iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

# Forecasting Models for Portugal’s Inbound Tourism

Nuno Miguel de Castro de Amorim Oliveira Dias

Master in Computer Engineering

Supervisor:
PhD Adriano Martins Lopes
Invited Assistant Professor,
Iscte-IUL

Oct, 2023

iscte

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

# Forecasting Models for Portugal’s Inbound Tourism

Nuno Miguel de Castro de Amorim Oliveira Dias

Master in Computer Engineering

Supervisor:
PhD Adriano Martins Lopes
*Invited Assistant Professor,*
*Iscte-IUL*

Oct, 2023

**Forecasting Models for Portugal’s Inbound Tourism**

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.7) [1].

# Acknowledgements

Words cannot express my gratitude to my superviser, Professor Adriano Lopes, ISTAR researcher at Iscte-IUL, whom without his patience, constant feedback and expertise this dissertation wouldn't be possible. I am also thankful to him for always finding time to help me through this journey.

Thanks should also go to Professor Fernando Brito e Abreu, ISTAR researcher at Iscte-IUL, who introduced me, together with Professor Adriano Lopes, to the theme of this dissertation by participating on the RESETTING project.

Many thanks to the whole RESETTING project team, but specially those involved in WP2, with whom technology was developed to help SMEs recovery financially from the COVID pandemic. I had the pleasure of working with my fellow MSc student André Garcia, also in WP2, with whom I shared not only the compromise and commitment to this honourable project but also lots fun in the process. I would like to extend my sincere thanks to BRU, at Iscte-iul, for their financial support through a research fellowship for the participation on the RESETTING project.

I could not have undertaken this journey without the help and love of my family and friends. I am very grateful specially for all the support from both of my parents and my brother, who helped me time and time again, without whom this journey wouldn't be possible. I would like to extend my sincere thanks to my godmother, by being so supportive every step of the way, to both my grandmothers, my aunties, uncles and cousins, for all of their kind words through this process.

# Resumo

Em Portugal, o mercado do turismo tem um peso considerável no PIB e, uma vez que as Pequenas e Médias Empresas (PMEs) do turismo foram as mais impactadas pela pandemia, qualquer contributo para a sua recuperação é bem-vindo. É neste contexto que surge o projeto europeu RESETTING, o qual visa disponibilizar ferramentas que possam auxiliar estas PMEs. Este trabalho de investigação enquadra-se neste propósito, estudando e disponibilizando modelos de previsão para o sector do turismo, de modo a que as PMEs portuguesas possam utilizar e, assim, ajustar os respetivos produtos e serviços à procura.

Após estudo da área de investigação relacionada com modelos de previsão do turismo, constatámos quais seriam os algoritmos mais utilizados e que dados deveriam ser utilizados. Este trabalho de investigação segue uma metodologia para manipulação de dados – CRISP-DM. Assim, começamos por obter, preparar os dados de interesse e analisar os mesmos para os entender melhor. Como resultado, construimos duas classes de modelos de previsão: modelos de referência e modelos de *deep learning*. Os modelos mais complexos de *deep learning* são baseados nos algoritmos *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU).

Finalmente, concluimos que ambos os modelos de deep learning funcionam muito bem com dados de frequência diária, superando os modelos de referência. No entanto, não funcionaram tão bem para dados de frequência mensal quando comparados com os modelos de referência. Concluindo, os modelos apresentados podem auxiliar as PMEs a obter previsões sobre procura de turismo, quer no horizonte de curto-prazo, quer no horizonte de médio-prazo.

## Palavras-chave

Previsão do Turismo, turismo de Portugal, modelos de previsão

# Abstract

In Portugal, the tourism sector has a considerable weight in the GDP and, as Small and Medium Enterprises (SMEs) were the most impacted by the pandemic, any contribution to help them to recover is welcome. It is in this context that the European RESETTING project has been setup, aiming to provide tools that can help these SMEs. This research work fits into the purpose, by studying and providing tourism forecasting models so Portuguese SMEs can use them and adjust their products and services to tourism demand.

After analysing the research field related to tourism forecasting, we have figured out the most popular algorithms and what data should be used. This research study follows a well-known methodology to deal with data – CRISP-DM. Hence, we collected and prepared the data of interest, then we analysed it so we could better understand it. Finally, we have created two classes of forecasting models: baseline and deep learning forecasting models. The more complex deep learning models were based on Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) algorithms.

From the experiments carried out we concluded that both deep learning models work well with data of daily frequency and were outperforming the baseline models. However, it did not work that well in the case of monthly data, when comparing with the baseline models. In the end, the models provided can help SMEs to obtain tourism demand predictions, both for the near-term – nowcasting – and the medium-term horizons.

## Keywords

Tourism forecasting, Portugal's tourism, forecasting models

# Contents

**1 Introduction** **1**
1.1 Motivation . . . 1
1.2 Objectives . . . 2
1.3 Scientific Methodology . . . 2
1.4 Document Structure . . . 3

**2 Related Work** **5**
2.1 Literature Review Process . . . 5
2.2 Related Work . . . 5
2.2.1 Statistical algorithms . . . 6
2.2.2 Econometric algorithms . . . 7
2.2.3 AI-based algorithms . . . 8
2.2.4 Bayesian Optimization . . . 11

**3 Methodology** **15**
3.1 Introduction . . . 15
3.2 Data Collection . . . 16
3.3 Data Preparation . . . 17
3.4 Baseline Forecasting . . . 17
3.5 Deep Learning Forecasting . . . 18
3.5.1 Long Short-Term Memory (LSTM) . . . 19
3.5.2 Gated Recurrent Unit (GRU) . . . 20

**4 Implementation** **21**
4.1 Data Collection . . . 21
4.1.1 SEF Data . . . 21
4.1.2 INE Data . . . 21
4.1.3 Google Trends Data . . . 22
4.2 Data Preparation . . . 23
4.2.1 SEF Data . . . 23

4.2.2 INE Data . . . . . . . . . . 24
4.2.3 Google Trends Data . . . . . . . . . . 24
4.3 Data Analysis and Visualization . . . . . . . . . . 26
4.3.1 SEF Data . . . . . . . . . . 26
4.3.2 INE Data . . . . . . . . . . 28
4.3.3 Google Trends Data . . . . . . . . . . 31
4.3.4 Comparative Analysis . . . . . . . . . . 31
4.4 Baseline Forecasting . . . . . . . . . . 32
4.4.1 SEF Data . . . . . . . . . . 32
4.4.2 INE Data . . . . . . . . . . 33
4.4.3 Google Trends Data . . . . . . . . . . 34
4.4.4 Comparative Analysis . . . . . . . . . . 35
4.5 Deep Learning Forecasting . . . . . . . . . . 35
4.5.1 SEF Model . . . . . . . . . . 35
4.5.2 INE Model . . . . . . . . . . 36
4.5.3 SEF + Google Trends Model . . . . . . . . . . 37
4.6 Models Evaluation . . . . . . . . . . 38
4.7 Forecasting . . . . . . . . . . 39
4.8 Summary . . . . . . . . . . 40

**5 Web Application for Deployment 41**
5.1 Introduction . . . . . . . . . . 41
5.2 Architecture . . . . . . . . . . 41
5.3 Implementation . . . . . . . . . . 42

**6 Conclusions and Future Work 45**
6.1 Contributions . . . . . . . . . . 45
6.2 Limitations and Future Work . . . . . . . . . . 45

**Bibliography 47**

**Appendices**

**A Appendix 1 51**
A.1 INE Baseline Prediction Visualization . . . . . . . . . . 51
A.2 INE Data Visualization . . . . . . . . . . 55
A.3 Google Trends Baseline Prediction Visualization . . . . . . . . . . 61
A.4 Google Trends Data Visualization . . . . . . . . . . 65
A.5 SEF Baseline Prediction Visualization . . . . . . . . . . 67
A.6 SEF Data Visualization . . . . . . . . . . 70

# 1 Introduction

In this Chapter, we will discuss the motivation and scope behind the theme of this dissertation, as well as the research questions we aim to answer. Furthermore, we present the structure of the work that is going to be done, the main objectives and the methodologies that we will use in order to accomplish those objectives.

## 1.1 Motivation

The tourism sector is increasingly relevant to the world's economy and more specifically in Portugal's case. According to a yearly document published by 'Turismo de Portugal' (entitled *Turismo em Números*), the Portuguese tourism authority responsible for the promotions, appreciation and sustainability of the sector, from 2010 until 2019 there was a 7.2% average annual growth rate. Then, due to the COVID pandemic, it suffered a huge decrease in 2020 and 2021, -64% and -48% respectively (both compared to 2019). Also, the tourism weight for the total Gross Domestic Product (GDP) has been growing from 2016 to 2019 (from 9.7% to 11.8%) with a much-expected decrease in 2020 and 2021, 6.6% and 8% respectively. The pandemic had a huge impact on the sector since there were restrictions regarding travelling, quarantine measures or the need for a negative COVID test or vaccination certificate (in Portugal, at least).

In 2021, international tourist arrivals grew 12.7% in comparison to the previous year, according to the World Tourism Organization. Even so, this value remained 68.7% below 2019's values. Regarding Europe, 2021 international tourist arrivals have grown 25.7% in comparison with the previous year, while being 59.4% bellow when comparing it with 2019 values. On the other hand, the south/Mediterranean Europe was the sub-region with the biggest share of tourist arrivals in Europe, amounting to 46%.

Concerning Portugal, and according to a yearly document published by the National Institute of Statistics (INE) (entitled *Estatísticas do Turismo*) the country was in the fifth position in terms of tourist imports and exports, surpassed only by Spain, Greece, Italy and Croatia.

Portugal is divided into seven distinct touristic regions: North, Center, Lisbon Metropolitan Area, Alentejo, Algarve, Autonomous Region of Madeira and Autonomous Region of Azores. The average bed occupation ratio in Portugal remained at 33.1%, with Madeira and Azores regions leading the way (47.5% and 37.9%, respectively), followed by Algarve (34.7%), Alentejo (34.3%), Lisbon (30.4%), North(30.3%) and Center (26.1%). Despite these values, Algarve had the most considerable value of total profits, 695.2 million euros, followed by Lisbon, North and Madeira, 443.5, 274.7 and 246.1 million euros respectively.

Alongside this solid tourism growth over the last few years, not only in Portugal but worldwide, there has been a growing number of tourism demand forecasting studies, focusing mainly on international tourist flows. These studies are important to worldwide tourism enterprises who want to quickly adapt to this volatile market. As mentioned before, tourism has a big share in Portugal's GDP and it took a big hit with the COVID pandemic. In particular, the Portuguese SMEs were badly hit. Hence, it is worth considering the following question: "How can we help SMEs, the most economically affected during the pandemic, to recover from the economic situation they were put in 2020 and 2021, with forecasting models?"

## 1.2 Objectives

The main goal of this work is to create forecasting models that can provide predictions about tourism flows, in particular for the Portuguese market. It is worth remembering that, according to the European Comission, European SMEs represent about 99% of all businesses in Europe. So in Portugal.

The RESETTING project, which aims at promoting the competitiveness of European SMEs in the tourism sector by using sophisticated and innovative technological solutions, gave us the framework to create and deliver the forecasting models.

Given the context above, in this dissertation we want to address the two following research questions:

- "How to provide tools to SMEs that allow them to adjust their products and services to international tourism demand?"
- "How to deliver tourism demand predictions fast enough and with the high quality possible, so it can help SMEs to stay competitive?"

## 1.3 Scientific Methodology

To accomplish the goals set before, the 'Design Science Research Methodology' (DSRM) was followed. This methodology fits into our work since it is a problem-solving paradigm that seeks to address our society's challenges through digital innovation, by producing new artefacts. In this case, one or more forecasting models.

As for the production of the forecasting models, we do so by following primarily the CRoss Industry Standard Process for Data Mining (CRISP-DM), a popular methodology in the field of data science [2].

CRISP-DM consists of six stages: (i) *Business Understanding* – in order to understand the objectives and requirements of the project from a business perspective; (ii) *Data Understanding* – in order to identify, collect, and analyze the data sets; (iii) *Data Preparation* – to determine which data sets will be used, to clean the data, adding new attributes if necessary, creating new data sets from multiple data sources and possibly to reformat data; (iv) *Modeling* – to select which algorithms to use by comparing different models with each other; (v) *Evaluation* – to evaluate which model best meets the business success criteria; and finally (vi) *Deployment* – to generate a report, plan monitoring and maintenance, and to deploy the model. Chapter 3 presents more details about the use of this methodology.

## 1.4 Document Structure

Apart from this Chapter, this document contains five more Chapters. In Chapter 2, we start by introducing the process of a systematic literature review, and then presenting the outcome of that review. After that, in Chapter 3, we describe a modelling approach to create the forecasting models for the tourism sector. Then in Chapter 4 we implement what was established in the Chapter 3. In Chapter 5 we explain how and where the predictions made by the forecasting models will be available to users, namely by presenting a web-based application to achieve that. Finally, in Chapter 6, we present an overview of the contributions of this work, as well as possible paths for further research.

# 2

# Related Work

## 2.1 Literature Review Process

In order to find useful information that helps reaching our goals of elaborating on related work, searching in the right places is paramount. That is why Scopus and Google Scholar were the main sources of articles for this Chapter.

First, we should define search queries. In Scopus, the following query was used to search for articles:

*("tourism" OR "tourists" OR "tourism companies" OR "tourism flow") AND ("prediction model" OR "prediction models" OR "forecasting model" OR "forecasting models") AND ("time series" OR "econometric" OR "ai-based").*

The results obtained from this search in Scopus were not good enough and complete to write properly the related work. Indeed, some articles simply did not exist in the Scopus database. Then we switched to Google Scholar and the same keywords were used again.

As a result of this search, there were more than 80 articles available for analysis. Actually, the abstracts of more than 80 articles were read but only 18 were considered as the main source of information for this literature review.

Another useful tool that was employed is Mendeley. It was a way to simplify the workflow of reference management, by storing and organizing every reference in one library. The *snowballing* technique was also used in our systematic literature review. This technique consists of using the citation list of an article to identify additional relevant articles.

## 2.2 Related Work

Given the process established above, this section presents the outcome of the work carried out. Each method is rigorously explained, and a few examples are given of articles that use a particular method, with information regarding the problem the article was attempting to solve, the data used and its source, what algorithms were used and the results obtained.

At the end of this section, Table 2.1 provides a summary about all the articles referred to in the section.

Based on the collected information after applying the process set, we can identify three main types of algorithms and/or models used in the context of tourism demand forecasting: statistical, econometrics and AI-based. Nonetheless, and as expected, sometimes the dividing line is not so clear as many of the solutions found are hybrid.

On the other hand, it is important to recognize that the data of interest is inherently time-series data. Recall that a time series uses past data to establish historical patterns and attempts to predict future events by identifying trends, slopes or cycles in certain variables over a certain period of time. That is why, unlike other types of algorithms that use random sample data, time series work with successive values that represent consecutive measurements over several days, months, semesters or years. These types of algorithms are frequently used in business and finance. For example, to make sales or stock price predictions. They can be split into two different groups: basic and advanced techniques.

### 2.2.1 Statistical algorithms

In the context of time-series data, there are a few basic statistical techniques, like the Naive Bayes, auto-regressive (AR), single exponential smoothing (ES), moving average (MA) and historical average (HA) are placed in this group due to their implementation simplicity and ability to capture historical patterns quite well.

One of the most popular basic time series techniques, the Naive Bayes approach, was used in 2011 by the author of [3] to understand the effect of temporal aggregation on forecast accuracy. Using 366 monthly series, 427 quarterly series and 518 yearly series supplied by Tourism Australia, the Hong Kong Tourism Board and Tourism New Zealand and comparing the Naive approach with several other techniques (like ARIMA, Forecast Pro, Theta method and Damped Trend), it was concluded that for yearly data, Naive produces the most accurate forecasts.

Advanced techniques differ from basic ones by the use of additional time series attributes, such as trends and seasonality. One of the most popular of this kind is the auto-regressive integrated moving average (ARIMA), a flexible tourism demand forecasting model that combines features from AR and MA techniques and adds an integrated component (I), where the data values are replaced by the difference between their values and previous values, to eliminate the seasonal component.

In 2012, in an attempt to forecast future arrivals in Greece, the author of [4] used ARIMA, comparing it with double exponential smoothing and the Holt-Winters method, and using the monthly data on tourist arrivals from January 1977 to December 2009. It was concluded that ARIMA outperformed both of the previously mentioned algorithms.

In 2018, to produce accurate forecasts using tourism data that is affected by social, economic and environmental factors, the author of [5] tested several algorithms, such as ARIMA, ANN and a hybrid of those two, to reach the optimal performance. Using

Malaysia tourist arrivals from 1998 to 2016 data obtained by the Department of Malaysia Tourism, it was concluded that the hybrid method had produced better results than its parts.

### 2.2.2 Econometric algorithms

Econometric algorithms use statistical and mathematical tools and economic theories to try and establish cause-effect relationships between economic variables and tourism demand, with a focus on determining how these variables affect tourism demand in the future. Unlike time series, which try to establish a trend or a pattern in past data, econometric algorithms focus on how certain variables affect future tourism demand.

The main types of econometric algorithms are auto-regressive distributed lag (ADLM), error correction (ECM), vector auto-regressive (VAR) and time-varying parameter (TVP).

In 2011, the auto-regressive distributed lag model, ADLM, was used by the author of [6] to forecast short to mid-term air traffic flows, helping airlines and regulatory authorities to make decisions. Using the data from the Annual Statement of movements, of both cargo and passenger, released by the UK's Civil Aviation Authority from 1961 to 2002 and comparing ADLM with TVP and VAR, it was concluded that the first one outperforms the alternatives.

The error correction model, ECM, was used in 2010 by the author of [7] to evaluate the impact of climate change on the appeal for Caribbean destinations, in comparison with a Naive model. Using the tourist arrivals data to the 18 Caribbean countries by the Caribbean Tourism Organisation's Annual Statistical Digest between the period of 1980 and 2004, it was concluded that ECM has outperformed the Naive model.

In 2006, the vector auto-regressive model, VAR, was used by the author of [8] in an attempt to examine the performance of various vector auto-regressive algorithms such as Bayesian VAR, VAR and AR. Using annual tourism arrivals data from 1973 to 2000 from Visitor Arrivals Statistics published by the Hong Kong Tourism Board, it was concluded that, by using a Bayesian approach to VAR, there was an improvement in the forecasting performance of the alternatives.

A time-varying parameter model, TVP, together with static regression, ECM, VAR and ADLM, was used in 2003 by the author of [9] to determine the accuracy of different econometric algorithms in the Denmark international tourism demand. The annual inbound data from six different countries (Germany, Netherlands, Norway, Sweden, UK and USA) from 1969 to 1997 and it was concluded that the TVP model produces the most accurate forecasts one and two years ahead.

As with all algorithms we will discuss, both time series and econometrics algorithms have some downsides such as the need for the data probability distribution and the specification of the chosen model before implementation.

As it was said in [10], "time-series and econometrics algorithms rely on the stability of historical patterns and economic structure".

### 2.2.3 AI-based algorithms

AI-based algorithms are more recent and importantly can capture nonlinear relationships between variables and patterns and between time series and variables with an external cause. They are considered to be accurate and with increasing forecasting performance. Non-parametric AI algorithms do not need specification or data probability distribution to make predictions, which happens with both time series and econometric algorithms. These algorithms are not the best at explaining their predictions, taking a 'black box' approach, that is, the exact process being 'unknown' to the researcher.

AI-based also has some downsides, as it was said in [10]: "Artificial intelligence algorithms are dependent on the quality and size of available training data". Therefore, data selection, understanding, cleansing and preparation are (most likely) the most important steps in building an AI-based tourism demand forecasting system.

