for the Region overall (33 cases per 100 000 versus 18 cases per 100 000 for the European Region) (WHO, 2023), see Figure 2. In this context, HPCs represent high

[1]By definition, incidence rate of TB is the reported number of new and relapse TB cases in a given year, expressed as the rate per 100 000 population (WHO, 2021b, p.7).

[2]HPCs includes 18 countries, namely Armenia, Azerbaijan, Belarus, Bulgaria, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, the Republic of Moldova, Romania, the Russian Federation, Tajikistan, Türkiye, Turkmenistan, Ukraine and Uzbekistan (WHO, 2021b).

incidence areas, defined as areas with estimated incidence of $\geq$ 20 cases of TB per 100 000 population (Mohajan, 2015, p.6). Although Portugal is excluded from HPCs, it still has a TB incidence rate of 14.6 per 100 000 population in 2021 (WHO, 2023). For instance, comparing to Spain, which has a incidence rate of 7.4 in the same year, Portugal turns out to be an area with relatively more notified cases of TB (WHO, 2023).

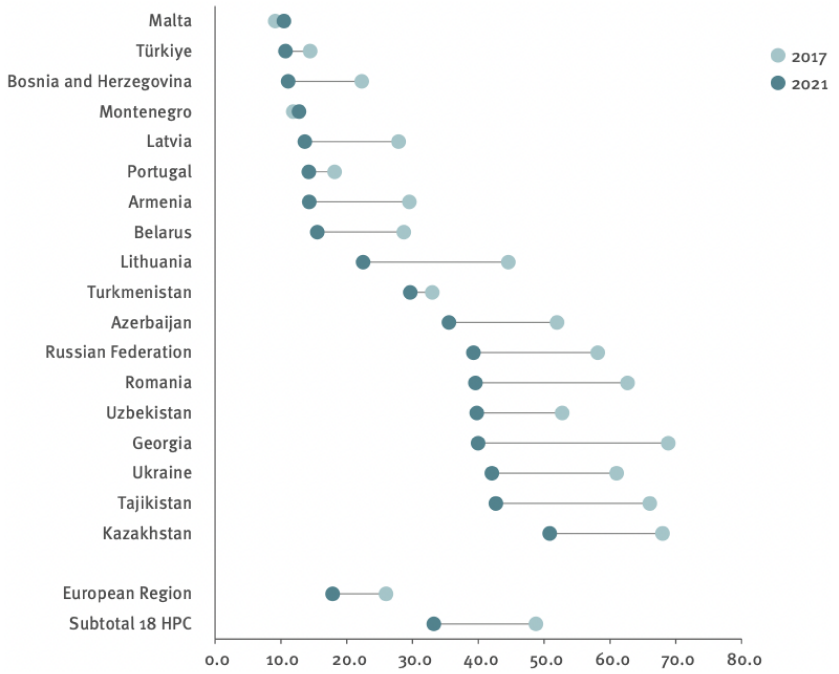

Figure 2: Trend in TB incidence rate for countries in the WHO European Region, 2017/2021. Source: WHO (2021a)

Despite many countries in the WHO European Region reporting a year-on-year increase in the number of notified TB cases, the Region still recorded 23% fewer TB cases in 2021 than in 2019. However the current rate of decline will not be sufficient to meet targets under the United Nations Sustainable Development Goals for ending the TB epidemic by 2030 (WHO, 2023).

Under these circumstances, the WHO recommends to increase new investments and develop multisectoral actions to tackle the causes that trigger TB epidemics as well as the need for new diagnostics, drugs, and vaccines (WHO, 2021a). As BCG is the only existing vaccine to fight TB, scientists argue that an improved vaccine could help achieve the goal of eradicating the disease by 2030 (Euronews, 2021).

2.2 Traditional methods on Tuberculosis Surveillance

As explained beforehand, TB is still considered, in several countries, a notifiable disease, meaning that healthcare providers need to constantly report all cases to public health authorities. This data is used to monitor trends in TB incidence and prevalence over time.

To implement effective strategies to detect and fight TB, its continuous surveillance plays a crucial role. Health surveillance is defined by ”the ongoing systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know” (Abat et al., 2016). While monitoring is defined by Thrusfield (2007) as ”the making of routine observations on health, productivity and environmental factors and the recording and transmission of these observations”. Following these terms, surveillance is a more intensive form of data recording than monitoring (Thrusfield, 2007).

Disease-specific surveillance is the traditional form of surveillance which involves tracking diseases through laboratory testing and reporting. The data gathered from these reports is then analysed and studied by public health departments to identify trends and patterns in disease occurrence and transmission (Abat et al., 2016). Laboratory testing provide a complete diagnosis and identifies the specific strain of a pathogen, which is useful to follow global trends of surveyed pathogens and infectious diseases (Abat et al., 2016). The authors also mention an example of a traditional surveillance system, namely the National Tuberculosis Surveillance System (NTSS), which records each newly case of TB in the United States since 1953. Succinctly, if a patient tests positive for *Mycobacterium tuberculosis*, local health departments anonymously report it to the NTSS (Abat et al., 2016). Afterwards, NTSS summarize and communicate the previous data to the CDC using its supplied software.

However, the traditional surveillance approach has its limitations. One of the primary challenges is the time lag between a patient’s diagnosis and reporting of the disease. This delay can vary from days to weeks, which can limit the efficacy of early interventions and augment the possibility of further transmission (Sreeramareddy et al., 2009). For instance, traditional surveillance systems, such as those employed by the CDC in the United States, typically detect TB several weeks after its occurrence (Zhou et al., 2011).

Another limitation is the dependence on timely and accurate reporting by healthcare providers, which can be influenced by factors such as a lack of resources or competing priorities (Effler

et al., 1999). For instance, a delayed reporting can result in incomplete or inaccurate data, compromising the effectiveness of surveillance systems (Jajosky and Groseclose, 2004).

Prior studies report that the traditional surveillance system is also limited by its capacity to spot emerging or unknown diseases, as it relies on prior knowledge of specific pathogens or diseases (Abat et al., 2016). As a result, this can make it difficult to identify new or novel diseases, which can delay the implementation of effective control measures.

Furthermore, the traditional surveillance system can be relatively resource intensive, requiring significant investments in laboratory facilities, staff, and equipment. This can be a challenge in low income regions or during periods of budget constraints (May et al., 2009).

In general, while the traditional approach can yield valuable information on TB surveillance, it has limitations that need to be addressed to ensure effective control and prevention of infectious diseases. Some authors suggest that a combination of disease-specific surveillance systems and syndromic surveillance systems seems to be increasingly feasible with the remarkable expansion of real-time data worldwide (Abat et al., 2016).

2.3 A Web-based tool for real-time Surveillance of Disease Outbreaks

Since the emergence of the World Wide Web in the early 1990s, technological advances have revolutionised several fields (Chapman et al., 2013). Millions of individuals search for online health-related information on a daily basis, making Web search queries a valuable source of information on collective health trends (Jun et al., 2018).

The surveillance of infectious diseases needs to be more global and instantaneous. This is a requirement for the detection and report of abnormal events related to these type of diseases (Abat et al., 2016). In the process, Google released a free sophisticated Web-based tool, named GT, for near real-time detection of disease outbreaks. This tool allows researchers to track and analyse internet search activity (Carneiro and Mylonakis, 2009). Woloszko (2020) explained that GT measure search intensity (number of searches for a given keyword divided by total searches) by region and time period. Users can search by keyword, category of keywords[3] or topic, being the last two harmonised across languages. For each keyword, the volume data are normalized from 1 to 100 (Zhou et al., 2011). Furthermore, GT provides a dataset of monthly

[3]The 1200 categories can be found in GitHub.

panel data covering observations from 2004 to 2020 for 46 countries, with 1200 searches categories. This tool assigns individual searches to (multiple) categories using a probabilistic algorithm (Choi and Varian, 2012). Some topics address certain ambiguity among keywords. For example, to overcome this problem, the topic "Apple (company)" was created to highlight searches related not to the fruit, but to the company, combining searches for keywords such as "apple watch" and "ipad". Moreover, to remove the problem of seasonality, GT variables are transformed in year-over-year growth rates (Woloszko, 2020).

The tracker built in this paper will exploit GT variables related to TB incidence. For instance, prior researchers identified changes in TB patterns for keywords related to TB, such as "TB symptoms", "TB treatment", "cough", in different geographic locations (Zhou et al., 2011). Note that breaks in certain periods caused by changes in the data collection process are handled by smoothing the year-on-year growth rates (Woloszko, 2020).