There are six main types of AI-based algorithms used in tourism demand forecasting: Artificial Neural Networks (ANNs), Rough Sets Approach, Support Vector Machines (SVMs), Fuzzy Time Series, Grey Theory and Genetic Algorithms (GAs) combined with Support Vector Regression algorithms (SVR) [11]. In the following section, we will discuss some of the related works.

#### 2.2.3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a type of AI-based method that was inspired and tries to simulate the structure and processes of the human brain. Like the human brain, the ANNs also have neurons linked to each other, called nodes, in the different layers of the network. This method usually has three distinct types of layers: input, hidden and output, where each node is linked to another by a weighted connection. Is through these connections that ANNs learn, by adjusting the weight of the links and by creating something like a 'neural pathway'. All these weight-adjusted values are imputed to some nodes using a scalar function to aggregate them and then using a transfer function to produce its output.

Back-propagation Neural Networks (BPNNs) are feed-forward networks, meaning that connections between the nodes do not form a cycle, and are one of the most widely used ANNs. It's a gradient steepest descent training algorithm (an iterative optimization algorithm for finding a local minimum of a differentiable function) that needs a learning rate (the step size) to be set out first. The learning rate determines how fast or if the algorithm can converge to a solution: if the learning rate is too high, the algorithm can't converge to a solution; if the learning rate is too low, the algorithm may take a long time or way too many iterations to converge to a solution.

In 2012, with the need for more accurate forecasts of tourism demand to reduce risk and uncertainty, the author of [12] used EMD (Empirical Mode Decomposition), ARIMA and BPNN algorithms, and the international tourist arrivals to Taiwan from Japan, Hong Kong and Macao, obtained from Taiwan tourist authorities from January 1971 to August

2009. The period from January 1971 to August 2001 was used as the training sample (80% of the total data) and the period from September 2001 to August 2009 was used as the testing sample (20% of the total data). As a result, EMD-BPNN produced the lowest prediction error and outperformed single BPNN and traditional ARIMA algorithms.

In 2018, in an attempt to produce precise forecasting of "tourism volume" data, the author of [13] used a hybrid solution (PCA-BPNN-ADE) that combined PCA (Principal Component Analysis, to reduce the dimension of the data set, and add interpretability and minimize information loss), BPNN and ADE (an algorithm to prevent the solution from falling into a local optimum and to balance between optimal solution and population diversity) and compared it with PCA-BPNN, BPNN, PCA-VAR, VAR and ARIMA to find which of those had the best performance. Using tourist arrivals to Beijing from 2011 to 2016 data, that was obtained from the Beijing Statistical Information Website, it was found that PCA-BPNN-ADE outperforms the other algorithms, being more efficient for tourist volume prediction problems.

As with other types of ANNs, multi-layer perceptron neural networks (MLPNN) consist of those three layers we spoke about earlier. Being similar to other feed-forward networks and using back-propagation, it tries to solve problems that are not linearly separable. In 2011, to map non-linear relationships between inputs and outputs of UK inbound tourism quarterly arrivals by purpose of visit from 1993 to 2007, the author of [14] used an MLPNN model. The results show that this model can outperform substantially linear combination forecasting algorithms, both in seasonal and semi-seasonal data.

Radial basis function network (RBF) is another type of feed-forward ANN used for function approximation problems that stand out for its fast learning speed. RBF networks have three layers, as many other ANNs we had previously mentioned, with the particularity of having only one hidden layer. This layer receives the data, which might be non-linearly separable, from the input layer and often transforms it into a higher dimensional space, so it can be more linearly separable. In 2015, using monthly data of tourist arrivals over the period from January 2001 to July 2012 to compare multiple ANN tourism demand forecasting algorithms (RBF networks, MLPNN and Elmans' NN), the author of [15] concluded that both RBF and MLPNN outperformed Elmans' NN. Additionally, RBF outperforms MLPNN when there are no additional lags introduced in the algorithm and vice versa.

As the method mentioned above, generalized regression neural network (GRNN), a feed-forward neural network, is often used for function approximation and it stands out from other ANNs for its four layers: input, hidden, summation and division. Also, this method does not require iterative training like other ANN algorithms (BPNs, for example). To predict tourism demand ahead of the year 2014, to welcome tourists at their cruise destination ports and prepare for their arrival, the author of [16] used MLPNN, RBF and GRNN forecasting algorithms. Monthly cruise tourist arrivals to Izmir cruise port in the period of January 2005 to December 2013 were used and the conclusion was that GRNN and MLPNN were outperformed by RBF.

Elman's neural network is also a feed-forward network based on BPN with the addition of an undertake layer to remember the output of the hidden layer, be sensitive to historical data and have the ability to adapt to time-varying characteristics. In 2003, to investigate the application of different algorithms to predict travel demand to Hong Kong from six different countries (USA, UK, Japan, Korea, Singapore and Taiwan), the author of [17] used Elman's NN and other forecasting algorithms like exponential smoothing and ARIMA. Using the yearly tourist arrival data of the Hong Kong Tourist Association from 1974 to 2000, it was found that Elman's NN outperforms the other options.

Recurrent Neural Networks (RNN) are another type of ANN, but bi-directional, not a feed-forward one, where the output from a certain step is fed as input to the next. What makes it different is the *memory state* or *hidden state*, which remembers the previous state inputted to the network. This neural network type works well with sequential data (time series). Two examples of RNNs are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). In 2018, to improve the current accuracy of tourism flow predictions, the author of [18] used LSTM and compared it to ARIMA and BPNN. Using travel data from 2013 to 2015, it was concluded that LSTM outperformed both ARIMA and BPNN.

#### 2.2.3.2 Support Vector Machines (SVMs) and Support Vector Regression algorithms (SVRs)

Support Vector Machine (SVM) is another AI-based method that is used for regression and classification assignments. The goal of this algorithm is to find a hyperplane in a space with N (number of features) dimensions to categorise each data point with the maximum distance between this hyperplane and the points of each class. The bigger the margin, the bigger the confidence in the classification.

These algorithms have been adapted to solve nonlinear regression estimation problems – the Support Vector Regression (SVR). SVRs, instead of trying to minimise the training error (like ANN algorithms do), try to minimise an upper bound on the generalisation error, this is, find the hyperplane with the maximum number of points.

In 2011, to help tourism planning and forecasting by using a less conventional AI-based model, SVR combined with a chaotic genetic algorithm (CGA), and comparing it to other forecasting algorithms (ARIMA and SARIMA), the author of [19] used annual tourist arrival data to Barbados from 1956 to 2003, obtained from Barbados Statistical Service and concluded that SVRCGA gives better results than the other forecasting algorithms.

#### 2.2.3.3 Rough Sets Approach

The rough set method is convenient for dealing with fuzzy data and discovering hidden patterns or relationships in the variables. The imprecise or vague data set is replaced by a pair, the lower and upper approximations, returning a crisp set. The lower approximation set contains all elements that confidently belong to a certain concept while the upper approximation set contains all elements that have the possibility of belonging to a certain concept. Also, an advantage of the rough set method is the ability to deal with both quantitative and qualitative variables.

In 2008, the author of [20] applied rough sets to study vague and imprecise data. Using U.S. and U.K. tourism demand for Hong Kong obtained by the Hong Kong Tourism Board from 1997 to 2002, it was concluded that this method generates transparent and comprehensible decision rules for planning and decision-making.

#### 2.2.3.4 Fuzzy Time Series and Grey Theory

In 2004, to forecast the tourism demand in a "capacity-constrained service industry", the author of [21] used a Fuzzy Time Series, a Grey Theory method and a hybrid Grey Theory model (Markov residual modified model), which don't need a large data set and long past time series. The data used in this article was the tourism arrivals to Taiwan from Hong Kong, the USA and Germany from 1989 to 2000, obtained by the Tourism Bureau of the Republic of China and it was concluded that the Fuzzy Time series had the worst performance of the three algorithms, with the Grey method performing very well with the Hong Kong and USA data set and being Markov the best performer with Germany's data set.

### 2.2.4 Bayesian Optimization

To tune the hyper-parameters for each model, like the hidden size (the number of features in the hidden state), the number of layers to stack or the learning rate, Bayesian optimization can be used. Since our goal is to get the best predictions possible, minimizing the error of the metrics, the more optimized our model is the better.

Bayesian Optimization works the following way: a certain number, defined under an inputted interval, of hyper-parameter values (called data points) are chosen randomly in the first iteration; then, for each iteration, the algorithm which we are trying to minimize is run several times with multiple random data points and the latter with the best result or the highest uncertainty (with the potential of finding an even better result) is found. More information about Bayesian Optimization can be found in this article: [22]

Table 2.1: Summary of most relevant articles.

| Article | Problem | Algorithms | Results |
|---|---|---|---|
| [3] | Understand the effect of temporal aggregation on the forecast accuracy | Naïve Bayes, ARIMA, Forecast Pro, Theta method and Damped Trend | For yearly data, Naive produces the most accurate forecasts |
| [4] | Forecast future arrivals in Greece | ARIMA, double exponential smoothing and Holt-Winters method | ARIMA outperformed both of the other used algorithms |
| [5] | Produce accurate forecasts using tourism data that is affected by social, economic and environmental factors | ARIMA, ANN and a hybrid of ARIMA and ANN | Hybrid method produced better results than its individual parts |
| [6] | Forecast short to mid-term air traffic flows, helping airlines and regulatory authorities to make decisions | ADLM, TVP and VAR | ADLM outperforms the other algorithms |
| [7] | Evaluate the impact of climate change on the appeal for Caribbean destinations | ECM, Naïve Bayes | ECM has outperformed the Naive model |
| [8] | Examine the performance of various vector auto regressive models | Bayesian VAR, VAR and AR | Using a Bayesian approach to VAR, there was an improvement over the forecasting performance of the alternatives |
| [9] | Determine the accuracy of different econometric models in the Denmark international tourism demand | TVP, static regression, ECM, VAR and ADLM | The TVP model produces the most accurate forecasts one and two years ahead |
| [12] | The need for more accurate forecasts of tourism demand to reduce risk and uncertainty on predictions | EMD, ARIMA and BPN | EMD-BPN produced the lowest prediction error and outperformed single BPN and traditional ARIMA models |

Continuation of Table 2.1

| Article | Problem | Algorithms | Results |
|---|---|---|---|
| [13] | Produce precise forecasting of "tourism volume" data | PCA-BPNN-ADE, CA-BPNN, BPNN, PCA-VAR, VAR and ARIMA | PCA-BPNN-ADE outperforms the other algorithms, being more efficient for tourist volume prediction problems |
| [14] | Map non-linear relationships between inputs and outputs of UK inbound tourism quarterly arrivals by purpose of visit | MLPNN | This model can outperform substantially linear combination forecasting algorithms, both in seasonal and semi-seasonal data |
| [15] | Compare multiple ANN tourism demand forecasting models | RBF networks, MLPNN and Elmans' NN | Both RBF and MLPNN outperformed Elmans' NN. Additionally, RBF outperforms MLPNN when there are no additional lags introduced in the algorithm and vice versa. |
| [16] | Predict tourism demand ahead of the year 2014, to welcome tourists at their cruise destination ports and prepare their arrival | MLPNN, RBF and GRNN | GRNN and MLPNN were outperformed by RBF |
| [17] | Investigate the application of different models to predict travel demand to Hong Kong from six different countries (USA, UK, Japan, Korea, Singapore and Taiwan) | Elman's NN, exponential smoothing and ARIMA | Elman's NN outperforms the other options |
| [18] | Accurate tourism flow predictions | LSTM, ARIMA and BPNN | LSTM outperformed both algorithms |

Continuation of Table 2.1

| Article | Problem | Algorithms | Results |
|---|---|---|---|
| [19] | Help tourism planning and forecasting by using a less conventional AI-based model | SVR combined with a chaotic genetic algorithm (CGA), ARIMA and SARIMA | SVRCGA gives better results than the other forecasting models |
| [20] | Study vague and imprecise data | Rough sets | This method generates transparent and comprehensible decision rules for planning and decision making |
| [21] | Forecast the tourism demand in a "capacity constrained service industry" | Fuzzy Time Series, a Grey Theory method and a hybrid Grey Theory model (Markov residual modified model) | Fuzzy Time series had the worst performance of the three models, with the Grey method performing very well with the Hong Kong and USA data set and being Markov the best performer with Germany's data set |
| | | | |

# 3 Methodology

## 3.1 Introduction

As in any research work, the process and the course of action need to be clearly defined. This is what we will do in this Chapter.

In order to build accurate forecasting models, there are a few aspects to be taken into account. First, it is necessary to collect as much data as possible (and with the highest quality). In this case, Portugal's tourism, social or economic data. Then such data needs to be properly cleaned and organized (pre-processing stage) and then analyzed so we can understand and make sense of the data. After that, feature engineering is a important step towards making accurate predictions later on. In that regard, we may extract additional information and create new features. Only then we can use the final data as input for the forecasting algorithms.

The models to be considered in this work are split into two classes: baseline and the deep learning models. Furthermore, the working pipeline to be followed is represented in Fig. 3.1.

As mentioned above, it is important to discover the nature of raw data we are dealing with. Hence, regardless of the processing stage we are at in a particular moment in time, the use of visual tools is paramount to figure out patterns and trends in data, whether there are ere big volumes of data or not. We elect Plotly as main tool due its sophistication and effortless interactive visualizations it provides. Indeed, Plotly was chosen over the popular data visualization library Matplotlib.

In the following sections, we will discuss how to proceed on each of the stages of the working pipeline.

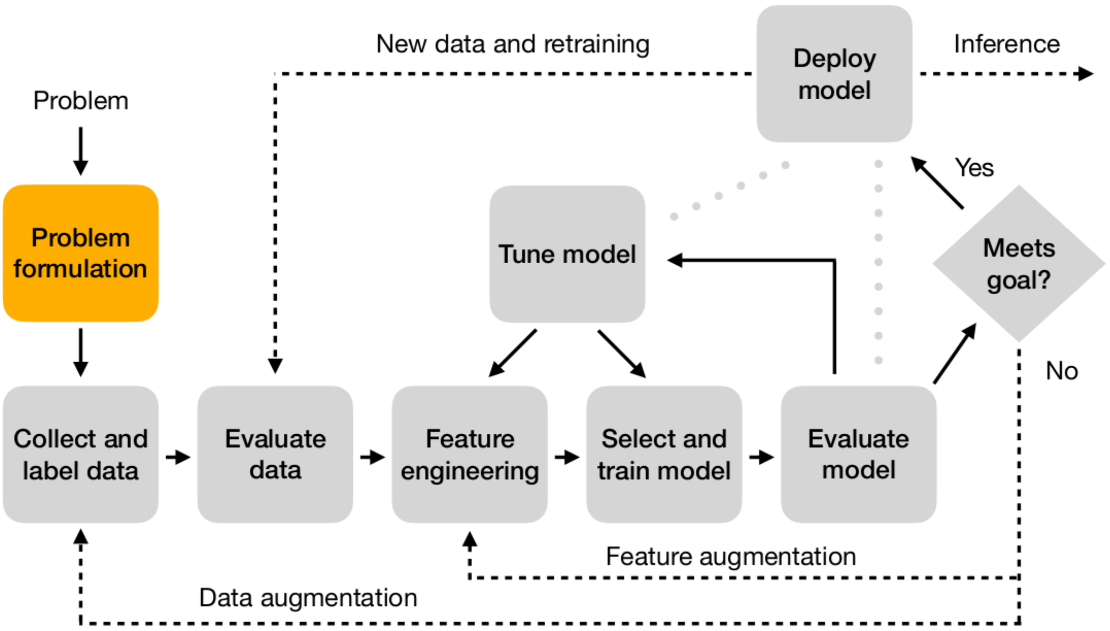


Figure 3.1: Pipeline underlying the process of creating and using forecasting models.

## 3.2 Data Collection

The data to be used is of utmost importance. Not only the topics it is related to but the sources themselves, as well as the frequency/timestamp it has been acquired and/or generated. Recall that the information we are dealing with is essentially classified as time-series data. Hence, regarding the data we have collected, we have to consider:

- The topic/issue of concern.
- The source of information.
- The frequency/timestamp that data has been acquired and/or generated.

To this end, we had to decide from where and which data was collected, so some criteria were needed. After careful thinking, the conclusion was that only monthly or more frequent data should be considered, and the data had to be related or have some influence on tourists' flow inside or to Portugal. That being said, the primary sources of data were the Portuguese Immigration and Border Service (SEF) , the National Institute of Statistics (INE) and Google Trends. These sources were chosen, not only because they met the criteria mentioned above but also for the easy and fast data access. In the case of INE and Google Trends, there was no need to ask formally for data and, on top of data, to wait several weeks to get a response. As for SEF, it was possible to use some data that was provided directly by Tourism of Portugal, under a research collaboration agreement between Tourism of Portugal and Iscte.

## 3.3 Data Preparation

After acquiring the raw data, the next key step is to clean, integrate and transform the dataset so it can be properly visualized and analysed. Some actions included here are the removal/imputation of missing values, dealing with outliers and duplicates (cleaning stage), combining data from multiple sources into a single dataset (integration stage), the normalization, standardization and discretization of data (transformation stage) and reducing the size of the dataset but retaining the crucial information (reduction step).

The data preparation was done using a wide range of tools. For example: Microsoft Excel, Jupyter Notebooks, some Python libraries like Pandas and Matplotlib and Pyspark, which is a Python API for Apache Spark. Afterwards, in Chapter 4, we will further discuss these tasks, with focus on each category of incoming data.

## 3.4 Baseline Forecasting

The purpose of baseline forecasting is to provide reasonable predictions without requiring a lot of time to build the models and use them. These models are important to understand whether or not we should proceed with the data we have chosen to use or not, and without spending too much time on building the models. A baseline prediction shows how more sophisticated models will perform on our time series [23].

To evaluate the quality of the predictions of these baseline models, we rely on some metrics. There are two main types of metrics:

- **Intrinsic** – to measure the performance by comparing the forecast with the actual values. This type divides itself into four major base errors: absolute error, squared error, percent error and symmetric error.

- **Extrinsic** – use of an external reference or benchmark in addition to the forecast and the actual values. This type divides itself into two major base errors: relative errors and scaled errors.

For this work, two types of metrics were chosen: *Absolute Error*, specifically the Mean Absolute Error (MAE), and *Squared Error*, specifically the Mean Squared Error (MSE). According to the author of [23], these metrics are both symmetric losses, unbiased from the under/over-forecasting perspective and best for single time series datasets. The mathematical equations are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

| | |
|---|---|
| $n$ | total number of data points |
| $Y_i$ | observed values |
| $\hat{Y}_i$ | predicted values . |

As for the models to be created, they were divided into three categories:

- **Baseline** – it includes different approaches of Naive models – Naive Mean, Seasonal, Drift and Moving Average.
- **Statistical** – it includes ARIMA, AutoARIMA and Exponential Smoothing.
- **Regression** – it includes Regression Model and Random Forest.

These models were chosen because, as it was concluded in Chapter 2, they are very popular and also used when looking at tourism forecasting-related. For example, Naive models are used in [3] and [7], and ARIMA models are used in [4] and [12].

As for the implementation of these models, the Darts Python library was used. This library is fast, easy to use and focuses on time series data predictions. Furthermore, not only does it contain a large variety of models, like ARIMA or Naive, but also deep neural networks, and supports both univariate and multivariate time series forecasting.

## 3.5 Deep Learning Forecasting

Deep learning is the type of model that was used to deliver the most accurate predictions possible. It was a more complex approach. With a strong baseline forecasting (from the section above) we are capable of comparing results with the deep learning ones. In that respect, the same two metrics were used – MSE and MAE. For the development and training of the deep learning models, it was used the Pythorch machine-learning library that focuses on neural network models.