Prior studies indicate that internet search data has proven to be effective in estimating the prevalence of several diseases. A popular example of using search volume data in health was the surveillance of influenza outbreaks with comparable accuracy to traditional methods (Nuti et al., 2014). Certain authors revealed that GT predicted the spread of influenza earlier than the CDC (Ginsberg et al., 2009). The system used in the previous paper processes billions of individual searches from five years of Google web search records to estimate influenza-like illness (ILI) occurrence in the United States (US). The proposed method computes a time series consisting of weekly counts of the most common search queries in this country. The conclusion drawn from the study is that this method is more accurate and effective in large populations of web search users. However, they argue that the benefits of real-time disease surveillance using GT outweigh this limitation.

In addition, another research refers to the use of Google Flu Trends for early detection of Influenza outbreaks. This tool builds on the more generic GT tool, which extrapolates data to estimate search volume and displays it in a search volume index graph. Moreover, GT can tracks the increases in volume of a web search query over time. It is also capable of ranking cities, regions and languages based on the search volume of a variable (Carneiro and Mylonakis, 2009). Moreover, this tool continues to show a strong correlation with the CDC's retrospective surveillance data. This research also leads to the same conclusion as the previous study. The authors complement it by stating that Google Flu Trends, and possibly GT, may spot infectious disease activity faster than traditional surveillance methods (Carneiro and Mylonakis, 2009).

There is also a paper that discusses the use of GT for medical nowcasting, which involves using real-time data to predict current and future health trends. The author focuses specifically on otolaryngology. To demonstrate the potential of GT for otolaryngology nowcasting, the author analysed search query data related to otolaryngology related symptoms and conditions over a 7-year period in German population (Braun and Harréus, 2013).

Previous analysis showed that GT data was highly correlated with existing epidemiological data since GT was able to capture variations in symptom prevalence and seasonality. For instance, Google search queries in Germany for 'sinusitis' showed a clear cyclic course, meaning dips in summer and peaks in winter, which significantly correlated with the existence clinical knowledge regarding the incidence of sinusitis. As expected, no seasonal differences were found for the term "acute hearing loss" (Braun and Harréus, 2013). Therefore, the paper highlights the potential of GT for medical nowcasting in otolaryngology and suggests that this approach may also be valuable in other areas of medicine.

2.3.1 Monitor Tuberculosis with Google Trends

As already mentioned in previous sections, several studies discuss the limitations of traditional surveillance systems in predicting and controlling TB outbreaks. To supplement TB surveillance, a past research propose the use of search volume data supplied by Google to monitor TB activity in America. The authors analyse how people search for medical information and demonstrate how Internet search data have been successfully used to estimate disease occurrence in the past. They then develop a dynamic surveillance system using a non-stationary Kalman filter with periodically changing parameters to capture the intrinsic transmission of TB, its relationship to contemporary search volume data, and its yearly cycle (Zhou et al., 2011).

Moreover, the authors of the paper highlight some advantages of their syndromic surveillance system, including weekly updates, the ability to predict outbreaks before they occur, its potential to be implemented at multiple levels, among others.

As shown in the previous research, the system was able to accurately predict the number of TB cases reported by the CDC with a low error rate. This can provide valuable information to public health departments, allowing them to prepare for potential outbreaks and allocate resources accordingly (Zhou et al., 2011).

Besides, this same system allows the public to have greater access to health surveillance in-

formation about TB and other infectious diseases accordingly (Zhou et al., 2011). By analysing search volume data, researchers can gain insights into the types of information a population is searching for and their level of concern about a particular disease. This system can also be adapted for other countries and even for other infectious diseases. While the authors focused on TB in the United States, the same approach can be used to monitor other diseases, especially those for which traditional surveillance systems are limited or inadequate, which is the case of developing countries (Zhou et al., 2011).

Even though Zhou et al. (2011) demonstrates the potential of using search volume data from Google for syndromic surveillance of TB, the authors still refer certain limitations to this approach, such as the potential of ”media effects” and changes in search behavior over time. For instance, if some celebrity is infected by TB, it is very likely that the news reports could cause a significant increase in searches for TB-related terms. Besides, events such as the World Tuberculosis Day always cause an immediate boost in searches that were not related to TB incidence. Under these circumstances, these searches can produce inaccurate estimations.

There is also a relevant paper that focuses on exploring the use of GT as a tool for TB disease surveillance in Indonesia. The study created a set of effective search terms that could be used in detecting early signs of TB outbreaks in this country. The method used in the study is the Pearson correlation to measure the relationship between real surveillance data collected from the Ministry of Health of Indonesia and Google search volume (Fudholi and Fikri, 2020).

The analysis of the study revealed a strong positive relationship between the studied search terms and TB cases in Indonesia, indicating that GT can be used as a rapid disease surveillance tool in Indonesia. For instance, Fudholi and Fikri (2020) suggest a set of search terms with the highest correlation score of 0.907.

3 Methodology, Methods and Tools

3.1 Research Design

In order to determine whether GT data can predict TB incidence, a quantitative approach was built, which involves the collection and analysis of numerical data to identify patterns and relationships.

The research design involves using Machine Learning (ML) methods to forecast the incidence of TB on a monthly basis from January 2021 to April 2023. To accomplish this, we collected historical data on TB incidence and exported GT search volume data. This data will be used to train and test multiple ML models. The goal is to identify the model(s) that provide the most accurate forecast, taking into account factors such as the model's ability to handle noisy data, its overall performance on both training and test data, among others. Once the optimal model have been identified, it will be used to generate monthly forecasts of TB incidence for the study period.

3.2 Data Collection and Description

This section describes the data sources, followed by a detailed description of dataset variables we used to fit the predicted models for TB surveillance.

ECDC Atlas - Historical data

Every year since 1995, European Centre for Disease Prevention and Control (ECDC) collects and publishes the TB incidence expressed as the rate per 100 000 population. From here, we extracted the annual incidence of TB in Portugal between the period 2007 to 2020, which is the last known public data (ANNEX A1). Figure 1 shows the evolution of TB incidence rate in Portugal between 1995 and 2020 obtained from previous source.

Google Search Volume data

One of the motivations of our study is to predict the number of TB cases using search volume data. GT supplies the search volume of a given term from January 2004 to the present. According to Woloszko (2020), the search intensities of most search categories show a downward trend over time because Google Search user base has increased dramatically since 2004. In this sense, the first three years of GT are considered meaningless data. As a result, this work

will only consider GT data from January 2007 to April 2023. In our research, a group of TB-related terms were chosen by domain knowledge, which included the names, causes, symptoms, diagnosis methods, and related diseases of TB. This includes 19 Google search terms (listed in Table 1), some of which were based on the article (Zhou et al., 2011), adapted to Portuguese. For instance, Figure 3 shows the search volume of the term ”tuberculose”. The numbers on the graph reflect how many searches have been submitted on Google every month since 2007 for “tuberculose” in Portugal.

Types	**Terms**
Name	tuberculose, TB
Symptom	sintomas tuberculose, palidez, dor peito, arrepios, febre, fadiga, tosse cronica, suor noturno, tossir sangue, perda apetite
Cause	mycobacterium, bacilo koch, silicose
Diagnosis	PPD, tuberculina, mantoux
Related Disease	diabetes mellitus

Table 1: List of Terms examined for Search Volume

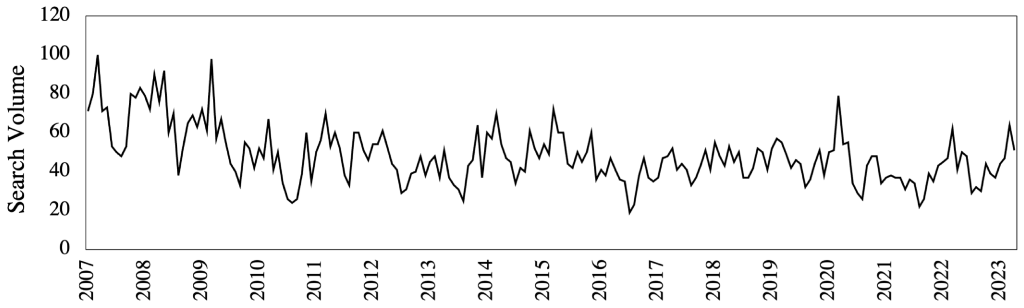

Figure 3: Monthly updated search volume of the term “tuberculose”, 2007-2023. Source: GT.

3.3 Data Preprocessing

Data quality is the first key to ensure the successful building of forecasting models. In this context, the ECDC data was submitted to a previous treatment. The article (Bras et al., 2014) shows the monthly seasonal incidence of PTB in Portugal between the period 2000-2010, being able to take the approximate monthly seasonal coefficients of 2007, see Figure 4. The seasonal coefficients represent the deviation from the average monthly incidence and can be used to estimate the monthly incidence of TB from annual data.

To do this, we first took the annual TB incidence for each year, divided it by the 12 months to get an approximate monthly incidence and then applied the approximate seasonal coefficients of 2007, one for each month, to account for the seasonal variation in the data (detailed process in ANNEX A2). We assume these seasonal coefficients to be constant over time, so the monthly coefficients from 2007 are used for all subsequent years. By using these coefficients, shown in ANNEX A2, we can capture seasonal patterns in the data, such as higher incidence in certain months of the year. This way it was possible to convert the annual TB incidence into monthly incidence.

The final dataset comprises 20 monthly variables, consisting of 168 observations for the incidence of TB and 196 records for each of the 19 TB-related searches. Additional details can be found in the ANNEX B1 for a more comprehensive understanding of the descriptive statistics of the dataset.

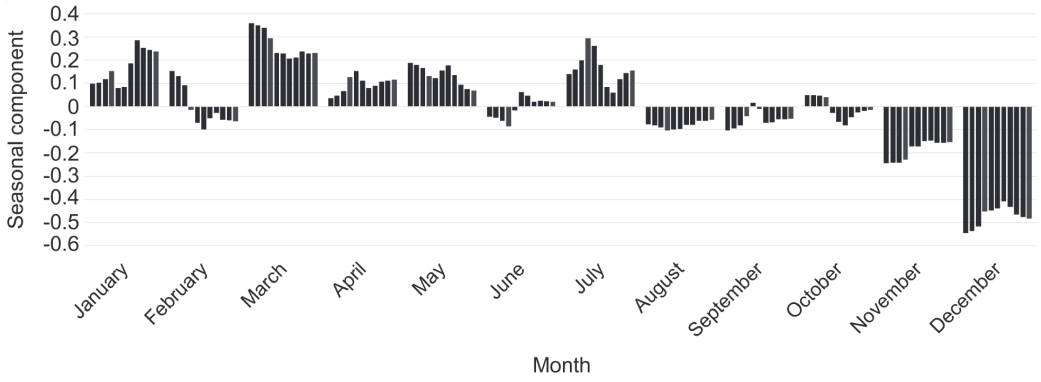

Figure 4: Seasonal subseries plot of the seasonal component estimated by STL decomposition of PTB monthly incidence rates in Portugal, 2000–2010. Source: Article (Bras et al., 2014).

3.4 Data Modeling

3.4.1 Cross-Validation on time series data

There are are specific considerations to take into account when working with time series data. The classical Cross-Validation (CV) approach becomes invalid in this context due to the temporal dependency within the series, meaning it makes no sense to forecast the past using future data (Chen et al., 2023). To overcome this issue, we used CV on rolling window (Bergmeir and Benítez, 2012). It works by dividing the data into a series of overlapping subsets, allowing to train the model on past data and validate its performance on future data. The training set is expanded sequentially over time to preserve the temporal relationship as shown in Figure 5.

In practice, the models were trained in a rolling window with 132 ($n - k = 168 - 36$) observations and tested in the following period (or fold). For each fold, the forecast error is computed and the overall error of the CV is the average of each fold error. To accomplish this, the data was divided into k = 36 folds, corresponding to the 3 years of the COVID-19 pandemic, which were included as part of our test period. For our dataset, we defined the TB-related terms as the attributes and the incidence rate of TB as the response (q_x = 19 and q_y = 1, being q = length). After splitting the data into training and testing sets with a [80%–20%] ratio, ML models will be selected and trained to forecast TB incidence (follow up in the next section).

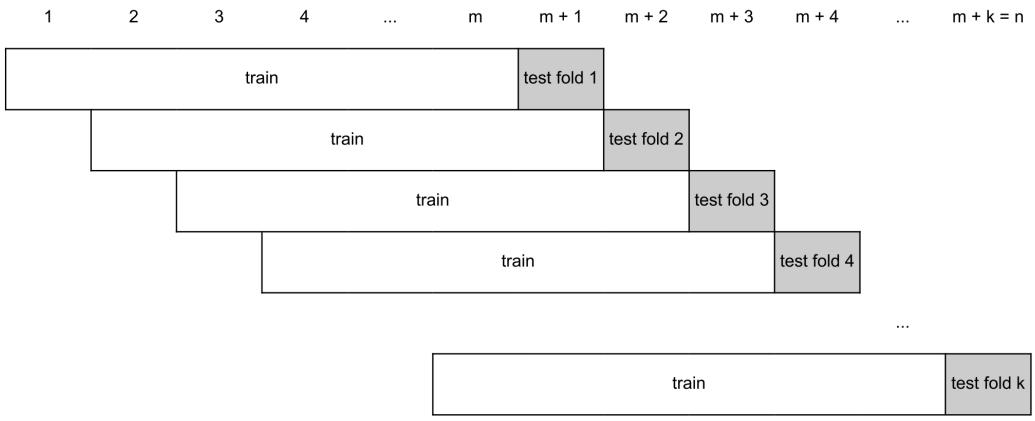

Figure 5: OSS Cross-Validation with time series.

3.4.2 Machine Learning Models

ML has become an increasingly popular field for analyzing complex data. Using ML models can be beneficial due to their ability to handle high-dimensional data (Murphy, 2023) and potentially improve prediction accuracy when compared to standard statistical models. In this paper, we will use multiple ML algorithms to analyze our dataset to be able to answer our research question. These algorithms are designed to automatically uncover generalizable pattern and relationships in the data (Mullainathan and Spiess, 2017, p.88), even in cases where the relationships could be non-linear or involve interactions between multiple attributes (x), which is the case with GT data. Besides, ML models can easily identify which Google search terms are most strongly associated with TB incidence.

Specifically, we will use five ML models, namely Ordinary Least Squares (OLS), Least Absolute Shrinkage and Selection Operator (LASSO), Partial Least Squares (PLS), Support Vector Machine (SVM) and Deep Neural Network (DNN). All these models were chosen for their abil-

ity to detect complex patterns and make accurate predictions. In the following subsections, we describe each of these models in more detail.

ORDINARY LEAST SQUARES (OLS)

OLS is a method for estimating the parameters of a linear regression model. This method estimates are commonly used to make predictions or even to test hypotheses regarding the relationship between the dependent variable and the independent variables. The equation for OLS is as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{1}$$

where y_i is the dependent variable, which is the variable we want to predict, x_{ij} are the independent or explanatory variables and β_j are the regression coefficients (Greene, 2003, p. 7). The latter represents the expected change in the dependent variable for a unit increase in the corresponding independent variable, holding all other independent variables constant. The term ε_i is the error term or the residual, which captures the part of the dependent variable that is not explained by the independent variables.

The OLS method seeks to estimate the values of the regression coefficients (β_j) that minimize the Sum of Squared Residuals (RSS) between the predicted values and the actual values (James et al., 2021, p. 72). It aims to find the line that best fits the data. The previous equation is derived by minimizing the RSS, which is defined as:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

where y_i is the actual value of the dependent variable and $\hat{y}_i$ is the predicted value of the dependent variable.

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

A popular approach for estimating parameters in sparse high-dimensional linear models is the LASSO, which was first introduced by Tibshirani (2011). Within the framework of the paper (Mullainathan and Spiess, 2017), LASSO corresponds to: a quadratic loss function; a class of linear functions; and a regularizer (sum of the absolute value of each coefficient). This method is a modification of the OLS model, where an additional penalty term is added to the

objective function. Following the formula proposed by James et al. (2021, p. 241), the LASSO coefficients, $\hat{\beta}_\lambda$, minimize the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3}$$

The penalty term $\lambda\sum_{j=1}^{p}|\beta_j|$ has the effect of forcing some of the regression coefficients to be exactly zero when the tuning parameter λ is sufficiently large, leading to variable selection (James et al., 2021, p. 242). In short, the parameter λ controls the strength of the penalty, and a larger value of λ leads to a more aggressive shrinking of the coefficients and more variable selection. Note that in the special case where $\lambda = 0$, the LASSO equation reduces to the least squares fit (equation (1)) and when λ becomes sufficiently large, LASSO outputs the null model in which all regression coefficients equal zero (James et al., 2021, p. 242).

PARTIAL LEAST SQUARES (PLS)

PLS can be used for both regression and classification problems and has been applied in various fields such as chemometrics, bioinformatics, and image analysis (Rosipal and Krämer, 2006, p. 34). This method was first developed by Herman (1966) and has since been extended by many other researchers.