Based on our own research, there are two Recurrent Neural Network (RNN) models that stand out when dealing with time-series data: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). RNNs work with sequential data (time series) step by step, keeping a *hidden state* which is updated at each time step with the new input and the previous state. The RNN architecture is described in Fig. 3.2. Further details can be found in the next two sections.

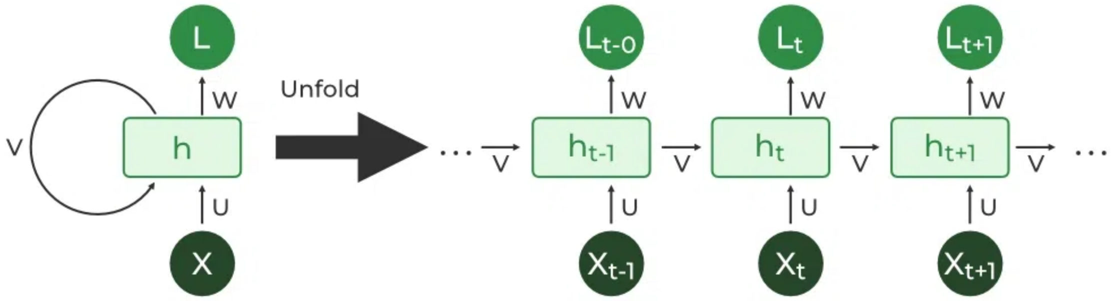


Figure 3.2: Recurrent Neural Network (RNN) architecture. Source: RNN

### 3.5.1 Long Short-Term Memory (LSTM)

LSTM is a type of RNN that allows long-term information to persist, which is a common problem of traditional RNNs. This model incorporates feedback connections, which makes understanding and predicting patterns in time-series data easier. The LSTM unit has three parts, known as *gates*: the *forget* gate, the *input* gate and the *output* gate. The first gate (forget) decides whether to keep or forget the information from the previous time step; the second gate (input) evaluates the importance of the new input information; the third gate (output) decides which information is relevant enough to go to the next state. The LSTM architecture is depicted in Fig. 3.3. More information about LSTM can be found in: [24].

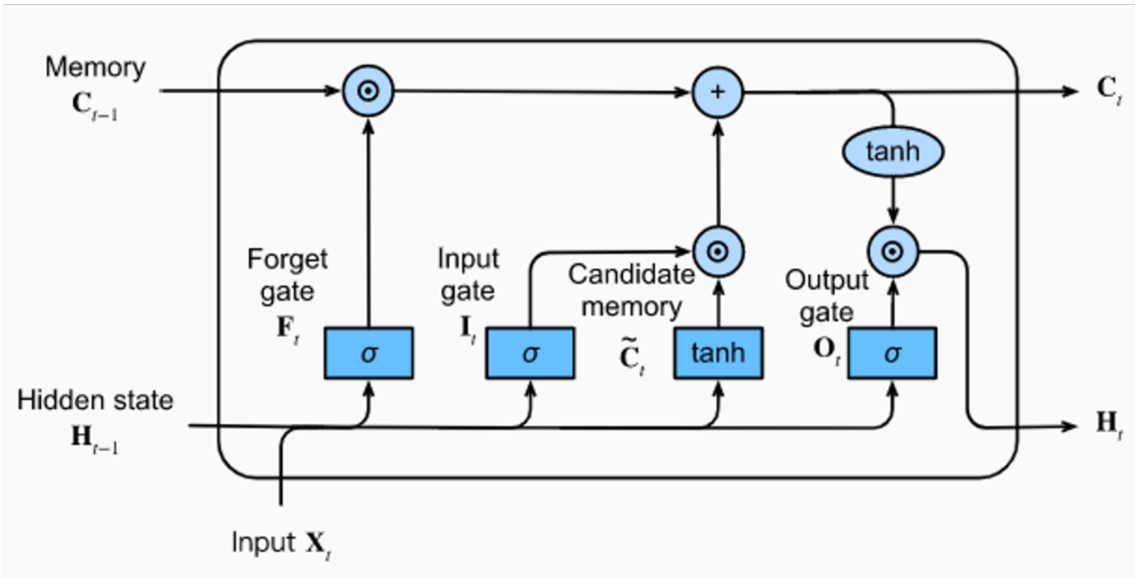


Figure 3.3: Long Short-Term Memory (LSTM) architecture. Source: LSTM

### 3.5.2 Gated Recurrent Unit (GRU)

GRU is another type of RNN with a similar architecture to LSTM, but offering an improvement. Like LSTM, GRU also controls the information flow using gates; the difference is that GRU has only two gates: the update gate and the reset gate. The first gate (update) is responsible for determining how much of the information from the previous steps needs to be forwarded to the future and the second gate (reset) decides how much information should be kept or forgotten. The GRU architecture is depicted in Fig. 3.4. More information about GRU can be found in: [25].

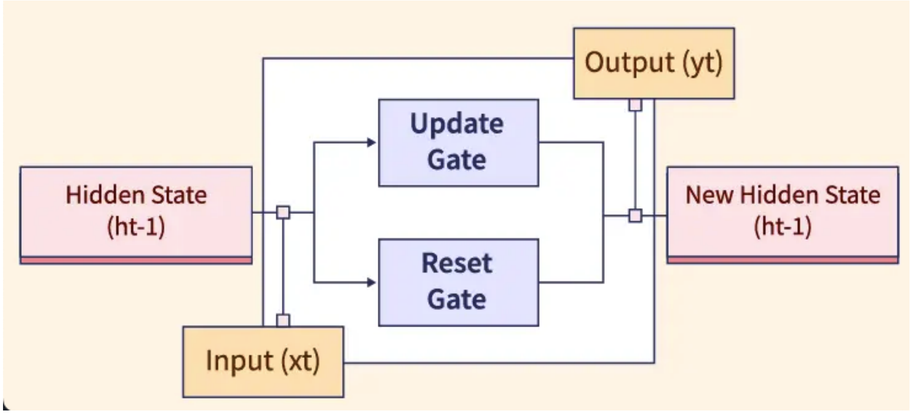


Figure 3.4: Gated Recurrent Unit (GRU) architecture. Source: GRU

# 4 Implementation

Given the methodology set in the previous Chapter, the purpose of this Chapter is to present its practical application. It starts by describing the data collection process from the sources previously defined, then it follows the preparation of the data and its visual analysis. The following stage is the creation of both the baseline and the deep learning models and, finally, there is an evaluation of the predictions delivered by those models.

## 4.1 Data Collection

As mentioned in Chapter 3, there are three sources of data: SEF, INE and Google Trends. The collection of raw data from these sources is the focus of this section.

### 4.1.1 SEF Data

The collected dataset from this organization was scattered across multiple CSV files, one for each year, from 2013 to 2019, and with a total of 6.3 GB and 4.3 Million rows. Every row of this dataset corresponds to a single tourism establishment accommodation. The features (columns) that did stand out were the establishment typology, municipality, check-in date, check-out date, age and nationality of the international tourist. Having the check-in/check-out date means that data can be extracted from this dataset with daily frequency. The schema for this collected raw data is depicted in Table 4.1.

### 4.1.2 INE Data

From this organization it was collected multiple datasets, of monthly frequency, from 1963 to 2022, and with a total of approximately 200 KB and 671 rows. Each dataset had a ’count’ of some variable by month and year related to tourist accommodation establishments: number of nights, average stay, entrance by country, number of guests, bed net rate, landed passengers, lodging earnings and total earnings. The schema for the collected raw data from INE is depicted in Table 4.2 and has a monthly frequency.

| Field | Datatype | Description |
|---|---|---|
| *rsn* | string | identifier |
| *Unidade_Hoteleira* | string | establishment identifier |
| *CAE* | string | identifier |
| *Tipologia* | string | establishment typology |
| *Classificacao* | string | accommodation classification |
| *Data_Checkin* | string | check-in date |
| *Data_Checkout* | string | check-out date |
| *Idade* | string | tourist age |
| *Concelho_UH* | string | establishment municipality |
| *Nacionalidade* | string | tourist nationality |
| *Pais_Origem* | string | tourist origin country |

Table 4.1: SEF raw data schema.

| Field | Datatype | Description |
|---|---|---|
| *Date* | string | value date |
| *TE_Total_Portugal* | string | total earnings |
| *Landed_Portugal* | string | landed passengers |
| *Total_Transport* | string | trips (transport) |
| *OS_Total_Portugal* | string | overnight stays |
| *Lod_Earn_Total_Portugal* | string | lodging earnings |
| *Avg_Stay_Total* | string | average stay |
| *BNR_Total* | string | bed net rate |
| *Guests_Total_Portugal* | string | number of guests |
| *Motive_Total* | string | trips (motive) |

Table 4.2: INE raw data schema.

### 4.1.3 Google Trends Data

The Google Trends tool was used to collect the total search volume, in Google's search engine, of specific keywords (related to Portugal's Tourism) from 2015 to 2018, and with a total of approximately 700 KB and 1826 rows. This means that each row contains the frequency of keyword searches (in the columns) on a given day.

The keywords were the following ones: *Portugal food*, *Portugal subway*, *Portugal tourist attractions*, *Portugal shopping*, *Portugal hotel booking*, *Portugal weather*, *Portugal map*, *Portugal snack*, *Portugal shopping map*, *Portugal tickets*, *Portugal hotel booking*, *Portugal accommodation*, *Portugal specialty*, *Portugal attractions*, *Portugal travel map*, *Portugal travel guide*, *Portugal travel*, *Portugal flights*, *Portugal hotels*, *Portugal airport*.

This keyword set that has been considered was inspired by [26], in which some queries had a bigger correlation with the 'Arrivals' data. The schema for the collected raw data from Google Trends is shown in Table 4.3 and has a daily frequency.

| Field | Datatype | Description |
|---|---|---|
| *date* | timestamp | description |
| *Portugal_shopping_map_norm* | double | shopping map |
| *Portugal_food_norm* | double | food |
| *Portugal_subway_norm* | double | subway |
| *Portugal_tourist_attractions_norm* | double | tourist attractions |
| *Portugal_shopping_norm* | double | shopping |
| *Portugal_travel_norm* | double | travel |
| *Portugal_tickets_norm* | double | tickets |
| *Portugal_hotel_booking_norm* | double | hotel booking |
| *Portugal_weather_norm* | double | weather |
| *Portugal_accommodation_norm* | double | accommodation |
| *Portugal_specialty_norm* | double | specialty |
| *Portugal_map_norm* | double | map |
| *Portugal_attractions_norm* | double | attractions |
| *Portugal_flights_norm* | double | flights |
| *Portugal_hotels_norm* | double | hotels |
| *Portugal_travel_map_norm* | double | travel |
| *Portugal_snack_norm* | double | snack |
| *Portugal_travel_guide_norm* | double | travel guide |
| *Portugal_airport_norm* | double | airport |

Table 4.3: Google Trends raw data schema.

## 4.2 Data Preparation

### 4.2.1 SEF Data

After getting the data, the first action taken was to merge the files into a single *.parquet* file (an "open source, column-oriented data file format designed for efficient data storage and retrieval", according to databricks.com), which would be the dataset to be used in the following stages of data pre-processing.

We noticed that this merging action produced some duplicates, not only fully identical (with the values of all the features exactly equal) but also with the same 'rsn' (the dataset 'ID' feature, supposed to be unique), all of which were removed.

Also, with a command that shows the number of null or NaN values by column, it was found that the feature 'Classificacao', which represents the rating given to a hotel by the tourist, had 1.5 million null values in a 4 million total. The final decision was to remove the feature. And this wasn't the only feature being removed – 'Unidade_Hoteleira' and 'CAE' were also removed as they were not useful for our study. Date columns like 'Checkin_Date' and 'Checkout_Date' were converted into Timestamp type. 'Age' and 'Checkin_Date' with senseless or meaningless values were removed ('Age' less than 1 or bigger than 99 and 'Checkin_Date before 2014 and after 2017).

The resulting dataset was written into a new *.parquet* file, for use in future processing steps. The resulting schema is depicted in Table 4.4.

| Field | Datatype |
|---|---|
| *Typology* | string |
| *Checkin_Date* | timestamp |
| *Checkout_Date* | timestamp |
| *Age* | double |
| *Municipality* | string |
| *Nationality* | string |
| *Origin_Country* | string |

Table 4.4: SEF processed data schema.

### 4.2.2 INE Data

In this case, the first action taken after the downloading of data was to join the different files by the 'Date' feature, which represents the month and year of the value.

Using a custom function, the feature 'Date', which had the values in written form ("Abril de 1967", for example), had its format converted into Timestamp type. The features are simply counts of some kind and strange values were dealt with using another custom function that cast Double Type to all values; if these values can't be cast into doubles, they become null.

With a command that shows the number of null or Nan values by column, some columns were dropped for the lack of enough data (less than 10 years). Also, every row with a date after 2020 was deleted because of "Covid values", which tends to produce inaccurate results (lack of consistent pattern, bad for prediction models). Every row with a date before 2009 was also deleted due to some years having lots of missing values, also inducing prediction models into bad results.

The resulting dataset was written into a new *.parquet* file, to use in future steps. The resulting schema is depicted in Table 4.5.

### 4.2.3 Google Trends Data

Here the process was similar to INE's data, joining the multiple files into a single dataset by the 'Date' feature. At first glance when looking into missing values, there was only one column with 20% null or Nan values, which wasn't that bad; after taking a deeper dive into the data it was found that nine columns had at least 50% of the values with 0. The decision was to remove all of these 9 columns from the dataset. Then a little visualization was done in order to check the patterns or the continuity of this data: the results showed that other seven columns weren't useful for the forecasting model, being removed from the dataset.

Finally, the resulting dataset was written into a new *.parquet* file, to use in future steps. The resulting schema is depicted in Table 4.6.

| **Field** | **Datatype** |
|---|---|
| *Date* | string |
| *Landed_Portugal* | double |
| *Landed_Lisboa* | double |
| *Landed_Faro* | double |
| *Landed_Porto* | double |
| *Landed_Madeira* | double |
| *Landed_Porto_Santo* | double |
| *Landed_João_Paulo_II* | double |
| *Landed_Horta* | double |
| *Landed_Santa_Maria* | double |
| *Landed_Flores* | double |
| *Landed_Graciosa* | double |
| *Landed_São_Jorge* | double |
| *Landed_Corvo* | double |
| *Landed_Pico* | double |
| *Landed_Lajes* | double |
| *Total_Transport* | double |
| *Air_Transport* | double |
| *Marine_Transport* | double |
| *Ground_Transport* | double |
| *Motive_Total* | double |
| *Motive_Leisure* | double |
| *Motive_Business* | double |
| *Motive_Family_Visit* | double |
| *Motive_Health* | double |
| *Motive_Religion* | double |
| *Motive_Others* | double |

Table 4.5: INE processed data schema.

| **Field** | **Datatype** |
|---|---|
| *date* | timestamp |
| *Portugal_travel_norm* | double |
| *Portugal_flights_norm* | double |
| *Portugal_hotels_norm* | double |
| *Portugal_airport_norm* | double |

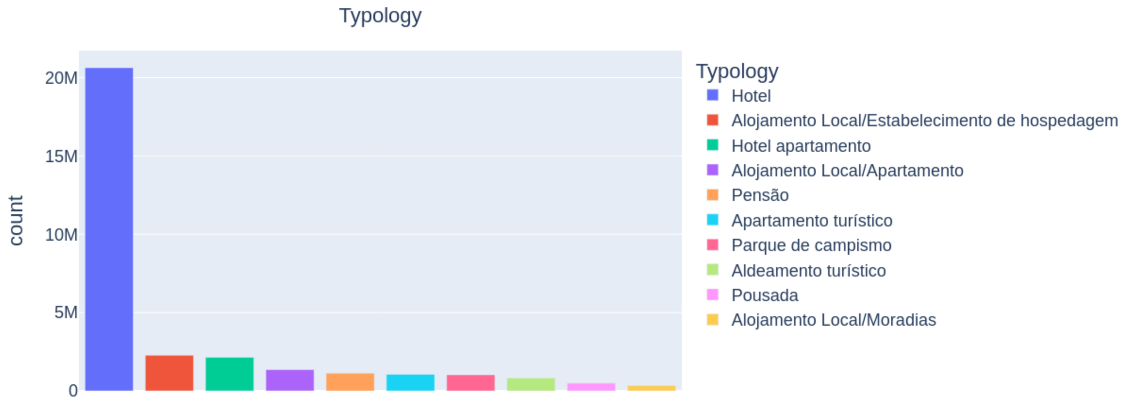
Table 4.6: Google Trends processed data schema.

## 4.3 Data Analysis and Visualization

In this section, we will try to understand the data using a variety of charts but always using the same library – Plotly.

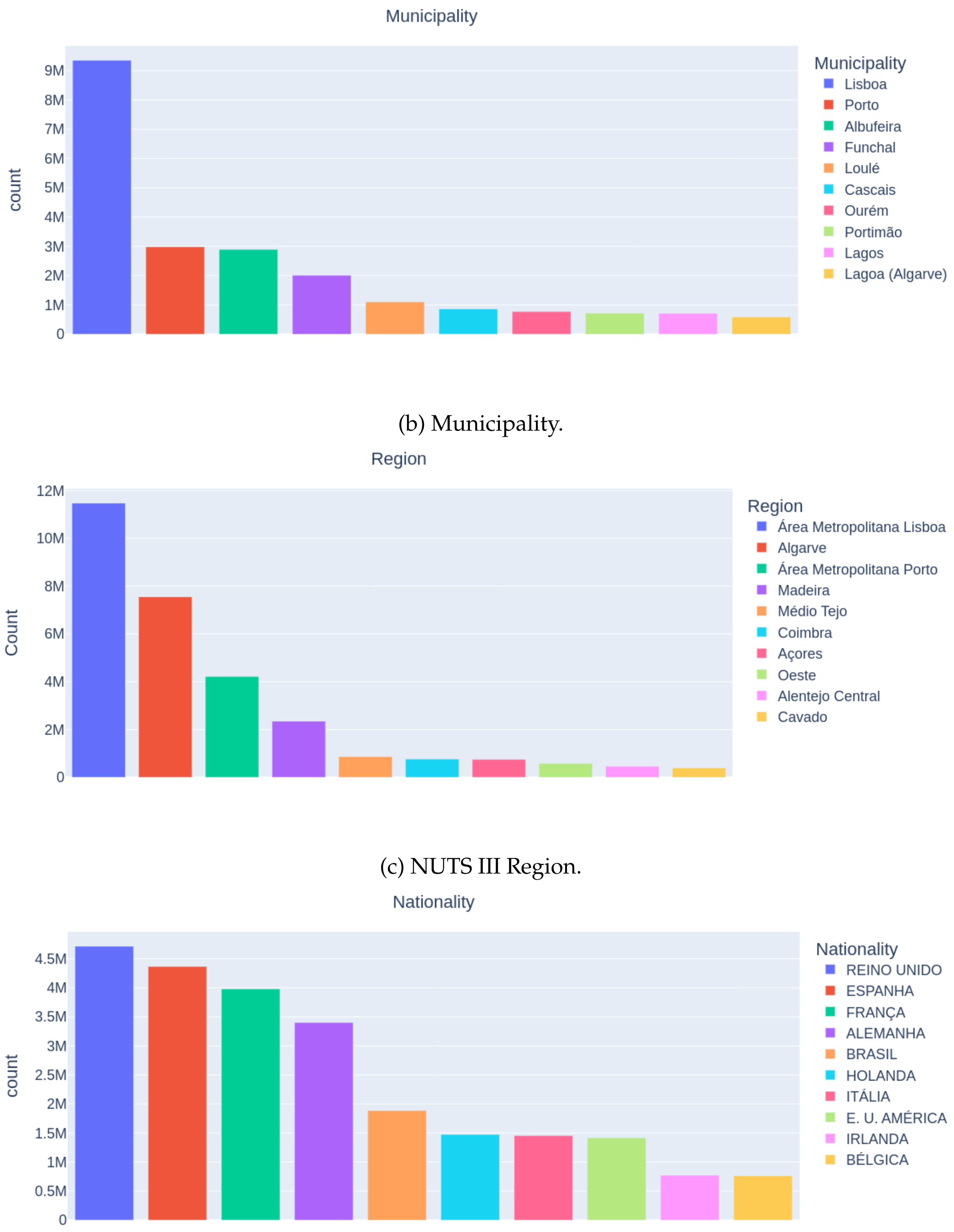
### 4.3.1 SEF Data

As can be seen in Fig. 4.1, the first kind of visualization performed on this data was to understand the values distribution of the multiple variables using a bar chart. From Fig. 4.1a it was concluded that "Hotel" is clearly the most popular type of accommodation typology, followed by some type of local accommodation ("Alojamento Local"). In Fig. 4.1b and 4.1c, it is seen that there are four big accommodations regions: Lisbon Metropolitan Area, Algarve, Porto Metropolitan Area and Madeira. Finally, in Fig. 4.1d and 4.1e, it is concluded that the tourists who most enjoy Portugal are or come from the United Kingdom, Spain, France, Germany and Brazil.