We specifically used the most basic algorithm of PLS, known as Marginal Regression followed by (Taddy, 2019, p. 226), which is recommended in regression scenarios where the number of covariates is high, and where the likelihood of multicollinearity [4] is expected (Abdi, 2010). In the context of our research, it works by calculating the OLS coefficient (β) for each GT search and a variable v_i is created that results from the sum of all searches weighted by the respective OLS coefficients. In practice, the so-called marginal regression is performed as the following:

$$\hat{TB} = \alpha + \beta * v_i + \varepsilon_i \tag{4}$$

This regression is applied successively to the e_i errors in order to improve the prediction for TB, in a process known as boosting.

[4] Statistical concept where two or more independent variables in a regression model are highly correlated with each other.

SUPPORT VECTOR MACHINES (SVM)

James et al. (2021, p. 367) presents a comprehensive introduction to SVM, a well-known supervised learning model in ML. The purpose of SVM in regression is to determine the hyperplane that provides the best fit for the data while maximizing the margin around it. This method is especially useful in high-dimensional data or when the data is non-linearly separable.

SVM can be performed with different kernel functions, this includes linear, polynomial, radial, sigmoid, among others. Using kernel functions, data expands to a feature space in order to accommodate non-linear class boundaries James et al. (2021, p. 386).

Having summarized the basics of SVM, let's focus on the polynomial kernel, which is the kernel function used in this study. This specific function represents the similarity of observations, mapping the data into feature space using a polynomial function. James et al. (2021, p. 382) defined it as the following:

$$K(x_i, x_j) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{5}$$

where $K(x_i, x_j)$ is the kernel function that quantifies the similarity of the two observations (x_i and x_j) and d represents the degree of the polynomial (where $d > 1$). Using higher values of d leads to a much more flexible decision boundary.

DEEP NEURAL NETWORK (DNN)

According to James et al. (2021, p. 623), the term DNN refers to any kind of differentiable function that can be represented as a computation graph, where the nodes are primitive operations such as matrix multiplication, while the edges represent numeric data in the form of vectors, matrices, or tensors.

Among the various types of DNN, we specifically focus on Multilayer Perceptron (MLP) for our research. The learning task of discovering a good solution is made much easier with multiple layers each of modest size (James et al., 2021, p. 407). A MLP, also known by feedforward neural network (FFNN), is one of the simple types of neural networks (Murphy, 2023, p. 632). It defines a mapping $y = f(x; \theta)$ and acquires the optimal values of the parameters θ that yield the best function approximation. These models are referred to as feedforward since information flows through the function being evaluated from x to the output y (Aaron et al., 2016, 164). As shown by Murphy (2023, p. 632), a MLP is composed of a series of L linear

layers, combined with elementwise nonlinearities:

$$f(x;\theta) = W_L\varphi_L(W_{L-1}\varphi_{L-1}(\dots\varphi_1(W_1x)\dots)) \tag{6}$$

where x is the input vector, which contains the values of the independent variables, θ represent the model parameters, $W_1, W_2, ..., W_L$ are the weight matrices for each layer. The W_L is the output layer (final layer), which produces the final prediction. Lastly, $\varphi_1, \varphi_2, ..., \varphi_L$ are the non-linear activation functions for each layer like the Rectified Linear Unit (ReLU). These functions introduce nonlinearity into the model, allowing it to capture complex relationships between the input and output variables.

In summary, the function represents the forward propagation of information through an MLP, where the input vector is transformed through a series of weighted connections and non-linear activation functions to produce an output.

3.5 Measures of Prediction Accuracy

When evaluating the ability of a forecasting model to make accurate predictions, there are many different measures of accuracy that are commonly used for time series data. Some of these measures are especially useful for evaluating how well a model can predict time series data.

However, each measure of accuracy has its own strengths and weaknesses, meaning that it is difficult to rely on just one measure to evaluate a model's performance. As a result, all the following measures of accuracy were used to assess the forecasting models in question:

MEAN ABSOLUTE ERROR (MAE)

The Mean Absolute Error (MAE) which is also called the Mean Absolute deviation is the average of the absolute value of the forecasts errors. The formula for MAE is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{7}$$

where N is the number of predictions, y_i are the observed values and $\hat{y}_i$ are the predicted values. The closer the MAE is to 0, the more accurate is the prediction (Tatachar, 2021). MAE tells us how close the line of best fit is to the data points.

MEAN SQUARED ERROR (MSE)

The Mean Squared Error (MSE) which is also called the Mean Squared Deviation is the mean squared error between the predicted and the actual values,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

where N is the number of predictions, y_i are the observed values and $\hat{y}_i$ are the predicted values. The square is taken to eliminate negative signs so MSE always takes positive values. The use of MSE to evaluate the accuracy of a specific forecasting model presents certain failures, namely its high sensitivity to outliers and scale dependency (Hyndman and Koehler, 2006).

ROOT MEAN SQUARED ERROR (RMSE)

The Root Mean Squared Error (RMSE) which is also called the Root Mean Squared Deviation is the square root of the mean of squares of all the errors. This measure of accuracy is simply the Standard deviation of the errors. The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{9}$$

where N is the number of predictions, y_i are the observed values and $\hat{y}_i$ are the predicted values. When comparing RMSE with MAE, the latter is less sensitive to outliers and easily interpreted as the squaring operation in RMSE amplifies the weight of larger errors. (Hyndman and Koehler, 2006).

AKAIKE INFORMATION CRITERION (AIC)

The Akaike Information Criterion (AIC) is a statistical measure of the relative quality of different models for a given set of data, following the simplification proposed by Greene (2003, p. 565):

$$AIC = ln\left(\frac{\epsilon'\epsilon}{k}\right) + \frac{2q}{k} \tag{10}$$

where ϵ is a column vector with the OOS errors, k is the number of folds and q is the number of parameters of each forecasting model. AIC penalizes models with more independent variables

(parameters) in order avoid overfitting and the one with the lowest AIC is generally considered to be the best fit for the data.

3.6 Tools

Throughout this research, we used Microsoft Excel for data extraction and to filter the annual data from ECDC into monthly TB incidence with the approximate seasonal coefficients. R Studio was also used to build the predictive models of TB incidence.

3.6.1 Software Support

The data analysis, processing and models building were done in the **R.Version(2022.02.0+443)**, using a small library of machine learning methods specifically designed for time series. To use these methods, it was necessary to load certain auxiliary packages, including **MASS**, **textir**, **gamlr**, **tree**, **kernlab**, and **h2o**.

We loaded another packages that helped us summarized data, **skimr**, customize our plots, **ggthemes**, look at relationships between variables, **GGally** and **corr**, and produce formatted tables, **knitr**. Alonge with these, we included **tidyverse**, a vast collection of R packages specifically designed for data science, and **stargazer**, a command that produces LaTeX code and outputs summary statistics. Additionally, the common libraries **ggfortify**, **corrplot**, **ggplot2** and **RColorBrewer** were used to generate time plots in R studio.

4 Findings

4.1 Relationship between Tuberculosis incidence and Google search terms

Firstly in our analysis, it is important to explore the strength and direction of the relationships among the variables of interest and explain what these associations suggest about the relationship between TB cases and TB-related searches. In this sense, a correlogram is a good starting point, which is a graphical representation of the correlation matrix, being a quick and easy way to spot relationships between pairs of variables and identify potential multicollinearity.

Based on the correlogram 6, there is only one important positive correlation to mention, namely between the term ”tuberculose” and monthly incidence of TB, with a strong correlation coefficient of ≈ 0.707. There are multiple weak positive correlations in this dataset. For instance, the one between the term ”diabetes mellitus” and incidence of TB, with a correlation coefficient of ≈ 0.301. There is also a weak positive correlation between the term ”sintomas tuberculose” and incidence of TB, with a correlation coefficient of ≈ 0.166. Curiously, the term ”febre” and the incidence of TB has a negative correlation. Note that correlation does not necessarily imply causality, and further analysis are necessary to determine the causal relationships among these variables.

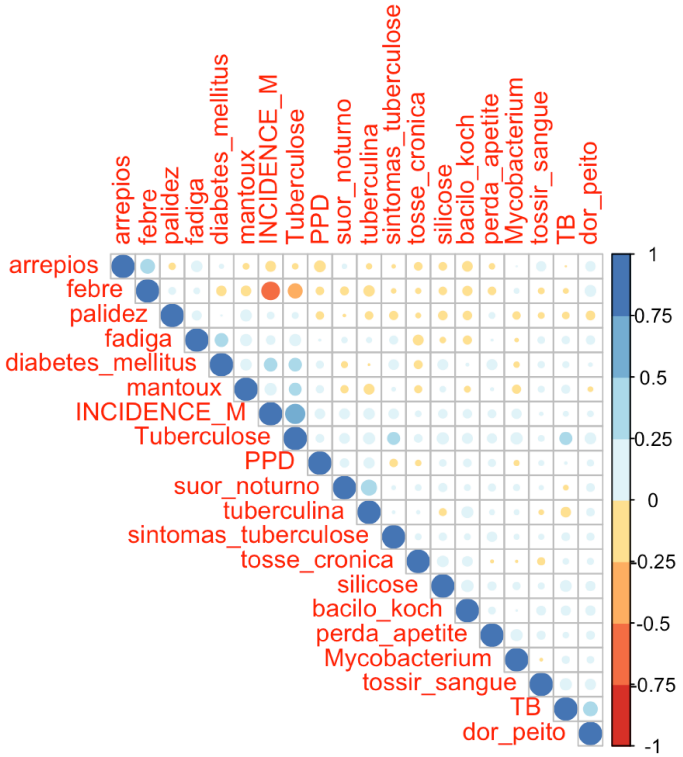

Figure 6: Correlogram of the dataset variables: TB incidence and TB-related search terms.

We find it relevant to create a scatter plot to better understand the relationship between the Google search term ”tuberculose” and the monthly incidence of TB since it was the only strong positive correlation, see Figure 7. A scatter plot is a useful way to visually represent the relationship between two variables and can help to identify patterns in the data. In this case, it can help to confirm if there is a positive or negative association between the two variables, or even if there are any outliers or unusual observations.

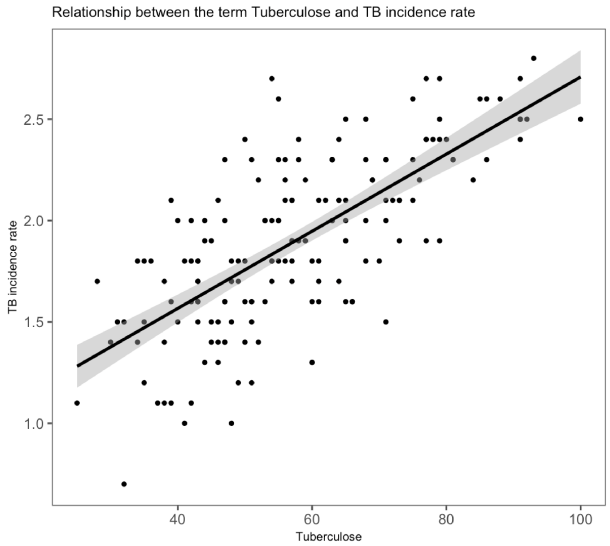

Figure 7: Relationship between the term ”tuberculose” and TB incidence rate.

The scatterplot suggests a positive (when one variable increases, the other increases as well) linear relationship between the two variables. The points cluster together in a diagonal pattern with little deviation from the direction of the line. This was expected given that we already knew the correlation coefficient between the two variables (≈ 0.707).

4.2 Simple Regression Results

Based on the previous results, we will conduct a simple regression analysis to test the hypothesis that there is an actual relationship between incidence rate of TB and the term ”tuberculose”. Regarding the Table 2, the regression equation is as follows:

$$\hat{Incidence}_M = 0.019 * tuberculose + 0.806 \tag{11}$$

Concerning the estimated slope of the model, $\hat{\beta}_1$ = 0.019, one can state that for every unit increase in the term ”tuberculose”, there is, on average, a predicted increase of 0.019 in the

incidence rate of TB, holding all other variables constant. The three stars next to the coefficient indicate that this value is statistically different from zero at the 1% level ($\alpha = 0.01$).

The estimated intercept or constant of the model is $\hat{\beta}_0$ = 0.806. This means that, on average, the constant term of 0.806 represents the predicted incidence rate of tuberculosis when the term "tuberculose" equals zero, holding all other variables constant.

The R-squared value of 0.500 indicates that term "tuberculose" explains 50% of the variation in the incidence rate of TB. In summary, the model seems to be a good fit, as the F statistic is significant and the residual standard error is relatively small.

As correlation is not sufficient to infer causality between variables, we computed a more complete regression model (see ANNEX C1), where we included all the variables of our dataset. The adjusted R-squared value of 0.645 indicates that all TB-related searches jointly explain 64.5% of the variation in the incidence rate of TB. Once again, the model appears to be a suitable fit, with an adjusted R-squared value close to 0.7 (standard value for a high level of correlation), a significant F statistic, and a relatively small residual standard error.

	Dependent variable:
	INCIDENCE_M
tuberculose	0.019***
	(0.001)
Constant	0.806***
	(0.088)
Observations	168
R^2	0.500
Adjusted R^2	0.496
Residual Std. Error	0.298 (df = 166)
F Statistic	165.672*** (df = 1; 166)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Simple Regression Model

4.3 Model Evaluation

As stated in section 3, the performance of the different ML models selected for our analysis will be compared based on the following measures of forecast error: MAE; MSE; RMSE; and AIC. Table 3 shows the results of each metric for each of the five models.

It is possible to denote that among all the ML models evaluated, PLS achieved the best results in the dataset for all selected accuracy metrics. This model has the lowest MAE, MSE, RMSE and even the lowest AIC value. Comparing with the DNN and OLS models, PLS was able to reduce the RMSE by $\approx 17\%$ and $\approx 4\%$, respectively. Therefore, PLS outperformed all other models, having the best predictive power.

The OLS and LASSO models have a similar performance, whereas the SVM and DNN models exhibit higher MAE, MSE, and RMSE values, as well as higher AIC values, indicating poorer predictive performance compared to all the other models. Meanwhile, the performance of the Median model appears to be on par with the OLS and LASSO models.

Models/Metrics	MAE	MSE	RMSE	AIC
OLS	0.235	0.086	0.293	-1.401
LASSO	0.237	0.085	0.291	-1.413
PLS	0.229	0.080	0.282	-1.474
SVM	0.233	0.094	0.307	1.307
DNN	0.245	0.095	0.330	1.294
Median	0.233	0.085	0.291	-1.413

Table 3: Models performance

4.4 Proposed PLS based model to forecast Tuberculosis incidence

The performance of the different models was compared earlier to select the best predicted model to forecast the incidence of TB in Portugal. Based on the previous results, one can state that the best model is the PLS. The Figure 8 shows the forecast made for the period between January 2021 to April 2023 using PLS.

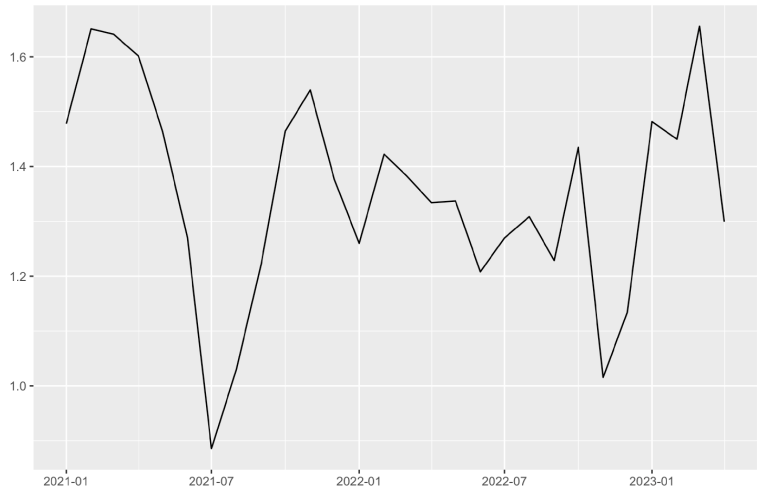

Figure 8: Forecast of TB incidence, 2021M1 - 2023M4.

Concerning the forecast, it can be noted that TB incidence exhibits strong monthly fluctuations over a period of 3 years, with the most pronounced troughs occurring in July 2021 and October 2022. As expected, the months of February and March present the highest values since the seasonal coefficients suggest a higher incidence in winter and early spring. The peaks in November and the lower incidence in July 2021 are more difficult to explain due to potential changes in historical seasonal patterns.

5 Discussion

The goal of this research was to investigate whether GT data could predict TB incidence in Portugal. By evaluating trends in search volume for TB-related terms and comparing them to historical data on TB incidence from ECDC, we aimed to test our hypothesis that there is a relation between existing TB cases and contemporary TB-related searches. In this section, our findings will be discussed to verify whether there is evidence to support our hypothesis.