(a) Establishment typology.

Figure 4.1: SEF data. Number of accommodations by each data variable presented (typology).

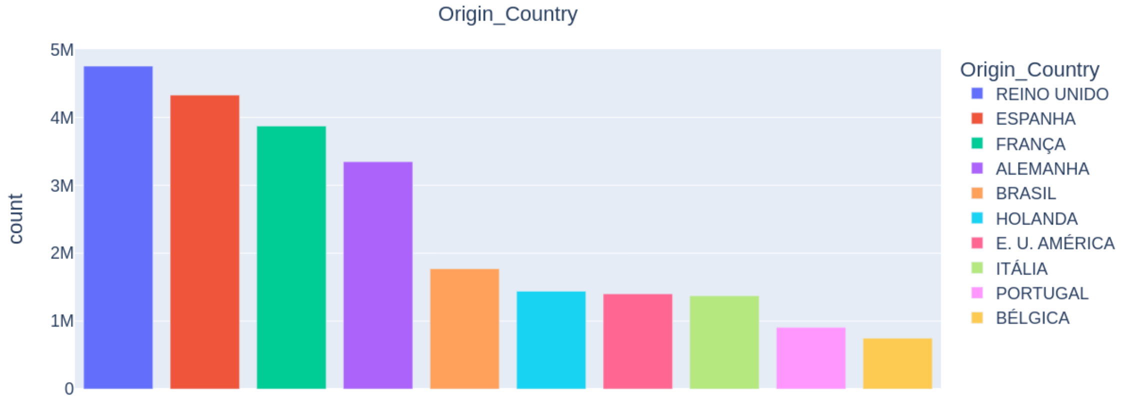


(b) Municipality.

(c) NUTS III Region.

(d) Nationality.

Figure 4.1: SEF data. Number of accommodations by each data variable presented ( municipality, NUTS III Region and nationality).



(e) Origin Country.

Figure 4.1: SEF data. Number of accommodations by each data variable presented (origin country).

In Fig. A.8, it was shown via a line chart the number of accommodations in a certain NUTS III region ( *Nomenclatura das Unidades Territoriais para Fins Estatísticos* ) over time; the regions used for this visualization were the same as mentioned above, in the bar chart. The most relevant insight we get is in Fig. 4.2, related to the number of accommodations in Algarve. It shows a clear pattern of lots of accommodations in the summer months and a calmer winter season (with a little spike on New Year's Eve).

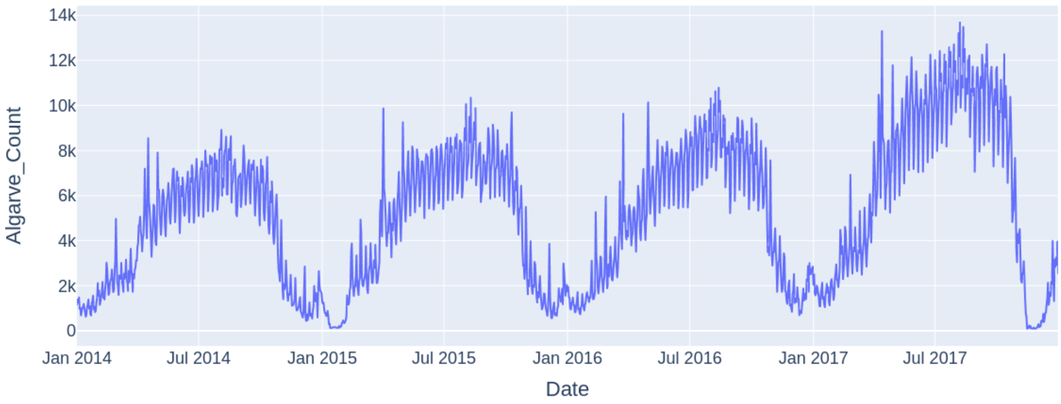


Figure 4.2: SEF data. Number of accommodations over time in Algarve.

### 4.3.2 INE Data

In Fig. A.2, it is shown line charts with passengers that landed in multiple Portuguese airports over time. As we can see, all charts are pretty similar in terms of patterns, with a steady growth over the years and a major decline in the pandemic years (2020 and 2021). As a note, data within the pandemic years is only in these charts to show what consequences COVID-19 had to each and every variable of this dataset. The most interesting chart of all is Faro (Fig. 4.3): Faro airport has a big discrepancy between the summer months and the rest of the year, due to its well-known ’beach tourism’.

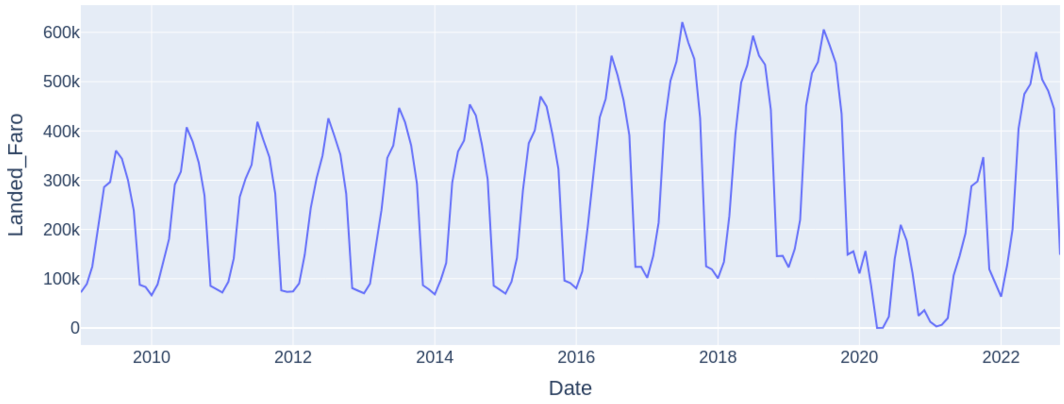
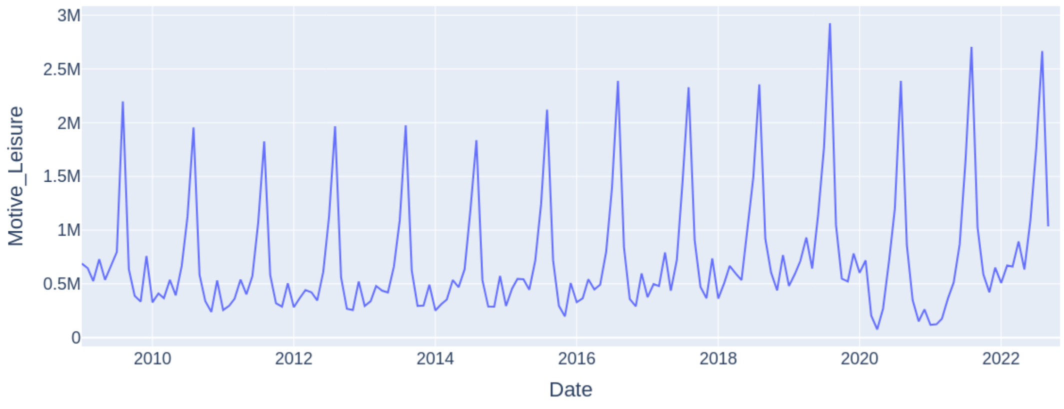
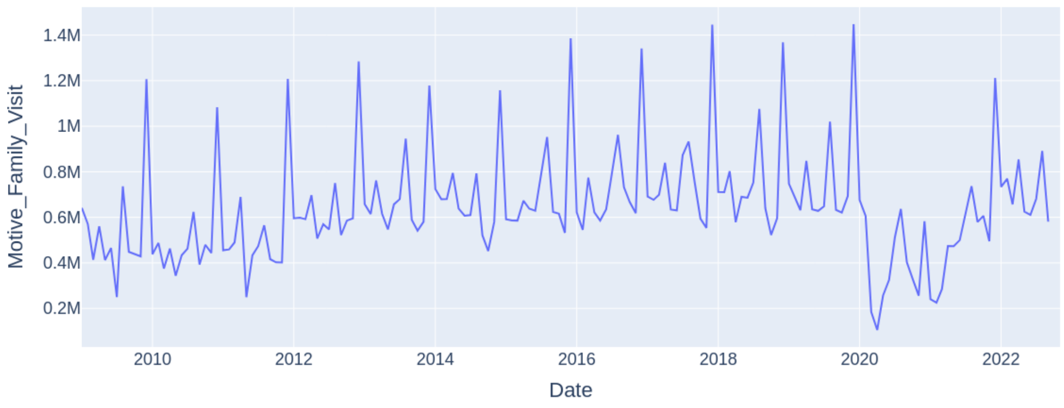


Figure 4.3: INE data. Number of passengers landed at Faro airport over time.

In Fig. A.3 there are two very interesting line charts that actually are the opposite to each other: while Fig. 4.4a have big spikes mainly in the summer months, Fig. 4.4b have the same spikes but in the winter months. This makes total sense because while in the summer months, people go on vacation (for beaches, pools and such), in the winter months (mainly in December) people get together with their families to celebrate Christmas and New Year eve.



(a) Leisure.



(b) Familiy Visit.

Figure 4.4: INE data. Number of travels for each motive over time.

In Fig. A.4 it is analysed, using line charts, the type of transport used for travel over time. What is most interesting but expected about this analysis is that air transportation (Fig. 4.5) was the most harmed during the pandemic, when compared to all the other forms of transportation.

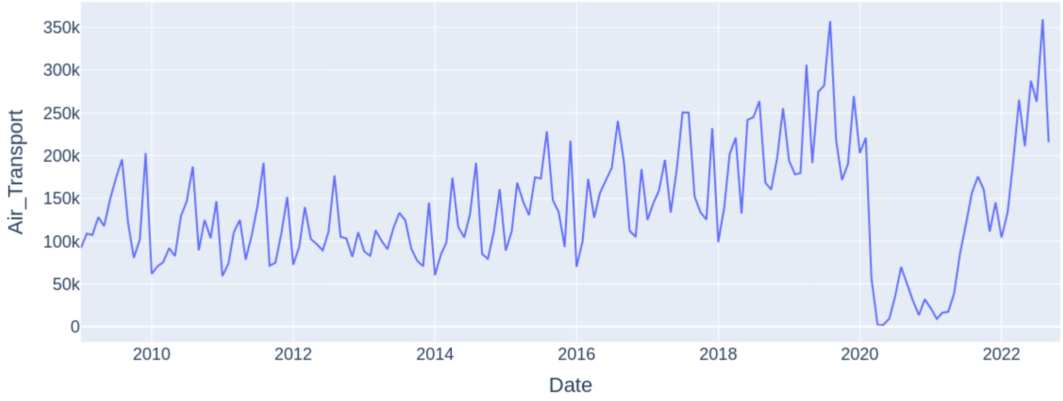


Figure 4.5: INE data. Number of travels by airplane over time.

### 4.3.3 Google Trends Data

As it can be seen in Fig. A.6, no matter what keywords are used, there is a clear pattern in the data, similar to what was seen before: high values during the summer months and a spike during New Year's Eve. There is also something interesting about Fig. 4.6: there is a constant growth over the years, which is also seen in INE's data (Fig. A.2a). This confirms that search engine keyword searches can anticipate the country's real arrivals. Therefore, it may suggest that we can have confidence in using this type of data, in combination with the SEF dataset for example, to predict the number of accommodations in a certain time frame.

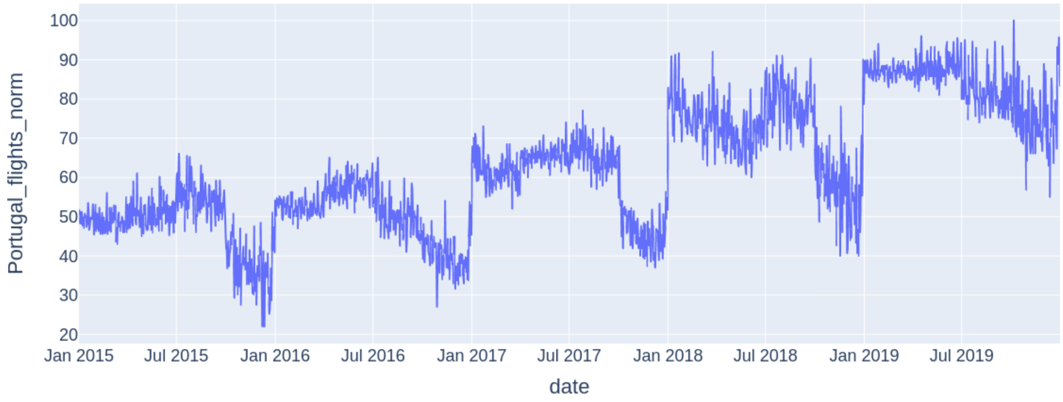


Figure 4.6: Search volume on Google's engine by *Portugal Flights* keyword over time.

### 4.3.4 Comparative Analysis

Alongside the individual analysis previously presented, it is worth to draw comparisons among those individual analyses. In that regard, when comparing Fig. A.8 and Fig. A.6, it can be seen that there are clear similarities between both patterns. It seems like the query search moment (Google Trends dataset) is followed by the accommodation (SEF dataset), which suggests that there is a high correlation between both.

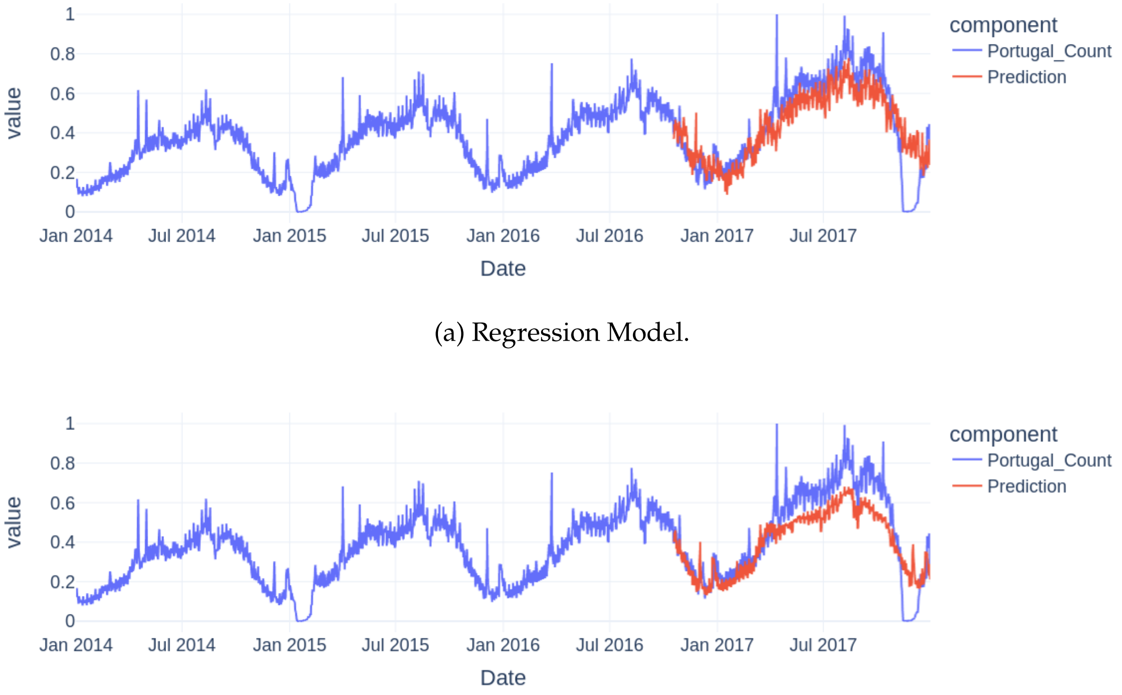
Even comparing two datasets with different data frequencies is not the easiest thing to do, we can conclude that the SEF accommodations in Lisbon (Fig. A.8a) show same pattern as the INE Lisbon arrivals (Fig. A.2b).

As data is now processed and ready-to-use, we can now create forecasting models and then making predictions.

## 4.4 Baseline Forecasting

### 4.4.1 SEF Data

To evaluate each baseline model under consideration, a derived variable was created, *Portugal_Count*, which represents the number of accommodations in Portugal over time. Indeed, by having just one variable it would simplify the process of evaluating the baseline models. Hence, having a variable that represents all the values of each variable instead of evaluating each NUTS III region or accommodation typology, for example. The predictions can be analysed in Fig. 4.7. As can be seen in Table 4.7, both regression models (Regression model and Random Forest) stand out from the rest in both metrics used for evaluation.



(a) Regression Model.

(b) Random Forest.

Figure 4.7: SEF baseline predictions by model (Regression Model, Random Forest).

| Baseline Model | MAE | MSE |
|---|---|---|
| Naive Mean | 0.218 | 0.068 |
| Naive Seasonal | 0.129 | 0.024 |
| Naive Drift | 0.197 | 0.054 |
| Naive Moving Average | 0.193 | 0.054 |
| ARIMA | 0.209 | 0.056 |
| AutoARIMA | 0.209 | 0.056 |
| Exponential Smoothing | 0.198 | 0.059 |
| Regression Model | 0.094 | 0.016 |
| Random Forest | 0.097 | 0.015 |

Table 4.7: SEF Baseline Forecasting models evaluated by metrics results.

### 4.4.2 INE Data

To evaluate each baseline model, *Landed_Portugal* ended up being chosen as a target variable because it is the most general variable of this dataset (and used in various articles as the main/target variable – [4], [7], [8], [12], for example). The predictions can be analysed in Fig. A.1. As can be seen in Table4.8 AutoARIMA stands out from the rest in both metrics. We can also see that all statistical models perform really well compared to the others.