Firstly, we used different statistical methods to explore the relationship between Google search volume related with TB and the monthly incidence of TB. The correlogram (Figure 6) revealed that the relationship between TB-related searches and monthly incidence of TB were not significant, except for the term ”tuberculose”. However, the scatter plot (Figure 7) and both regression analysis (Tables 2 and C1) out weighted the previous result, since TB-related searches can explain more than half of the variation in the incidence rate of TB in Portugal.

The predictive model, namely PLS, demonstrated a strong forecast accuracy comparing with the remaining ML models considered in the analysis. It successfully identified the underlying patterns of TB incidence. This high predictive performance of the PLS model highlights its accuracy for this task, providing confidence in the reliability of the forecast.

The forecast (Figure 8) showed a resistant annual TB incidence lower than the one observed in 2019 but higher than that in 2020 (Figure 1), which was abnormally low. This could be due to the COVID-19 pandemic, which may have limited the diagnosis of other diseases, which goes along with the projection of DGS (2018) as previously discussed in section 2.

In light of this, our findings end up supporting our hypothesis, concluding that this research might provide valuable insights about the relationship between Google search volume and TB incidence. Thus, the results obtained may confirm the potential of using searches to estimate the incidence of TB, as previously explored by Zhou et al. (2011). It is important to highlight that our research adds value by focusing specifically on Portugal and employing alternative methods instead of relying on a non-stationary Kalman filter, as implemented by previous authors.

Building on these findings, we developed a TB disease surveillance system based on data provided by the Google search engine. While TB is used here as an example, the proposed approach can also be applied to other diseases. Besides, the official data on TB incidence in Portugal is restricted to Direção-Geral da Saúde (DGS), making it inaccessible to the public. However, GT data are updated at the end of each week, enabling our surveillance results to be

ahead of DGS data. By providing weeks of lead time and using real-time GT data, this approach may empower public health authorities to identify early potential outbreaks and allocate surveillance resources more efficiently on this matter. Moreover, it does not depend on constant reports from the National Health department, the information provided is publicly available online and it is cost-effective as it does not require additional amount of resources. Above all, this approach may have the potential to increase public awareness to health surveillance information.

In short, our analysis might offer early detection capabilities that may serve as an important line of defense against future outbreaks in Portugal. Embracing these advancements may enable public health authorities to ultimately mitigate the burden of TB in Portugal and beyond.

6 Conclusion and further developments

6.1 Summary

This research shows that using Google search volume may help predict the occurrence of TB in Portugal. The study found that there is a relationship between TB-related search terms and the number of TB cases. This TB surveillance system can perhaps provide with up-to-date TB incidence projection for a near future to health departments and help them to earlier identify potential outbreaks. If this approach was effectively implemented, this might eventually reduce the burden of TB in Portugal and beyond, ultimately contributing to the goal of eradicating TB by 2030 (Euronews, 2021), as the surveillance system also becomes more rigid and robust.

6.2 Limitations

It is important to acknowledge certain limitations of our research. Firstly, the scope of of our study was restricted to Portugal, so the application of the previous findings to other countries may vary. Secondly, GT data does not provide information on specified TB cases. Additionally, our research is significantly limited by the ”media effect” of GT. It is important to consider the ”media effect” when analyzing data from GT, as it can directly influence search volume for specific terms and consequently lead to inaccurate results.

Another limitation is that we assumed seasonality to be constant over time, which may not be accurate due to several factors, such as the impact of COVID-19. Additionally, by using approximate seasonal coefficients, we are dealing with estimations of the monthly incidence of TB instead of measuring the actual monthly incidence.

6.3 Future Research

Future research should aim to address these limitations and further explore the understanding of GT data for TB surveillance. It would be interesting to replicate this study in different geographical contexts in order to filter the geographic variation on the subject.

An alternative approach to improve the analysis would be to also adopt the method applied by Zhou et al. (2011), which used a dynamic surveillance system using a non-stationary Kalman filter instead of relying solely on ML models for the prediction of TB incidence. It would also be important to apply the proposed method with the official DGS data that is not easily available

and whose guarantee would not have been possible in the time foreseen for the development of this research. Such developments could represent important avenues for future research.

References

AARON, C., G. LAN, AND B. YOSHUA (2016): "Deep Learning - Ian Goodfellow, Yoshua Bengio, Aaron Courville - Google Books," .

ABAT, C., H. CHAUDET, J. M. ROLAIN, P. COLSON, AND D. RAOULT (2016): "Traditional and syndromic surveillance of infectious diseases and pathogens," .

ABDI, H. (2010): "Partial least squares regression and projection on latent structure regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2.

AREIAS, C., T. BRIZ, AND C. NUNES (2015): "Pulmonary tuberculosis space-time clustering and spatial variation in temporal trends in Portugal, 2000-2010: An updated analysis," *Epidemiology and Infection*, 143.

BERGMEIR, C. AND J. M. BENÍTEZ (2012): "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, 191.

BRAS, A. L., D. GOMES, P. A. FILIPE, B. D. SOUSA, AND C. NUNES (2014): "Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal," *International Journal of Tuberculosis and Lung Disease*, 18.

BRAUN, T. AND U. HARRÉUS (2013): "Medical nowcasting using Google trends: Application in otolaryngology," *European Archives of Oto-Rhino-Laryngology*, 270.

CARNEIRO, H. A. AND E. MYLONAKIS (2009): "Google trends: A web-based tool for real-time surveillance of disease outbreaks," *Clinical Infectious Diseases*, 49.

CHAPMAN, A. L., T. C. DARTON, AND R. A. FOSTER (2013): "Managing and monitoring tuberculosis using web-based tools in combination with traditional approaches," .

CHEN, W. J., J. J. YAO, AND Y. H. SHAO (2023): "Volatility forecasting using deep neural network with time-series feature embedding," *Economic Research-Ekonomska Istrazivanja*, 36.

CHOI, H. AND H. VARIAN (2012): "Predicting the Present with Google Trends," *Economic Record*, 88.

DANIEL, T. M. (2006): "The history of tuberculosis," *Respiratory Medicine*, 100.

DGS (2018): "Tuberculose em Portugal Desafios e Estratégias," .

——— (2022): "Relatório de Vigilância e Monitorização da Tuberculose em Portugal," .

EFFLER, P., M. CHING-LEE, A. BOGARD, M.-C. IEONG, T. NEKOMOTO, AND D. JERNIGAN (1999): "Statewide System of Electronic Notifiable Disease Reporting From Clinical Laboratories," *JAMA*, 282.

EURONEWS (2021): "Um século de BCG. Covid tira recursos à vacina que salvou milhões," .

FUDHOLI, D. H. AND K. FIKRI (2020): "Towards an Effective Tuberculosis Surveillance in Indonesia through Google Trends," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 299–308.

GINSBERG, J., M. H. MOHEBBI, R. S. PATEL, L. BRAMMER, M. S. SMOLINSKI, AND L. BRILLIANT (2009): "Detecting influenza epidemics using search engine query data," *Nature*, 457.

GREENE, W. H. (2003): *Econometric Analysis*, Upper Saddle River, New Jersey: Prentice Hall, 5th ed.

HERMAN, W. (1966): "Estimation of principal components and related models by iterative least squares," *Krishnaiaah, P.R*, Multivaria.

HYNDMAN, R. J. AND A. B. KOEHLER (2006): "Another look at measures of forecast accuracy," *International Journal of Forecasting*, 22.

JAJOSKY, R. A. AND S. L. GROSECLOSE (2004): "Evaluation of reporting timeliness of public health surveillance systems for infectious diseases," .

JAMES, G., T. HASTIE, R. TIBSHIRANI, AND D. WITTEN (2021): "An introduction to statistical learning (2nd ed.)," *Springer texts*, 102.

JUN, S. P., H. S. YOO, AND S. CHOI (2018): "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications," *Technological Forecasting and Social Change*, 130.

LUSA, A. (2023): "Tuberculose. Autoridades antecipam dificuldades em cumprir metas da OMS nos próximos anos," .

MAY, L., J. P. CHRETIEN, AND J. A. PAVLIN (2009): "Beyond traditional surveillance: Applying syndromic surveillance to developing settings-opportunities and challenges," *BMC Public Health*, 9.

MOHAJAN, H. K. (2015): "Tuberculosis is a Fatal Disease among Some Developing Countries of the World," *American Journal of Infectious Diseases and Microbiology*, 3.

MULLAINATHAN, S. AND J. SPIESS (2017): "Machine learning: An applied econometric approach," vol. 31.

MURPHY, K. P. (2023): "Probabilistic machine learning : advanced topics," *Gaussian processes for machine learning*.

NUTI, S. V., B. WAYDA, I. RANASINGHE, S. WANG, R. P. DREYER, S. I. CHEN, AND K. MURUGIAH (2014): "The use of google trends in health care research: A systematic review," .

ROSIPAL, R. AND N. KRÄMER (2006): "Overview and recent advances in partial least squares," vol. 3940 LNCS.

SREERAMAREDDY, C. T., K. V. PANDURU, J. MENTEN, AND J. V. DEN ENDE (2009): "Time delays in diagnosis of pulmonary tuberculosis: A systematic review of literature," .

TADDY, M. (2019): *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*, McGraw Hill, 1st ed.

TATACHAR, A. V. (2021): "Comparative Assessment of Regression Models Based On Model Evaluation Metrics," *International Research Journal of Engineering and Technology*, 8.

THRUSFIELD, M. (2007): *Veterinary Epidemiology*, vol. 4, Blackwell Science Ltd., third ed.

TIBSHIRANI, R. (2011): "Regression shrinkage and selection via the lasso: A retrospective," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73.

WHO (2021a): "Tuberculosis deaths rise for the first time in more than a decade due to the COVID-19 pandemic," .

——— (2021b): "Tuberculosis surveillance and monitoring in Europe 2021 - 2019 data," .

——— (2022): "Tuberculosis deaths and disease increase during the COVID-19 pandemic," .

——— (2023): “Tuberculosis surveillance and monitoring in Europe 2023 - 2021 data,” .

WOLOSZKO, N. (2020): “Tracking Activity in Real Time with Google Trends,” .

ZHOU, X., J. YE, AND Y. FENG (2011): “Tuberculosis surveillance by analyzing google trends,” *IEEE Transactions on Biomedical Engineering*, 58.

A Evolution of Tuberculosis incidence in Portugal

Appendix Table A1: Annual TB incidence and evolution rate in Portugal between the period 1995 - 2020

1995 - 2020		
	Rates	
Year	Incidence	Evolution
1995	55,722	NULL
1996	52,252	-6,23%
1997	50,693	2,39%
1998	51,906	-2,98%
1999	50,655	-2,41%
2000	43,848	-13,44%
2001	42,582	-2,89%
2002	43,301	1,69%
2003	39,714	-8,28%
2004	36,799	-7,34%
2005	33,903	-7,87%
2006	32,877	-3,03%
2007	29,803	-9,35%
2008	28,446	-4,55%
2009	27,180	-4,45%
2010	25,677	-5,53%
2011	24,677	-3,90%
2012	24,719	0,17%
2013	22,980	-7,04%
2014	21,846	-4,93%
2015	21,157	-3,16%
2016	18,721	-11,51%
2017	18,565	-0,83%
2018	18,813	1,33%
2019	18,800	-0,07%
2020	14,035	-25,35%

Author's Calculation. This Table presents the statistics related to the annual incidence and evolution rate of TB in Portugal. Note that TB incidence rate is the estimated rate of TB expressed as the rate per 100 000 population. Source: ECDC Atlas.

Appendix Table A2: Data Preprocessing - filter the annual TB incidence to monthly incidence with monthly seasonality of 2007

2007			
	Data Preprocessing		
Month	Annual incidence (a)	Seasonal coefficients (b)	Monthly incidence (c)
January	29,8	0,2	2,7
February	29,8	0,0	2,5
March	29,8	0,3	2,8
April	29,8	0,1	2,6
May	29,8	0,2	2,7
June	29,8	0,0	2,5
July	29,8	0,2	2,7
August	29,8	-0,1	2,4
September	29,8	-0,1	2,4
October	29,8	0,0	2,5
November	29,8	-0,2	2,3
December	29,8	-0,5	2,0

Author's Calculation. This Table presents the data preprocessing related with filtering the annual data to monthly data. Implementing the process for 2007, we first took the annual TB incidence of this year (a), which is 29,8, divided it by 12 to get an approximate monthly incidence and then summed the corresponding approximate seasonal coefficient (b) to account for the seasonal variation in the data ($c_m = a_m/12 + b_m$, being m the corresponding month). We assumed seasonality to be constant over time, so these monthly coefficients from 2007 (b) are used for all subsequent years. Note that this process is done for all the years until 2020. Source: (Bras et al., 2014).

B Descriptive Statistics

Appendix Table B1: Descriptive Statistics of the Dataset

Statistics	N	Mean	St. Dev.	Min	Max
INCIDENCE_M	168	1.901	0.419	0.700	2.800
tuberculose	196	55.214	15.853	25	100
TB	196	47.026	13.033	0	100
sintomas_tuberculose	196	26.643	20.216	0	100
palidez	196	10.816	15.215	0	100
dor_peito	196	49.704	17.437	0	100
arrepios	196	35.311	16.880	0	100
febre	196	25.699	10.989	9	100
fadiga	196	32.969	11.613	0	100
tosse_cronica	196	6.847	13.852	0	100
suor_noturno	196	6.617	13.605	0	100
tossir_sangue	196	5.796	11.832	0	100
perda_apetite	196	11.821	16.118	0	100
mantoux	196	9.332	14.319	0	100
PPD	196	21.362	17.412	0	100
tuberculina	196	11.495	17.127	0	100
silicose	196	11.082	15.763	0	100
Mycobacterium	196	16.311	18.133	0	100
bacilo_koch	196	6.964	15.340	0	100
diabetes_mellitus	196	30.582	16.212	0	100

Author's Calculation. This Table presents the summary statistics of the monthly variables of the dataset, including the TB incidence and the 19 TB-related terms. Note that the scale of GT ranges from 0 to 100, with 0 indicating no search volume for a specific term during a certain period and 100 indicating the maximum search volume based on its historical search volume. In short, the TB-related terms presented in the Table can only range between 0 and 100. Source: ECDC Atlas and GT.

C Multiple Regression Results

Appendix Table C1: Multiple Regression Analysis

	Dependent variable:
	INCIDENCE_M
Tuberculose	0.014***
	(0.002)
tuberculina	−0.002
	(0.001)
sintomas_tuberculose	0.001
	(0.001)
TB	−0.006***
	(0.002)
palidez	0.003*
	(0.001)
dor_peito	0.003**
	(0.001)
arrepios	−0.00000
	(0.001)
febre	−0.019***
	(0.003)
fadiga	−0.0004
	(0.002)
tosse_cronica	0.0001
	(0.001)
suor_noturno	0.002
	(0.002)
tossir_sangue	−0.001
	(0.002)

Continued on next page

Appendix Table C1 – continued from previous page

	Dependent variable:
	INCIDENCE_M
perda_apetite	−0.002
	(0.001)
mantoux	0.0002
	(0.001)
PPD	0.002*
	(0.001)
silicose	−0.001
	(0.001)
Mycobacterium	0.002*
	(0.001)
bacilo_koch	0.0001
	(0.001)
diabetes_mellitus	0.003*
	(0.001)
Constant	1.524***
	(0.127)
Observations	168
R^2	0.686
Adjusted R^2	0.645
Residual Std. Error	0.250 (df = 148)
F Statistic	16.988*** (df = 19; 148)
Note:	*p<0.1; **p<0.05; ***p<0.