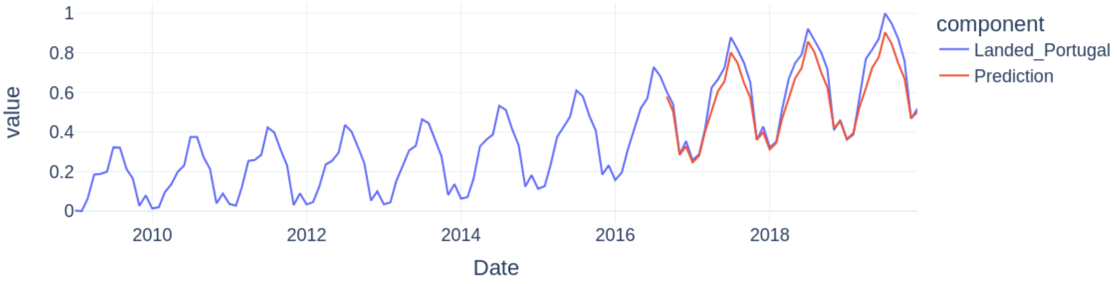


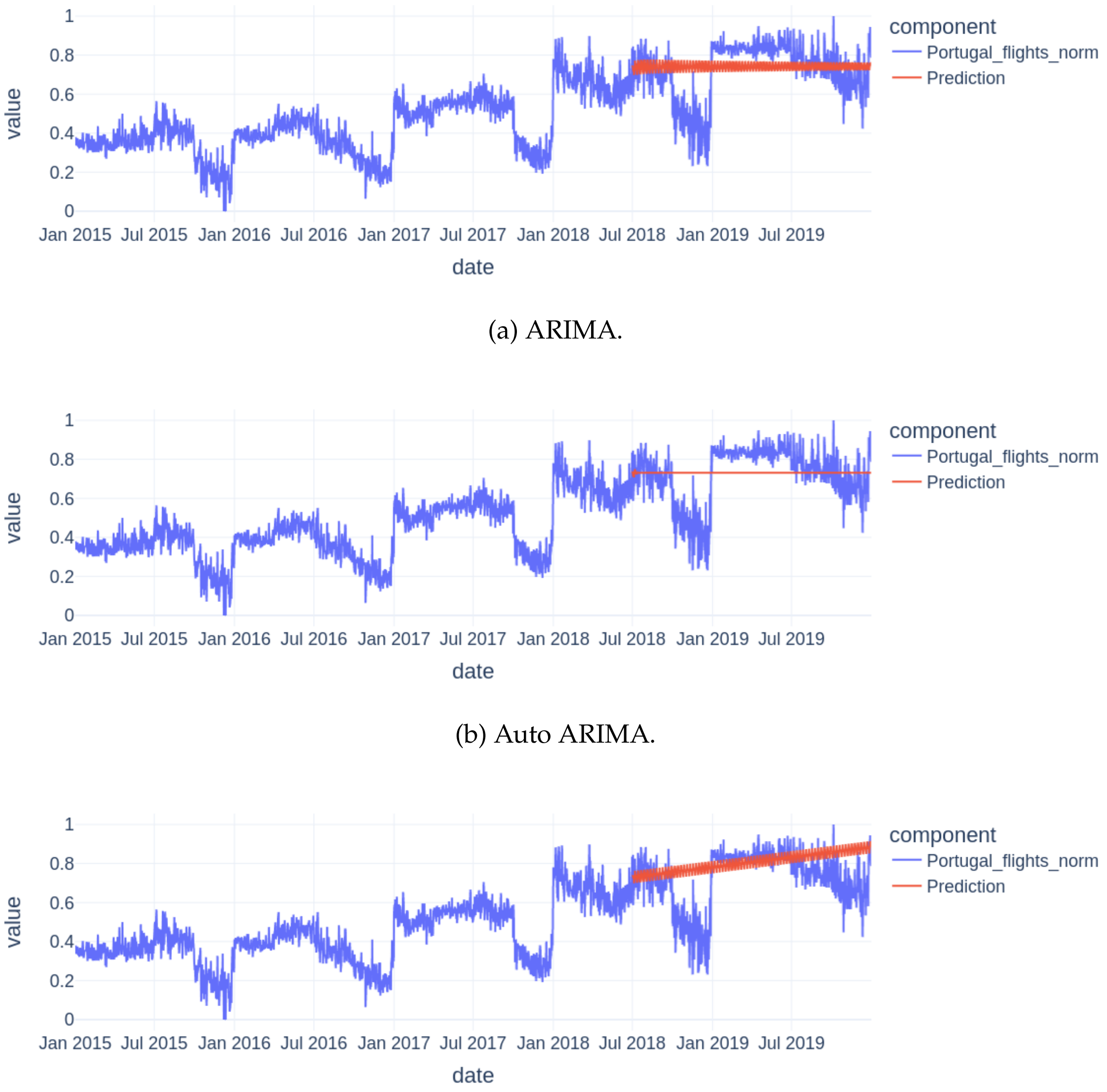
Figure 4.8: INE baseline prediction with the AutoARIMA model.

| Baseline Model | MAE | MSE |
|---|---|---|
| Naive Mean | 0.364 | 0.178 |
| Naive Seasonal | 0.214 | 0.052 |
| Naive Drift | 0.236 | 0.088 |
| Naive Moving Average | 0.212 | 0.064 |
| ARIMA | 0.089 | 0.011 |
| AutoARIMA | 0.054 | 0.005 |
| Exponential Smoothing | 0.082 | 0.011 |
| Regression Model | 0.089 | 0.014 |
| Random Forest | 0.113 | 0.017 |

Table 4.8: INE baseline forecasting models evaluated by metrics results.

### 4.4.3 Google Trends Data

To evaluate each baseline model, *Portugal_flights_norm* ended up being used as the target variable because, being ’arrivals’ the most used variable to make predictions in general in this field, the search of those flights might have a similar pattern. The predictions can be analysed in Fig. 4.9. As it can be seen in Table4.9, all statistical models perform better than the others, but not with a big difference.



(a) ARIMA.

(b) Auto ARIMA.

(c) Exponential Smoothing.

Figure 4.9: Google Trends baseline predictions by model (ARIMA, Auto ARIMA and Exponential Smoothing).

| Baseline Model | MAE | MSE |
|---|---|---|
| Naive Mean | 0.301 | 0.107 |
| Naive Seasonal | 0.212 | 0.058 |
| Naive Drift | 0.215 | 0.068 |
| Naive Moving Average | 0.187 | 0.043 |
| ARIMA | 0.113 | 0.022 |
| AutoARIMA | 0.117 | 0.023 |
| Exponential Smoothing | 0.120 | 0.028 |
| Regression Model | 0.137 | 0.033 |
| Random Forest | 0.126 | 0.023 |

Table 4.9: Google Trends baseline forecasting models evaluated by metrics results.

### 4.4.4 Comparative Analysis

We can conclude that, supported by the figures and tables from this section, there is a big difference in results when comparing monthly (INE) with daily (SEF and Google Trends) data. This is because things like the time of the year, holidays or other intricacies from daily data really affect the performance of the models when making predictions. Also, based on the analysis exposed in the tables in this section, we can conclude that the NAIVE models were the worst performers in all kinds of data, making their use questionable even when doing baseline forecasting.

## 4.5 Deep Learning Forecasting

In this section, we show the results of building and training the deep learning forecasting models for each data source chosen. Also, we analyse the predictions we get, and comparing them with the baseline forecasting predictions of the previous section. Finally, we discuss the use of these models to achieve the goals defined in Chapter 1.

### 4.5.1 SEF Model

The derived variable *Portugal_Count* was used to make predictions but, this time, using deep learning models. In this case, *Portugal_Count* was used not only as the target variable but also as the only feature to make these predictions. The outcome of predictions can be seen in both Fig. 4.10 and Fig. 4.11. As can be seen in Table 4.10, and when comparing the two better performers from the baseline forecasting (Regression and Random Forest model), we can conclude that both deep learning algorithms provide us much better predictions.

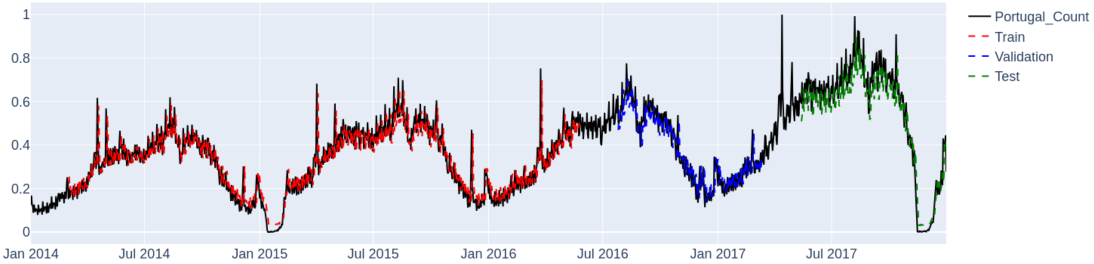


Figure 4.10: SEF LSTM prediction.

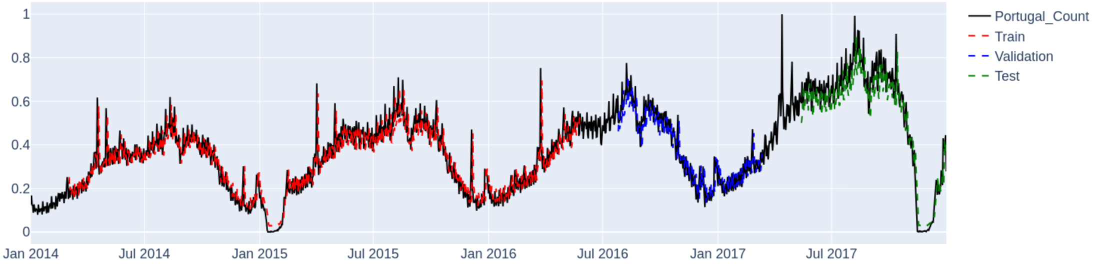


Figure 4.11: SEF GRU prediction.

| Baseline Model | MAE | MSE |
|---|---|---|
| LSTM | 0.067 | 0.007 |
| GRU | 0.065 | 0.007 |
| Regression Model | 0.094 | 0.016 |
| Random Forest | 0.097 | 0.015 |

Table 4.10: SEF deep learning vs. baseline forecasting results by metric.

### 4.5.2 INE Model

The same derived variable that was created before, *Landed_Portugal* was used to make predictions using deep learning models. *Landed_Portugal* was used both as a target and as only feature variable. The predictions can be seen in both Fig. 4.12 and Fig. 4.13. As can be seen in Table 4.11, both deep learning algorithms performed much worse when compared with Auto ARIMA, the best performer of the baseline models.

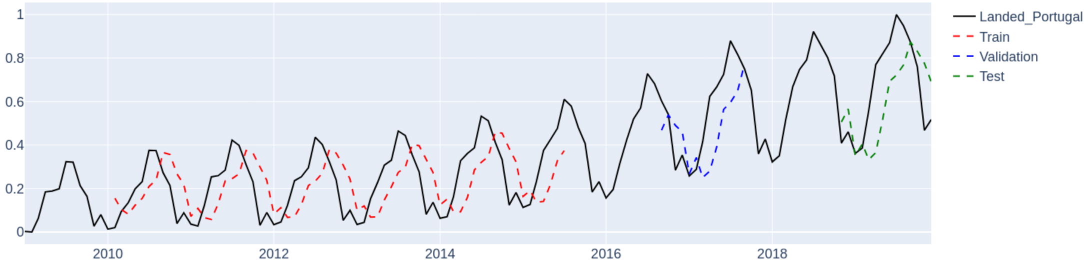


Figure 4.12: INE LSTM prediction.

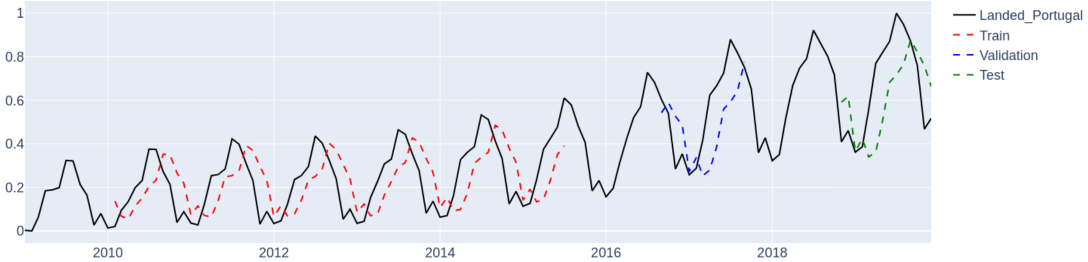


Figure 4.13: INE GRU prediction.

| Baseline Model | MAE | MSE |
|---|---|---|
| LSTM | 0.1490 | 0.0303 |
| GRU | 0.1283 | 0.0229 |
| Auto ARIMA | 0.054 | 0.005 |

Table 4.11: INE deep learning vs. baseline forecasting results by metric.

### 4.5.3 SEF + Google Trends Model

The same variable used to make deep learning predictions on SEF data, *Portugal_Count*, is used here as a feature and as the target variable. The difference is that along with *Portugal_Count* (renamed as *Portugal_stays_count*), 4 other variables from Google Trends Data were used as features: *Portugal_travel_norm*, *Portugal_flights_norm*, *Portugal_hotels_norm* and *Portugal_airport_norm*. The predictions can be seen in both Fig. 4.14 and Fig. 4.15. As can be seen in Table 4.12, the results using a multivariate dataset to make predictions are very similar to using only the same variable as a feature and as the target (4.10), but still a big improvement over baseline models.

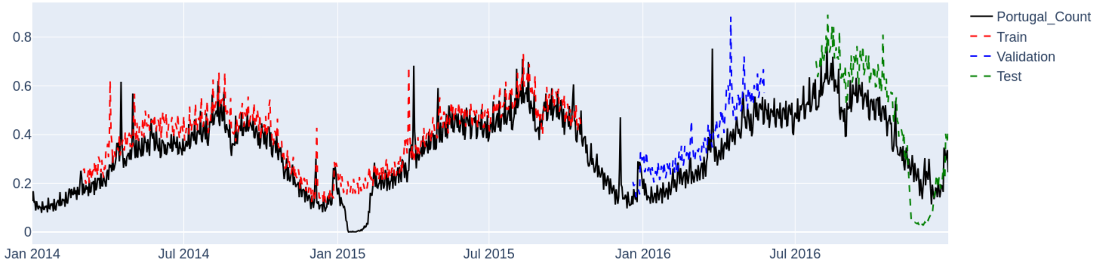


Figure 4.14: SEF + Google Trends LSTM predictions.

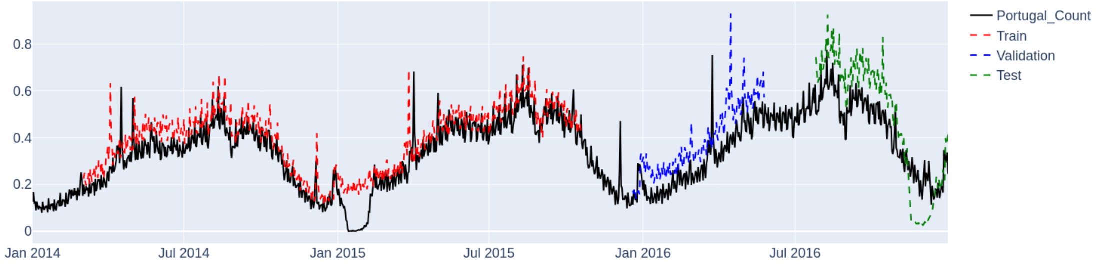


Figure 4.15: SEF + Google Trends GRU predictions.

| Baseline Model | MAE | MSE |
|---|---|---|
| LSTM | 0.068 | 0.007 |
| GRU | 0.066 | 0.007 |
| ARIMA | 0.113 | 0.022 |
| AutoARIMA | 0.117 | 0.023 |
| Exponential Smoothing | 0.120 | 0.028 |

Table 4.12: SEF + Google Trends deep learning vs. baseline forecasting results by metric.

## 4.6 Models Evaluation

The purpose of this section is to evaluate and analyse the results from both baseline and deep learning models, and conclude whether we met the research goals set in Chapter 1 or not.

Regarding SEF data, both deep learning models outperformed the highest performers of the baseline models (regression model and random forest). The same happened with SEF + Google Trends data, with both deep learning models outperforming the baseline statistical models (ARIMA, AutoARIMA and Exponential Smoothing models). In the case of INE's data, the deep learning models did not perform better than the baseline models, being almost two times worse than the latter. In this case, the advantage of using deep learning models with the datasets we have is when using daily frequency data.

## 4.7 Forecasting

Although deep learning forecasting models produce good results when we are predicting short-term results, we cannot say the same for the medium-term or long-term horizon. To do this, it's better to use statistical models because, although the short-term accuracy is not the best, the long-term horizon is better than the one from deep learning models. In this case, we used ARIMA to predict the arrivals to Portugal the following year (we have data until the end of 2019 and we tried to predict 2020s arrivals). The results are shown in Fig. 4.16. This forecast has not only the prediction for that year but also a cone of uncertainty, inside which the results can vary.

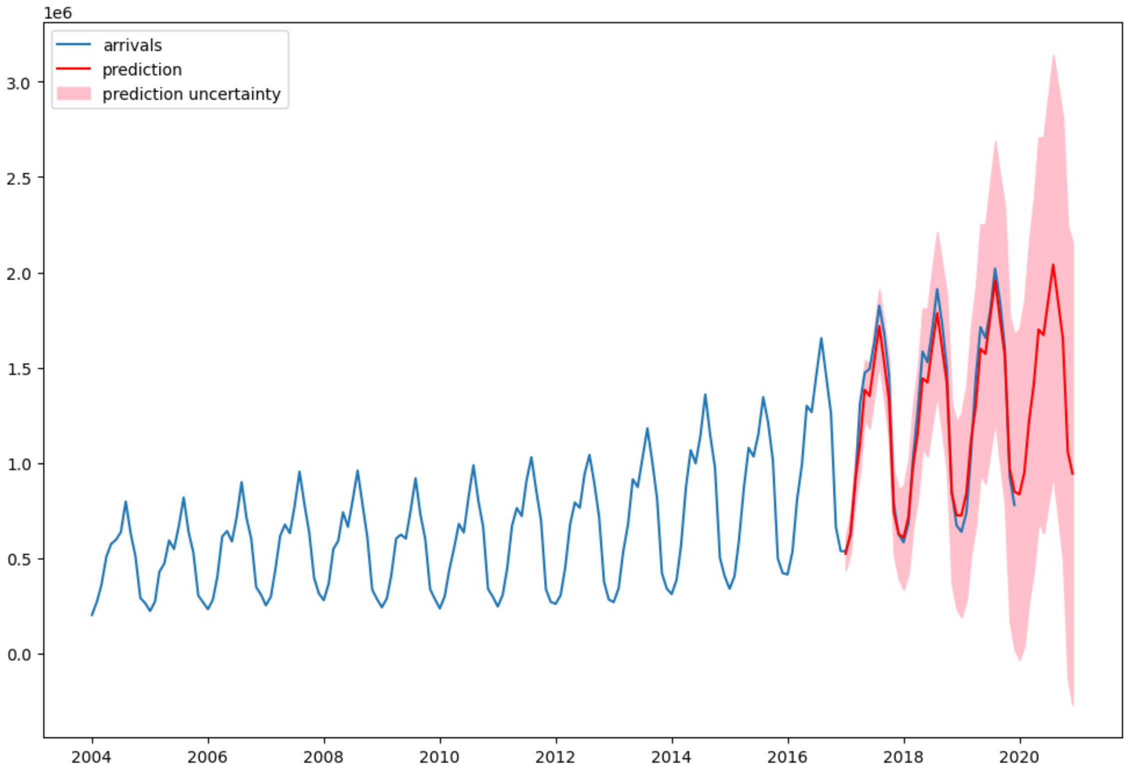


Figure 4.16: ARIMA arrivals forecast (2020).