D Sample Code about forecasting with Machine Learning: Tuberculosis Dataset

```
# Generic functions (source libmlts.R)
# cv.ols(x,y,k)
cv.ols <- function(x,y,k=4){
  n <- length(y)
  m <- n - k
  y <- as.matrix(y)
  x <- as.matrix(x)
  pred  <- rep(0,k)
  error <- rep(0,k)
  for (i in 1:k){
    ytrain <- y[i:(m+i-1),1]
    xtrain <- x[i:(m+i-1),]
    xtest  <- x[m+i,]
    mod <- lm(ytrain ~ xtrain)
    pred[i]  <- xtest %*% coefficients(mod)[-1] +
     ↪  coefficients(mod)[1]
    error[i] <- y[m+i] - pred[i]
  }
  mae <- sum(abs(error))/k
  mse <- sum(error*error)/k
  q <- ncol(x)
  aic <- log(sum(error*error)/k) + 2*q/k
  list("prediction" = pred, "MAE" = mae, "MSE" = mse, "AIC" =
   ↪  aic)
}

# cv.lasso(x,y,k)
cv.lasso <- function(x,y,k=4){
  n <- length(y)
```

```
  m <- n - k
  y <- as.matrix(y)
  x <- as.matrix(x)
  pred  <- rep(0,k)
  error <- rep(0,k)
  for (i in 1:k){
    ytrain <- y[i:(m+i-1),1]
    xtrain <- x[i:(m+i-1),]
    xtest  <- x[m+i,]
    mod <- gamlr(xtrain, ytrain, gamma=0)
    pred[i]  <- predict(mod, newdata=t(as.matrix(xtest)))
    error[i] <- y[m+i] - pred[i]
  }
  mae <- sum(abs(error))/k
  mse <- sum(error*error)/k
  q <- ncol(x)
  aic <- log(sum(error*error)/k) + 2*q/k
  list("prediction" = pred, "MAE" = mae, "MSE" = mse, "AIC" =
   ↪  aic)
}

# cv.svm(x,y,k)
cv.svm <- function(x,y,k=4){
  n <- length(y)
  m <- n - k
  y <- as.matrix(y)
  x <- as.matrix(x)
  pred  <- rep(0,k)
  error <- rep(0,k)
  for (i in 1:k){
    ytrain <- y[i:(m+i-1),1]
    xtrain <- x[i:(m+i-1),]
```

```
    xtest  <- x[m+i,]
    mod <- ksvm(ytrain ~ xtrain, kernel ="polydot")
    pred[i]  <- predict(mod, newdata=t(as.matrix(xtest)))
    error[i] <- y[m+i] - pred[i]
  }
  mae <- sum(abs(error))/k
  mse <- sum(error*error)/k
  q <- ncol(x)
  aic <- log(sum(error*error)/k) + 2*q/k
  list("prediction" = pred, "MAE" = mae, "MSE" = mse, "AIC" =
   ↪  aic)
}

# cv.pls(x,y,k)
cv.pls <- function(x,y,k=4){
  n <- length(y)
  m <- n - k
  y <- as.matrix(y)
  x <- as.matrix(x)
  q <- ncol(x)
  pred  <- rep(0,k)
  error <- rep(0,k)
  mae_aux <- 1000000000
  for (d in 1:5){
    for (i in 1:k){
        ytrain <- y[i:(m+i-1),1]
        xtrain <- x[i:(m+i-1),]
        xtest  <- x[m+i,]
        mod <- pls(xtrain, ytrain, K=d)
        pred[i]  <- predict(mod, newdata=xtest)
        error[i] <- y[m+i] - pred[i]
    }
```

```
    mae <- sum(abs(error))/k
    mse <- sum(error*error)/k
    aic <- log(sum(error*error)/k) + 2*q/k
    if (mae <= mae_aux){
        pred_aux <- pred
        mae_aux  <- mae
        mse_aux  <- mse
        aic_aux  <- aic
        d_aux    <- d
    }
  }
  list("prediction" = pred_aux, "MAE" = mae_aux, "MSE" =
   ↪  mse_aux, "AIC" = aic_aux, "directions" = d_aux)
}

# cv.dnn(y,df,k)
cv.dnn <- function(response,df,k=4){
  n <- nrow(df)
  m <- n - k
  pred  <- rep(0,k)
  error <- rep(0,k)
  for (i in 1:k){
    dtrain <- as.h2o(df[i:(m+i-1),])
    dtest  <- as.h2o(df[m+i,])
    mod <- h2o.deeplearning(y = response, training_frame =
     ↪  dtrain)
    pred[i]  <- as.vector(h2o.predict(mod, newdata = dtest))
    error[i] <- y[m+i] - pred[i]
  }
  mae <- sum(abs(error))/k
  mse <- sum(error*error)/k
  q <- ncol(df) - 1
```

```
  aic <- log(sum(error*error)/k) + 2*q/k
  list("prediction" = pred, "MAE" = mae, "MSE" = mse, "AIC" =
   ↪  aic)
}

# Load library of machine learning methods for time series
source("libmlts.R")

# Load necessary packages
library(tidyverse)
library(stargazer)
library(ggthemes)
library(ggfortify)
library(GGally)
library(skimr)
library(corrr)
library(knitr)
library(ggplot2)
library(corrplot)
library(RColorBrewer)

# Import Data
data <- read.csv("data.csv")

# Records
t <- nrow(data) # Total number of records
n <- 168   # Number of records with data for the response
 ↪  variable (INCIDENCE_M)
n1 <- n+1

# Tables
```

```
dt <- data[1:n,-1] # Data for training and validation, without
 ↪  the first column (OBS)
dt2 <- data[n1:t,-1] # Data for forecasting, without the first
 ↪  column (OBS)
dt3 <- data[1:t,-1] # Data without the first column (OBS)

# Response (y) and attributes/covariates (x)
y <- dt[,1]     # INCIDENCE_M
x <- dt[,-1]    # Google Trends data
q <- length(x)  # Number of attributes (GT's)

# Quick summary statistics table (with median)
stargazer( as.data.frame(dt3, median = TRUE , header = FALSE ,
 ↪  type = "latex", title = "Summary Statistics"))

# Correlogram
M <-cor(dt)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))

# Scatterplot
plot.scatter <- ggplot(data = dt, aes(x = Tuberculose, y =
 ↪  INCIDENCE_M)) + geom_point(size = 1) + theme_few() + labs(x
 ↪  = "Tuberculose", y = "TB incidence rate", title =
 ↪  "Relationship between the term Tuberculose and TB incidence
 ↪  rate") + theme(title = element_text(size = 8))
plot.scatter # Print scatterplot
# Add regression line to scatterplot
plot.scatter + stat_smooth(method = lm,colour="black")

# Regression between TB incidence and search term "tuberculose"
lm.incidence <- lm(INCIDENCE_M ~ Tuberculose, data = dt)
```

```
stargazer(lm.incidence, type = 'text', no.space = TRUE)

# Multiple Regression between TB incidence and Google searches
lm.incidence <- lm(INCIDENCE_M ~ Tuberculose + tuberculina +
 ↪  sintomas_tuberculose + TB + palidez + dor_peito + arrepios
 ↪  + febre + fadiga + tosse_cronica + suor_noturno +
 ↪  tossir_sangue + perda_apetite + mantoux + PPD + silicose +
 ↪  Mycobacterium + bacilo_koch + diabetes_mellitus, data = dt)

stargazer(lm.incidence, type = 'text', no.space = TRUE)

# CODE ML METHODS
# Number of folds for validation
k <- 36   # 3 years
m <- n - k

# Predictions and errors for cross-validated ML time series
 ↪  models
pred_ols   <- cv.ols(x, y, k)
pred_lasso <- cv.lasso(x, y, k)
pred_svm   <- cv.svm(x, y, k)
pred_pls   <- cv.pls(x, y, k)
pred_dnn   <- cv.dnn("INCIDENCE_M", dt, k)

# Median model (simple combining procedure)
P <- cbind(ols=pred_ols$prediction,
           lasso=pred_lasso$prediction,
           svm=pred_svm$prediction,
           pls=pred_pls$prediction,
           dnn=pred_dnn$prediction)
pred_median <- apply(P,1,median)    # Median by row of P
```

```
error_median <- y[(n-k+1):n] - pred_median
MAE_median <- sum(abs(error_median))/k
MSE_median <- sum(error_median*error_median)/k
AIC_median <- log(sum(error_median*error_median)/k) + 2*q/k
P <- cbind(P, median=pred_median)

# Race between models
MAE <- cbind(ols=pred_ols$MAE,
             lasso=pred_lasso$MAE,
             svm=pred_svm$MAE,
             pls=pred_pls$MAE,
             dnn=pred_dnn$MAE,
             median=MAE_median)
MSE <- cbind(ols=pred_ols$MSE,
             lasso=pred_lasso$MSE,
             svm=pred_svm$MSE,
             pls=pred_pls$MSE,
             dnn=pred_dnn$MSE,
             median=MSE_median)
RMSE <- sqrt(MSE)
AIC <- cbind(ols=pred_ols$AIC,
             lasso=pred_lasso$AIC,
             svm=pred_svm$AIC,
             pls=pred_pls$AIC,
             dnn=pred_dnn$AIC,
             median=AIC_median)
round(P,1)
round(MAE,3)
round(MSE,3)
round(RMSE,3)
round(AIC,3)
```

```
# Forecasting 2021M1 - 2023M4 with the best model
best <- pls(x, y, K=5)
z <- dt2[,-1] # Data without the response
fcast  <- ts(predict(best, newdata=z), frequency=12,
↪  start=c(2021,1))
autoplot(fcast)
```