As a note, the process of forecasting short-term results might also be called as nowcasting. More precisely, nowcasting is the prediction of the very near future, this is, in the next time step. For example, if the time series has a monthly frequency we will try to predict the next month; if the time series has a daily frequency we will try to predict the next day. Furthermore, nowcasting is also possible due to the constant flow of information and the continuous update of the prediction model with new data. And here is where our deep learning models can shine.

## 4.8 Summary

We have presented the application of the methodology set in Chapter 3. That is, we collected raw data from SEF, INE and Google Trends sources, processed it, analysed it and then built some models with it. We have seen that both baseline and deep learning models had good performance with the data we have used, and also that, mainly in the case of deep learning models, they perform really well with daily data.

As a final note, recall that a goal of this research study was also to contribute to helping SMEs to recover from the economic situation after COVID-19. Actually, in the context of the RESETTING project, the predictions from these models that have been built will accomplish what is expected, either by using medium-term forecasting or a more short-term nowcasting.

# 5

# Web Application for Deployment

## 5.1 Introduction

As expected and shown in Fig. 3.1, the results of forecasting models should be available to users in general. And as mentioned before, the RESETTING project is guided towards helping SMEs recover from the economic downfall of the pandemic. The RESETTING project was also keen on creating conditions that would allow SMEs to use various types of information to their advantage, including predictions on tourism flows.

Hence, given the forecasting models that have been created, and so the results they can generate, it was decided that the best way forward to make them available to SMEs was via a web-based application, where the predictions could be visualised and explored. That is, the model predictions were placed in a web application, where businesses and authorized people might use them as appropriate, in particular in order to adjust their services to the predictions of future tourism flow.

In the following we will present details of the web application, hereafter called *Resetlytics*. We should point out from the beginning that, as there are two main components in the web application – front-end and back-end – the work of implementation has been carried by a team of two people, myself and my colleague André Garcia. Hence, the front-end is described here in more detail as it was my own responsibility, whereas implementation work in the back-end and datatase side was André's responsibility.

## 5.2 Architecture

In order to implement *Resetlytics*, we have designed a three-tier architecture to support it, which is one of the most popular architectures when developing this type of applications. Among other aspects that have been considered, the application: (i) will be more secure since the client side does not access the data directly; (ii) it leads to higher scalability, as it can be deployed in multiple machines and scale each tier independently; (iii) delivers better performance; and (iv) code is re-usable, since each layer is connected to each other through an interface and, for instance, if we change the implementation of one, there is no automatic need to change the implementation of the others. In Fig. 5.1 we can see a diagram of the three-tier architecture.

The three tiers of the supporting architecture are as follows:

- **Presentation/Client:** the layer through which the user interacts with the application.
- **Application/Business:** the business logic layer that uses certain protocols to listen to client requests via a browser, process them and deliver the response back to the client.
- **Data:** the database layer, which stores and manages information from the application.

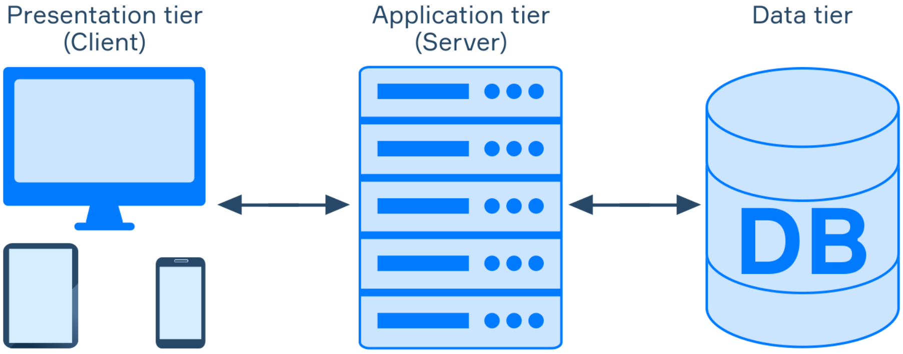


Figure 5.1: Three-tier architecture supporting the *Resetlytics* web application.

## 5.3 Implementation

Given the architecture above established, the implementation of *Resetlytics* web application has relied primarily on the following frameworks:

- The presentation layer (front-end) runs on a browser and was built using NextJS, which is a React framework.
- The application layer has been built using the Django framework.
- The data layer is supported by a MySQL database. Notice that we are working with structured data, that is why we elected a SQL-based database.

Prior to any details about modules and functionalities, notice that since *Resetlytics* is also for authorized entities, as defined in the RESETTING project, and to whom were given access credentials directly, there was no need to create a *Register* page. However, there was a need for a *Login* page, as depicted in Fig. 5.2.

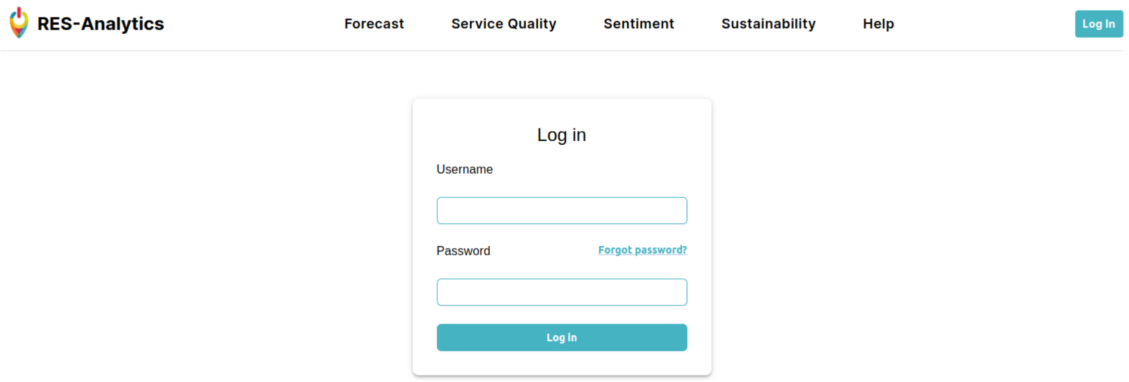


Figure 5.2: *Resetlytics* login page.

The module we are mostly interested on relates to the *Forecast* page. Wee have opted for a simplistic view of visualizing the model predictions, by using one of the most popular JavaScript charting libraries – ChartJS. Just to give a glimpse, in Fig. 5.3 we can see that the actual values from the time series are in black, while the predictions from the model are in green. Not only that but those predictions have a cone of uncertainty, which represents the potential error in the forecast.

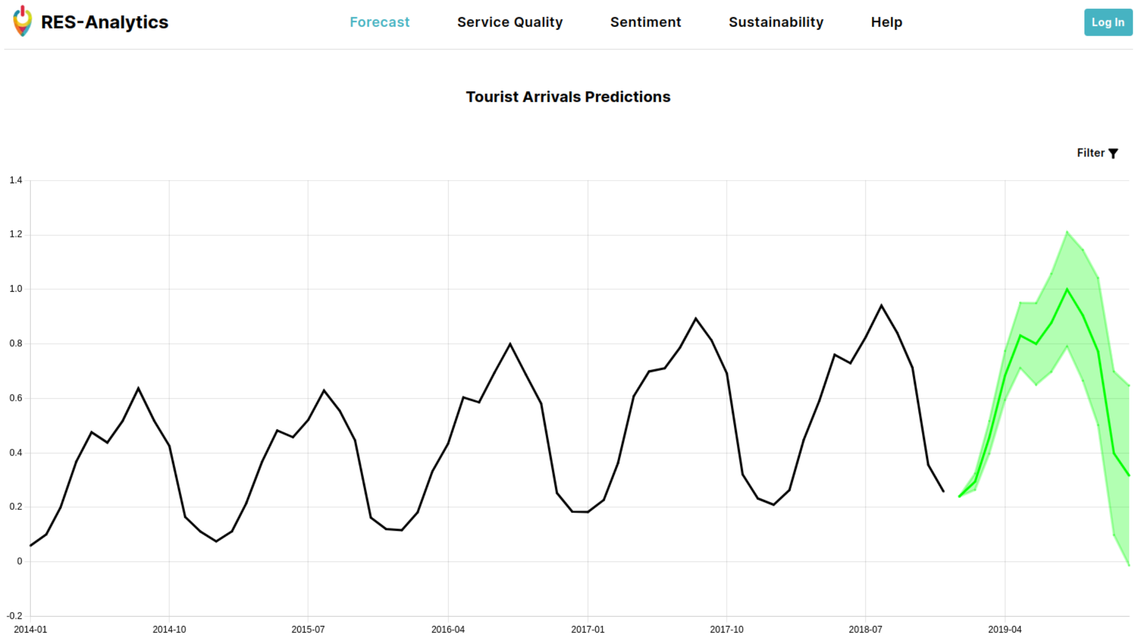
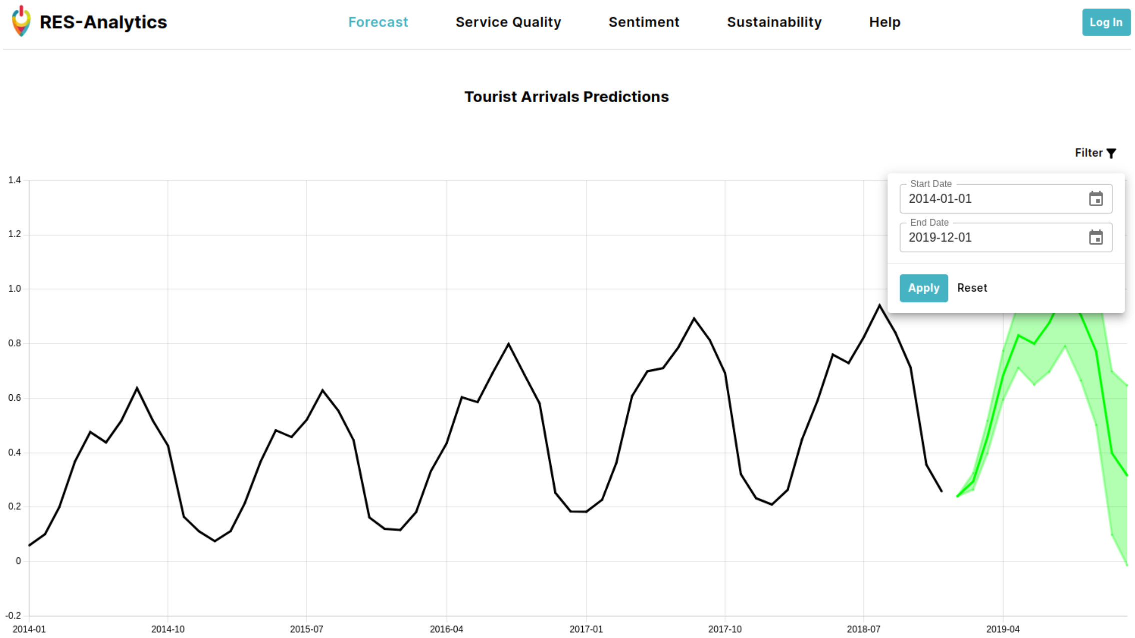


Figure 5.3: *Resetlytics* forecast page.

In addition, we also have a *Filter* button on the right-hand side of the page that enables the user to pick the time interval he pretends to visualize. We can see the structure of the filter in Fig. 5.4.



Start Date 2014-01-01

End Date 2019-12-01

Figure 5.4: *Resetlytics* forecast page with a filter to be applied in the time feature.

Finally, besides the forecasting module that has been incorporated in *Resetlytics*, there are other modules that are also part of *Resetlytics*. They are out of scope of this research study but still, it is important to mention them here. As can be seen in the navigation bar in Fig. 5.3, there are also references to:

- *service quality*, containing information on whether the SME meet the needs of customers;
- *sentiment analysis*, containing charts about the overall sentiment of the customers regarding products and services of particular SME;
- *sustainability*, which relates to information about SME's sustainability.

These are modules that have already been implemented, despite being in the early stages of development.

# 6 Conclusions and Future Work

In this Chapter, we will present an overview of major contributions of this research study to the field, as well as its limitations. Furthermore, we will discuss what could be done in the future to enhance this research work.

## 6.1 Contributions

In Chapter 1, we have introduced the motivation and scope of this dissertation, exposing how relevant the tourism market is to the world and, more specifically, to the Portuguese economy. Finally, we asked ourselves how could we make a contribution to help the Portuguese SMEs to recover economically from the COVID-19 pandemic, but with focus on tourism forecasting models.

After a proper literature review of the research field has been established, the process and the course of action was planned and then implemented, in accordance to what has been described in Chapters 3 and 4. We started by defining the working pipeline to be followed in order to build the models, collecting the data from the chosen sources (SEF, INE and Google Trends), then to process the data so it could analysed and, afterwards, to use the clean data for creating both the baseline and the deep learning forecasting models. We concluded that, after evaluating the results from the models, deep learning models worked best with daily frequency data, whereas statistical/baseline models worked best with monthly frequency data.

As for providing results to users, in Chapter 5 it was shown how predictions can be visualized, via the *Resetlytics* web application that has been designed and implemented – its architecture and functionalities were conveyed.

## 6.2 Limitations and Future Work

There are some limitations in the work presented in this dissertation. The first one is basically data-related: the amount of data from SEF could be increased so that the algorithms could pick up the patterns more easily.

The second limitation is more related to performance, that is: in order to optimize the deep learning models' hyperparameters, we could have used bayesian optimization because there is a high probability that the prediction results might have been improved.

The third limitation is related to the *Resetlytics* web application: we could have improved the visualization of the forecasting charts (maybe with other types of data visualization) and also have added a higher degree of complexity to the filter. All these limitations mentioned can be improved/added in future work. And, of course, additional functionalities to enhance the usability and added-value of the application.

# Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] C. Schröer, F. Kruse, and J. M. Gómez. "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 526–534. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.01.199. URL: https://www.sciencedirect.com/science/article/pii/S1877050921002416 (cit. on p. 3).

[3] G. Athanasopoulos et al. "The tourism forecasting competition". In: *International Journal of Forecasting* 27 (3 2011-07), pp. 822–844. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2010.04.009 (cit. on pp. 6, 12, 18).

[4] D. Gounopoulos, D. Petmezas, and D. Santamaria. "Forecasting Tourist Arrivals in Greece and the Impact of Macroeconomic Shocks from the Countries of Tourists' Origin". In: *Annals of Tourism Research* 39 (2 2012-04), pp. 641–666. ISSN: 01607383. DOI: 10.1016/j.annals.2011.09.001 (cit. on pp. 6, 12, 18, 33).

[5] M. E. Nor, A. I. Nurul, and M. S. Rusiman. "A Hybrid Approach on Tourism Demand Forecasting". In: *Journal of Physics: Conference Series* 995 (1 2018-04). ISSN: 17426596. DOI: 10.1088/1742-6596/995/1/012034 (cit. on pp. 6, 12).

[6] R. Fildes, Y. Wei, and S. Ismail. "Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures". In: *International Journal of Forecasting* 27 (3 2011-07), pp. 902–922. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2009.06.002 (cit. on pp. 7, 12).

[7] W. R. Moore. "The impact of climate change on Caribbean tourism demand". In: *Current Issues in Tourism* 13 (5 2010), pp. 495–505. ISSN: 13683500. DOI: 10.1080/13683500903576045 (cit. on pp. 7, 12, 18, 33).

[8] K. K. Wong, H. Song, and K. S. Chon. "Bayesian models for tourism demand forecasting". In: *Tourism Management* 27 (5 2006), pp. 773–780. ISSN: 02615177. DOI: 10.1016/j.tourman.2005.05.017 (cit. on pp. 7, 12, 33).

[9] H. Song, S. F. Witt, and T. C. Jensen. "Tourism forecasting: accuracy of alternative econometric models". In: *International Journal of Forecasting* 19.1 (2003), pp. 123–141. ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(01)00134-0. URL: https://www.sciencedirect.com/science/article/pii/S0169207001001340 (cit. on pp. 7, 12).

[10] R. Law et al. "Tourism demand forecasting: A deep learning approach". In: *Annals of Tourism Research* 75 (2019-03), pp. 410–423. ISSN: 01607383. DOI: 10.1016/j.annals.2019.01.014 (cit. on pp. 7, 8).

[11] E. X. Jiao and J. L. Chen. "Tourism forecasting: A review of methodological developments over the last decade". In: *Tourism Economics* 25 (3 2019-05), pp. 469–492. ISSN: 13548166. DOI: 10.1177/1354816618812588 (cit. on p. 8).

[12] C. F. Chen, M. C. Lai, and C. C. Yeh. "Forecasting tourism demand based on empirical mode decomposition and neural network". In: *Knowledge-Based Systems* 26 (2012-02), pp. 281–287. ISSN: 09507051. DOI: 10.1016/j.knosys.2011.09.002 (cit. on pp. 8, 12, 18, 33).

[13] S. Li et al. "Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index". In: *Tourism Management* 68 (2018-10), pp. 116–126. ISSN: 02615177. DOI: 10.1016/j.tourman.2018.03.006 (cit. on pp. 9, 13).

[14] S. Cang. "A non-linear tourism demand forecast combination model". In: *Tourism Economics* 17 (1 2011-02), pp. 5–20. ISSN: 13548166. DOI: 10.5367/te.2011.0031 (cit. on pp. 9, 13).

[15] O. Claveria, E. Monte, and S. Torra. "Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information". In: *International Journal of Tourism Research* 17 (5 2015-09), pp. 492–500. ISSN: 15221970. DOI: 10.1002/jtr.2016 (cit. on pp. 9, 13).

[16] D. M. Ã§uhadar, I. Cogurcu, and C. Kukrer. "Modelling and Forecasting Cruise Tourism Demand to Izmir by Different Artificial Neural Network Architectures". In: *International Journal of Business and Social Research* 4.3 (2014-03), pp. 12–28. URL: https://ideas.repec.org/a/mir/mirbus/v4y2014i3p12-28.html (cit. on pp. 9, 13).

[17] V. Cho. *A comparison of three different approaches to tourist arrival forecasting*. 2003, pp. 323–330 (cit. on pp. 10, 13).

[18] Y. Li and H. Cao. “Prediction for Tourism Flow based on LSTM Neural Network”. In: *Procedia Computer Science* 129 (2018). 2017 INTERNATIONAL CONFERENCE ON IDENTIFICATION,INFORMATION AND KNOWLEDGEIN THE INTERNET OF THINGS, pp. 277–283. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.03.076. URL: https://www.sciencedirect.com/science/article/pii/S1877050918303016 (cit. on pp. 10, 13).

[19] W. C. Hong et al. “SVR with hybrid chaotic genetic algorithms for tourism demand forecasting”. In: vol. 11. 2011-03, pp. 1881–1890. DOI: 10.1016/j.asoc.2010.06.003 (cit. on pp. 10, 14).

[20] C. Goh, R. Law, and H. M. Mok. “Analyzing and forecasting tourism demand: A rough sets approach”. In: *Journal of Travel Research* 46 (3 2008), pp. 327–338. ISSN: 15526763. DOI: 10.1177/0047287506304047 (cit. on pp. 11, 14).

[21] C. H. Wang. “Predicting tourism demand using fuzzy time series and hybrid grey theory”. In: *Tourism Management* 25 (3 2004), pp. 367–374. ISSN: 02615177. DOI: 10.1016/S0261-5177(03)00132-8 (cit. on pp. 11, 14).

[22] B. Shahriari et al. “Taking the Human Out of the Loop: A Review of Bayesian Optimization”. In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175. DOI: 10.1109/JPROC.2015.2494218 (cit. on p. 11).

[23] J. Manu. *MODERN TIME SERIES FORECASTING WITH PYTHON explore industry-ready time series forecasting using modern machine learning and deep learning*. PACKT PUBLISHING LIMITED, 2022. ISBN: 9781803246802 (cit. on p. 17).

[24] B. Lindemann et al. “A survey on long short-term memory networks for time series prediction”. In: *Procedia CIRP* 99 (2021). 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020, pp. 650–655. ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2021.03.088. URL: https://www.sciencedirect.com/science/article/pii/S2212827121003796 (cit. on p. 19).

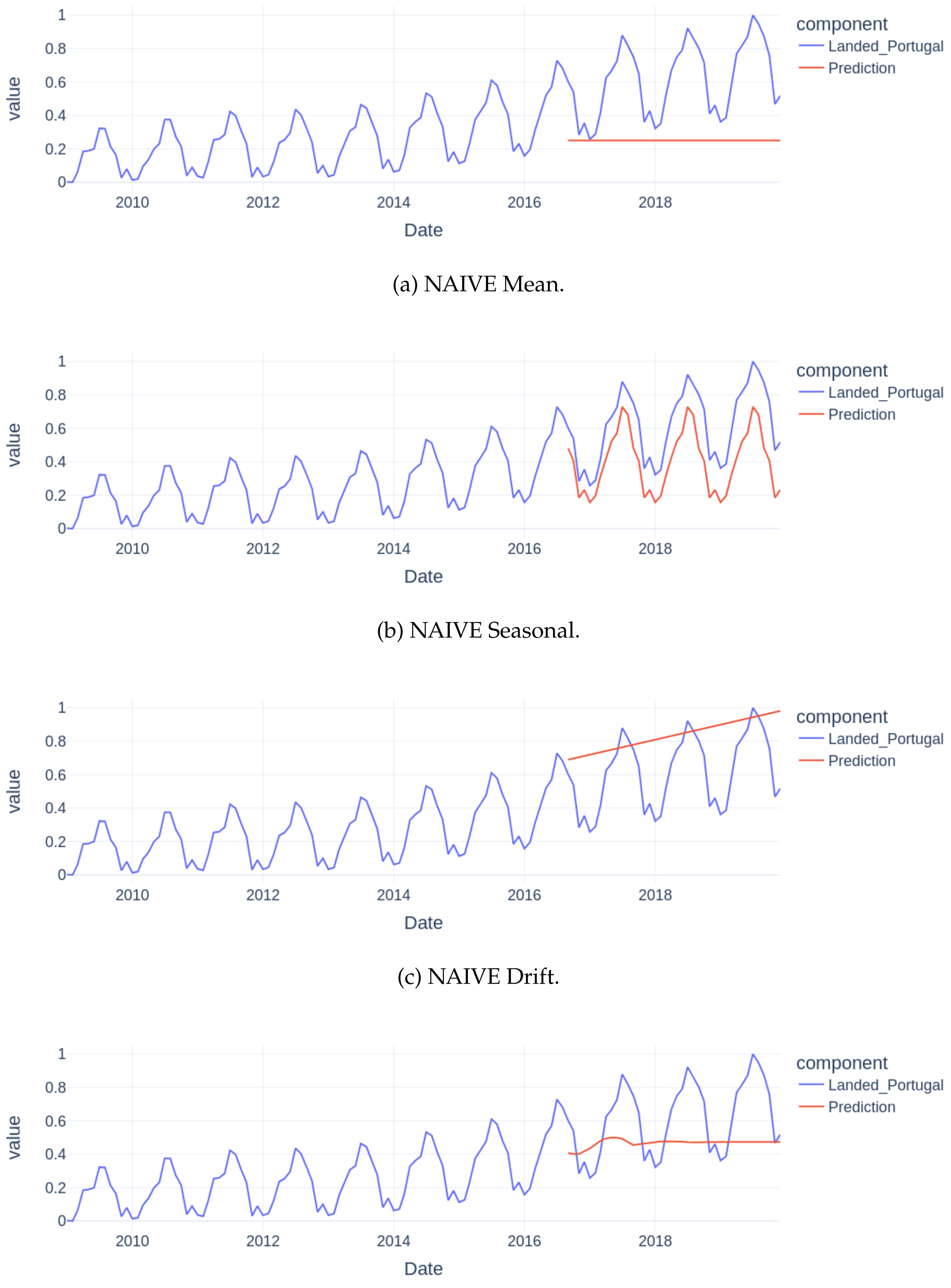
[25] R. Dey and F. M. Salem. “Gate-variants of Gated Recurrent Unit (GRU) neural networks”. In: (2017), pp. 1597–1600. DOI: 10.1109/MWSCAS.2017.8053243 (cit. on p. 20).

[26] L. Wen, C. Liu, and H. Song. “Forecasting tourism demand using search query data: A hybrid modelling approach”. In: *Tourism Economics* 25 (3 2019-05), pp. 309–329. ISSN: 13548166. DOI: 10.1177/1354816618768317 (cit. on p. 22).

# A

# Appendix 1

## A.1 INE Baseline Prediction Visualization



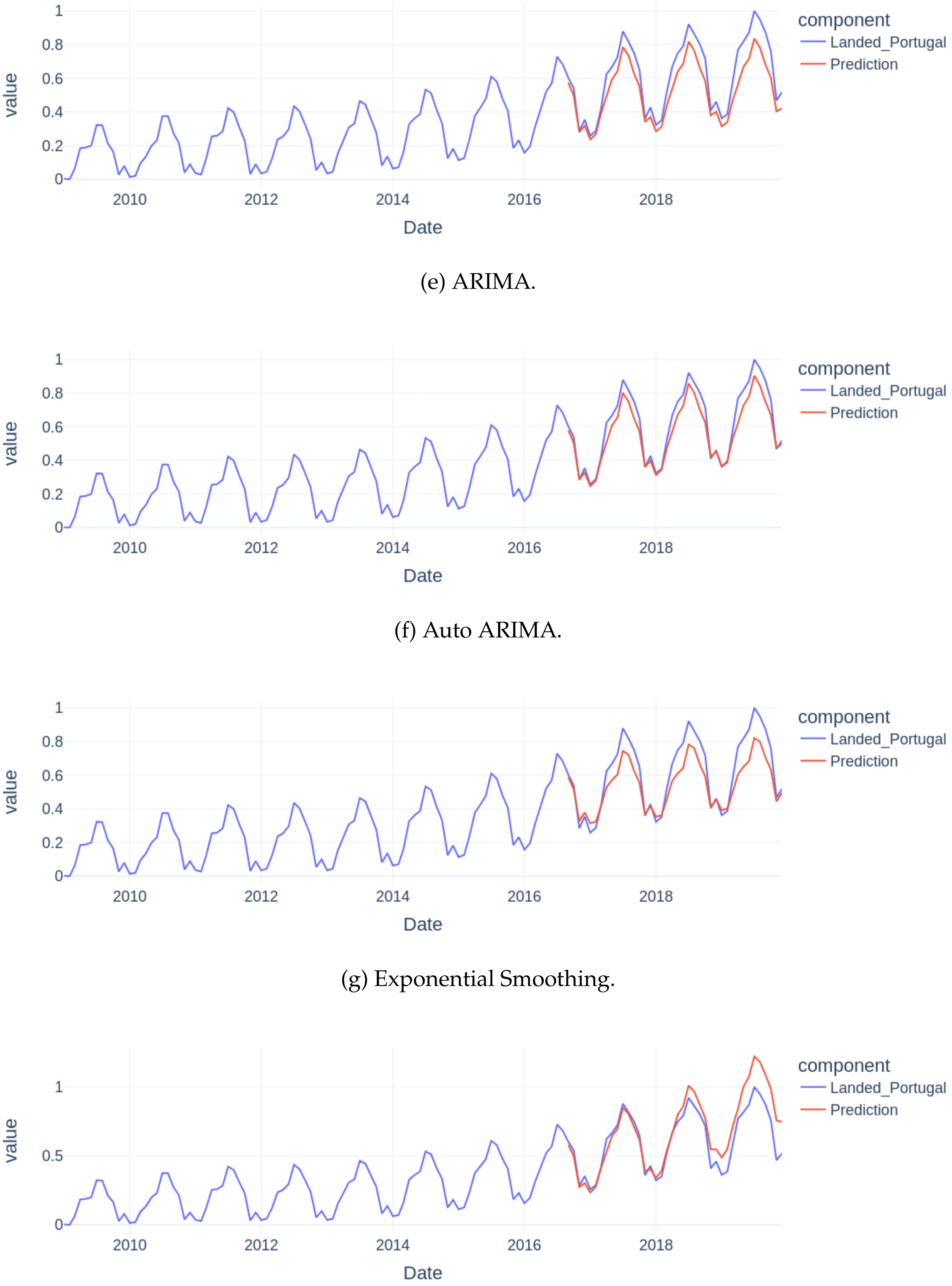
(a) NAIVE Mean.

(b) NAIVE Seasonal.

(c) NAIVE Drift.

(d) NAIVE Moving Average.

Figure A.1: INE baseline prediction by model (NAIVE Mean, NAIVE Seasonal, NAIVE Drift and NAIVE Moving Average).



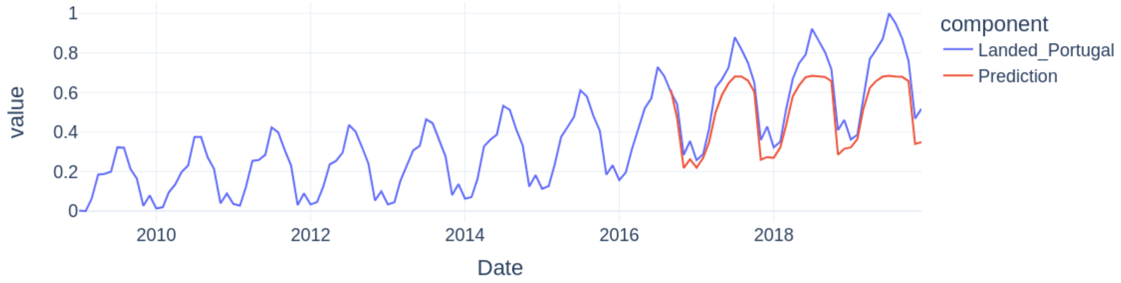
(e) ARIMA.

(f) Auto ARIMA.

(g) Exponential Smoothing.

(h) Regression Model.

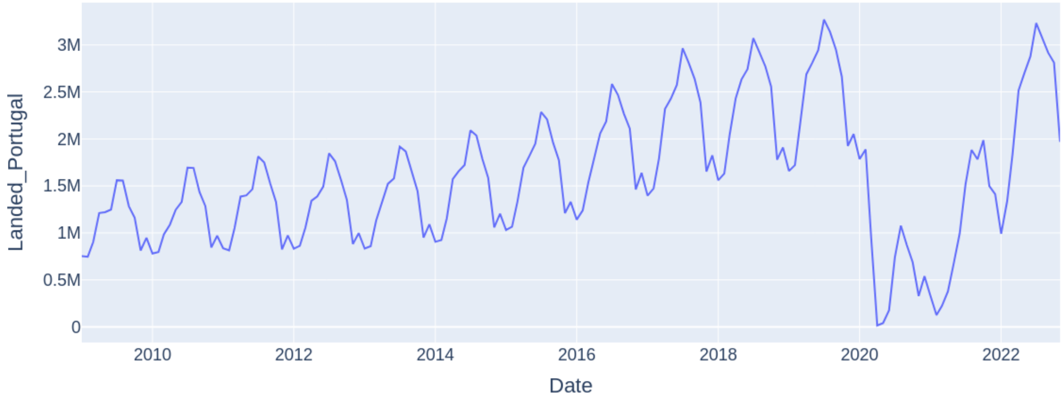
Figure A.1: INE baseline prediction by model (ARIMA, Auto ARIMA, Exponential Smoothing and Regression Model).



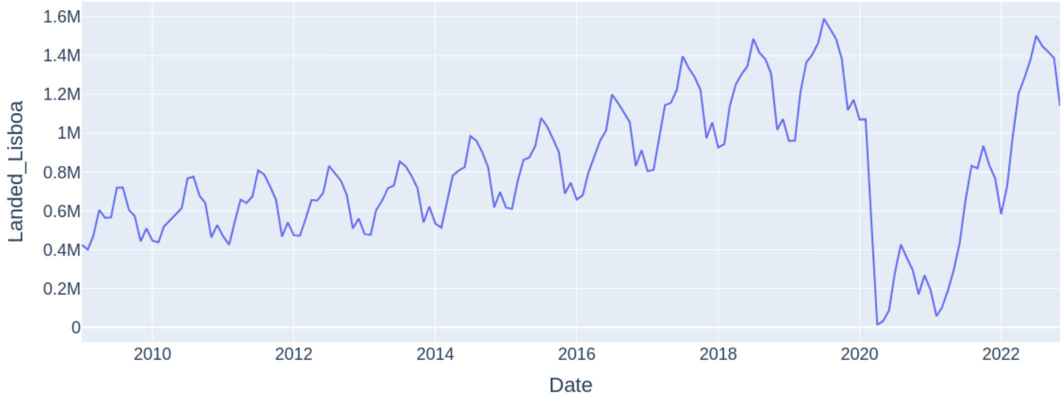
(i) Random Forest.

Figure A.1: INE baseline prediction by model (Random Forest).

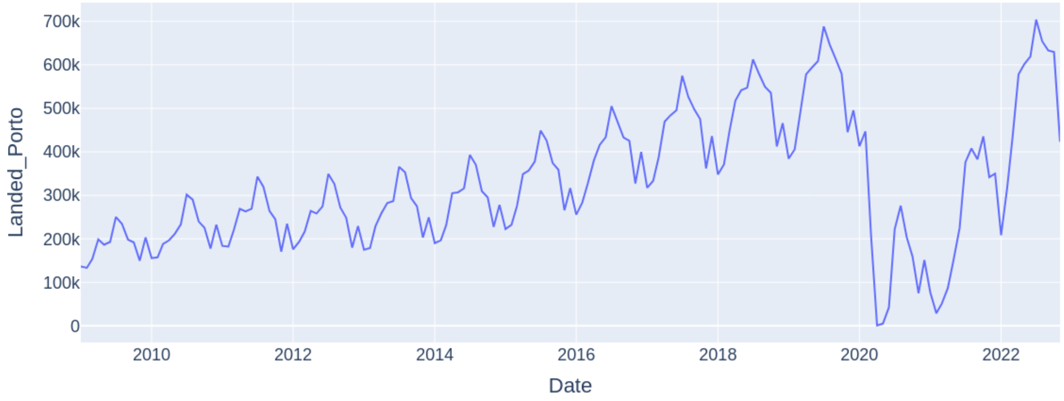
## A.2 INE Data Visualization



(a) Portugal.

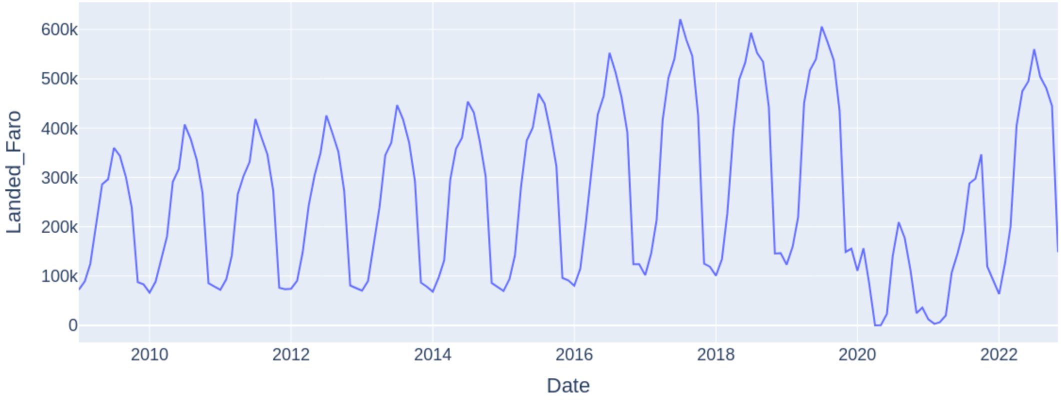


(b) Lisbon.

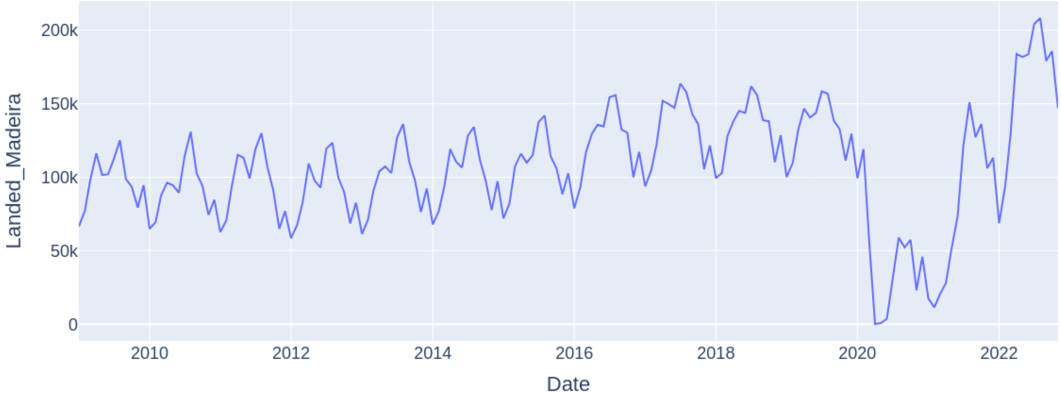


(c) Porto.

Figure A.2: INE data. Number of passengers landed in each airport over time.

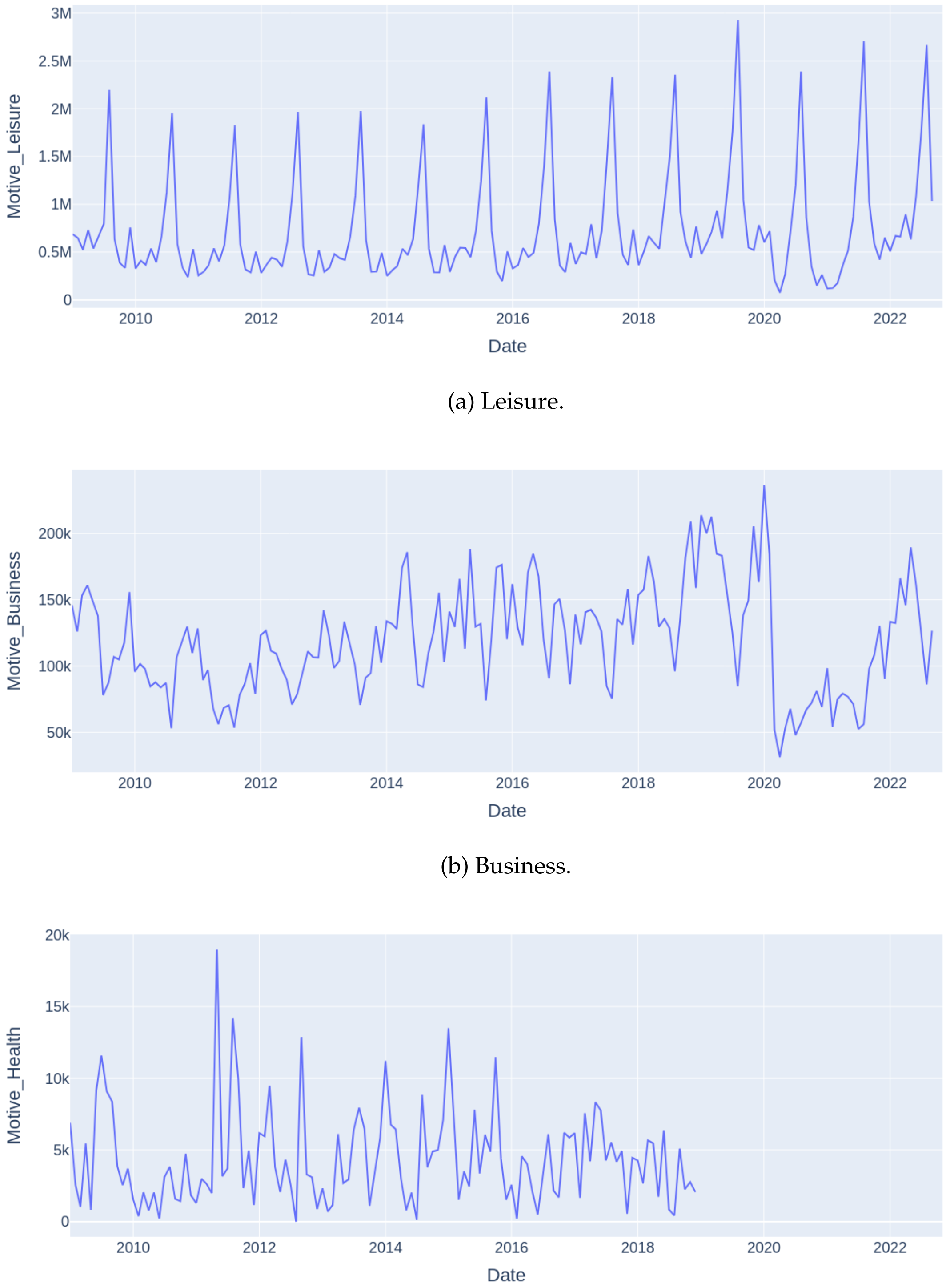


(d) Faro.



(e) Madeira.

Figure A.2: INE data. Number of passengers landed in each airport over time.

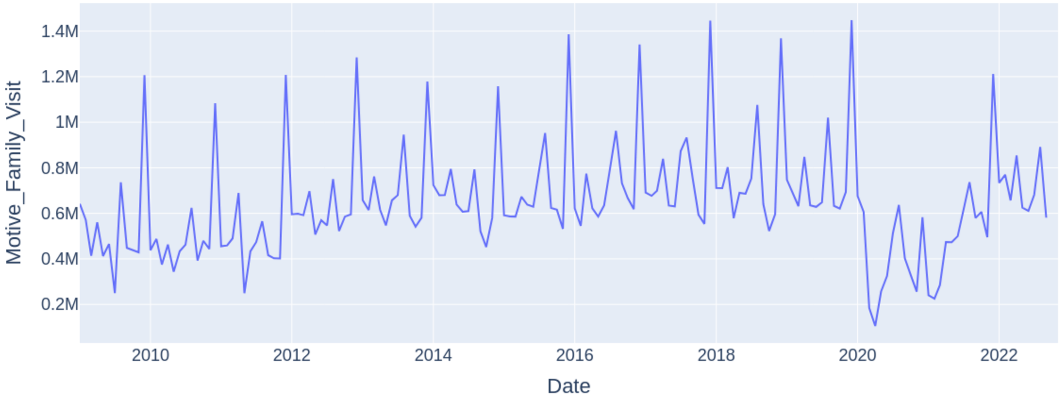


(a) Leisure.

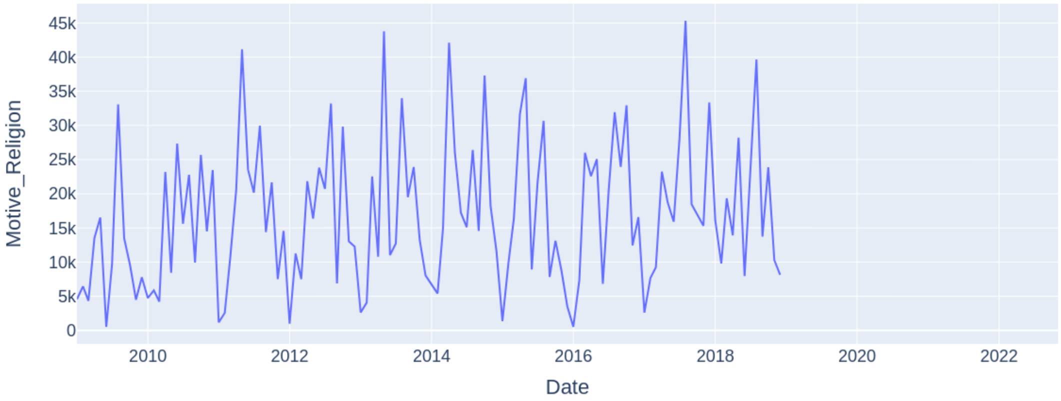
(b) Business.

(c) Health.

Figure A.3: INE data. Number of travels for each motive over time.

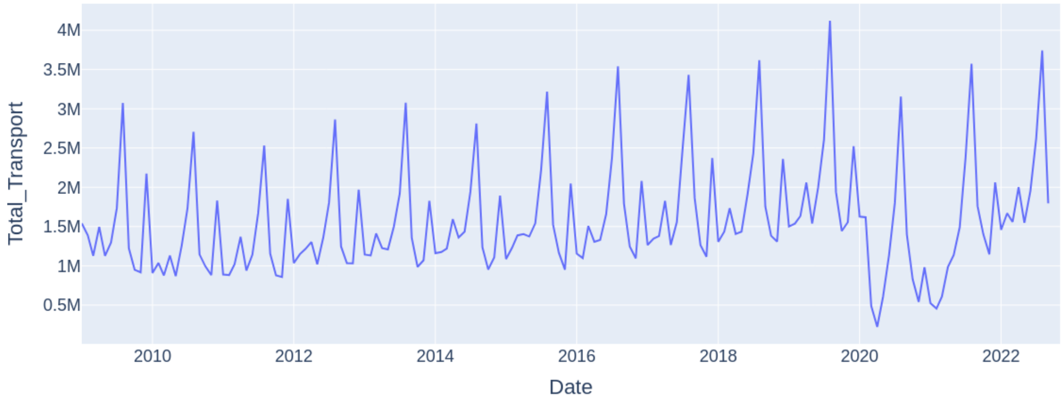


(d) Familiy Visit.

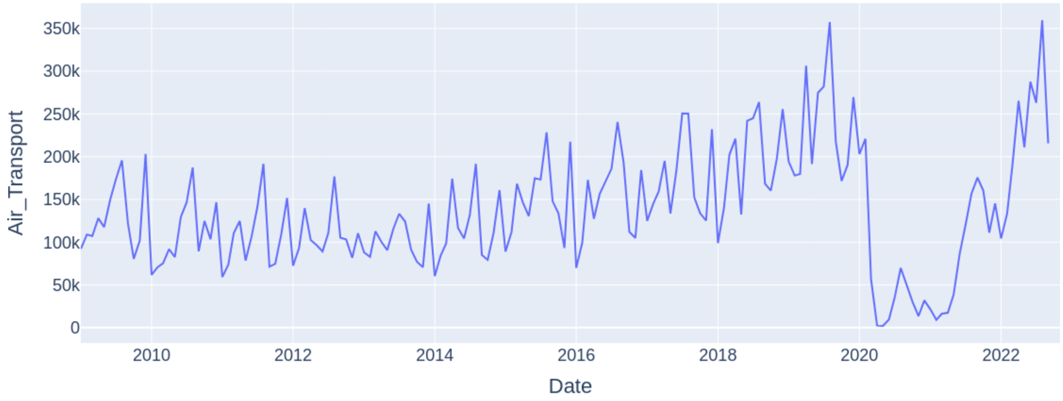


(e) Religion.

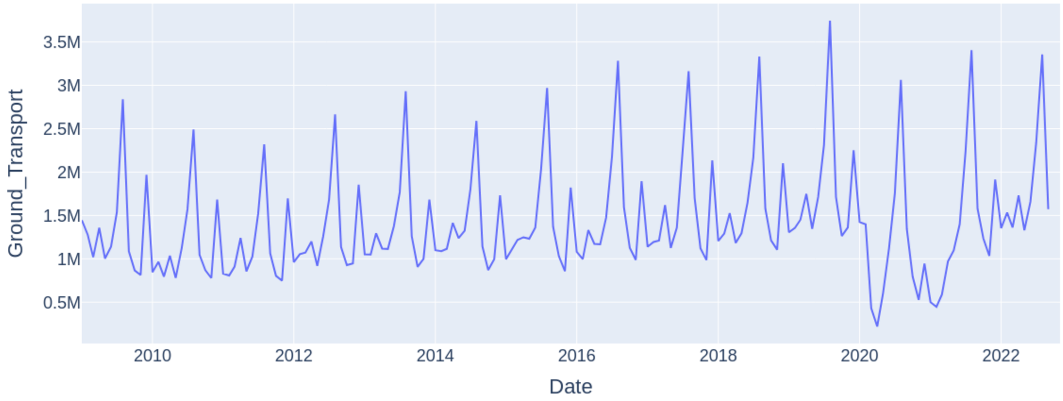
Figure A.3: INE data. Number of travels for each motive over time.



(a) Total.

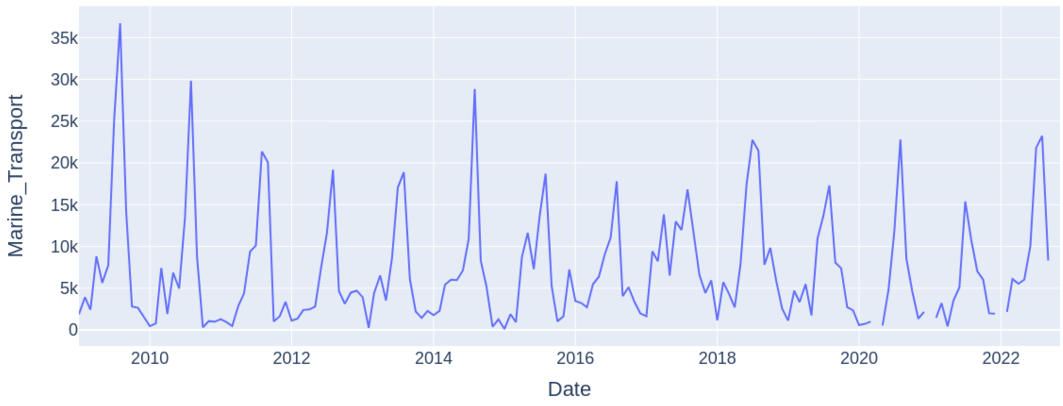


(b) Air.



(c) Ground.

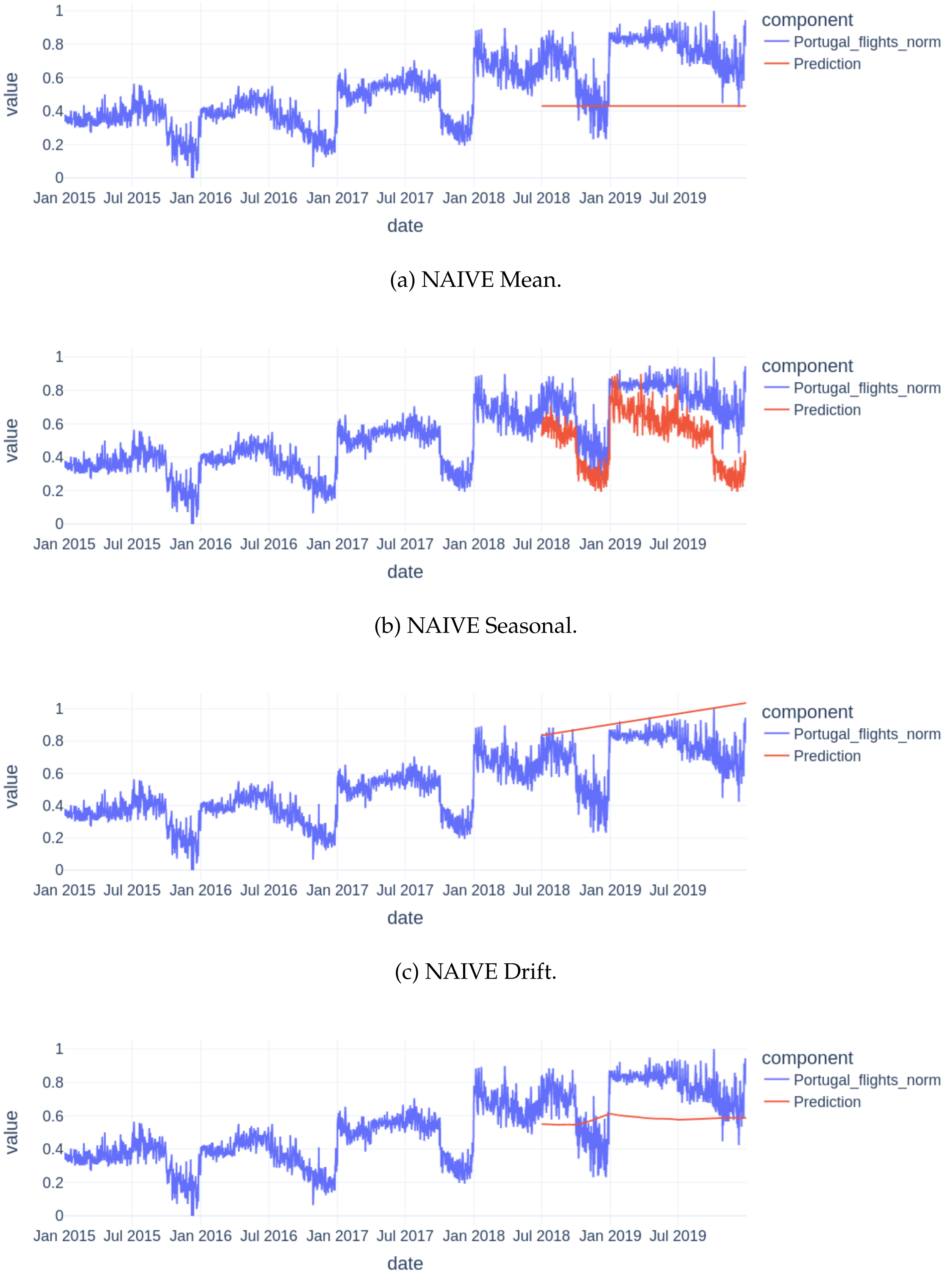
Figure A.4: INE data. Number of travels for each transportation type over time.



(d) Marine.

Figure A.4: INE data. Number of travels for each transportation type over time.

## A.3 Google Trends Baseline Prediction Visualization



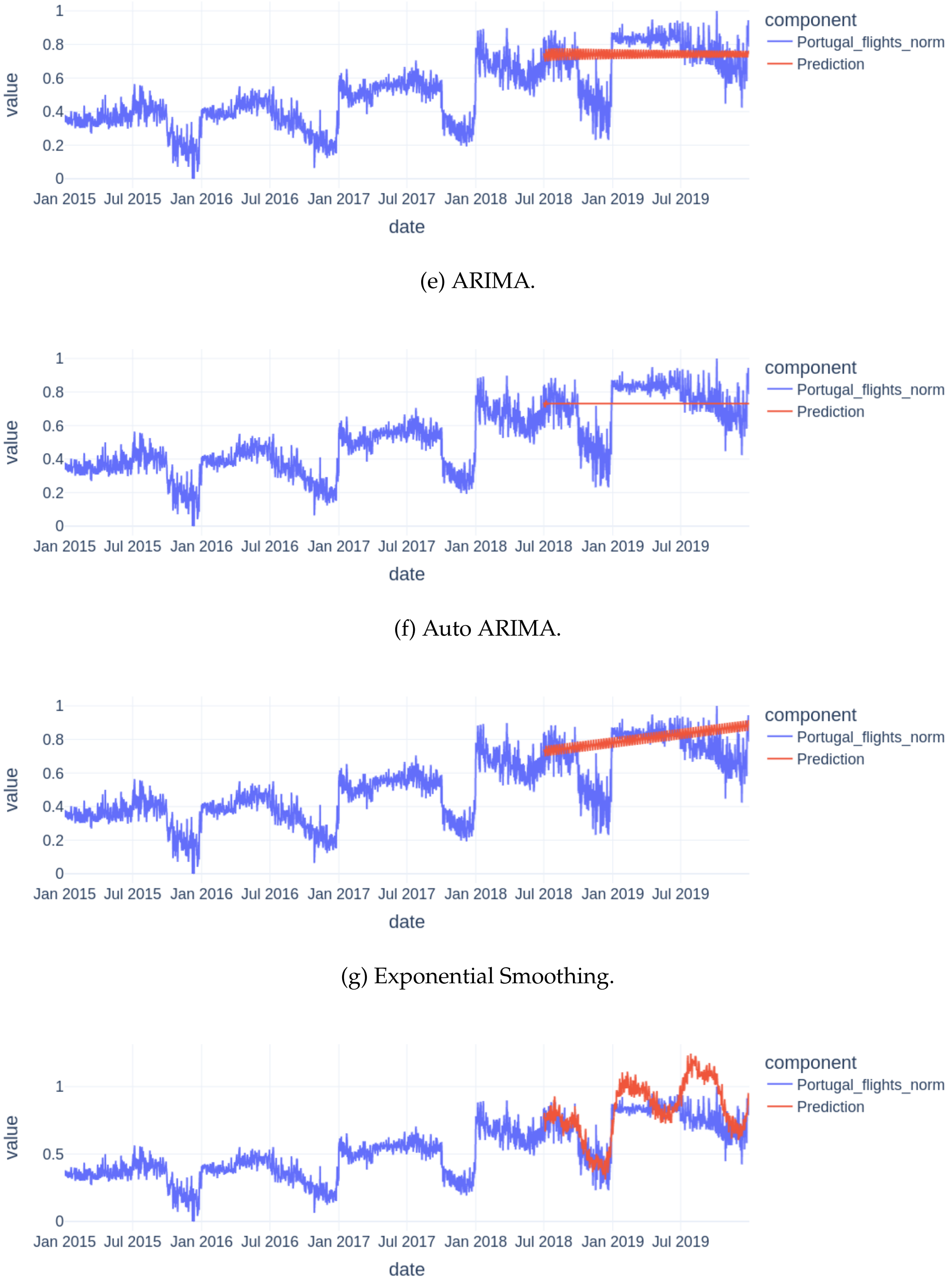
(a) NAIVE Mean.

(b) NAIVE Seasonal.

(c) NAIVE Drift.

(d) NAIVE Moving Average.

Figure A.5: Google Trends baseline prediction by model (NAIVE Mean, NAIVE Seasonal, NAIVE Drift and NAIVE Moving Average).



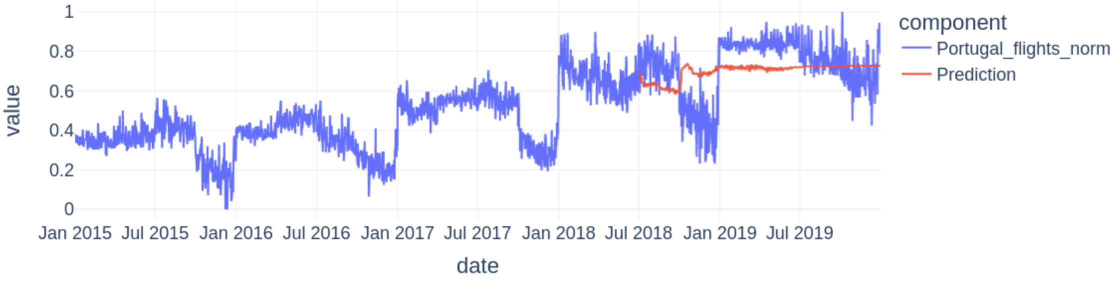
(e) ARIMA.

(f) Auto ARIMA.

(g) Exponential Smoothing.

(h) Regression Model.

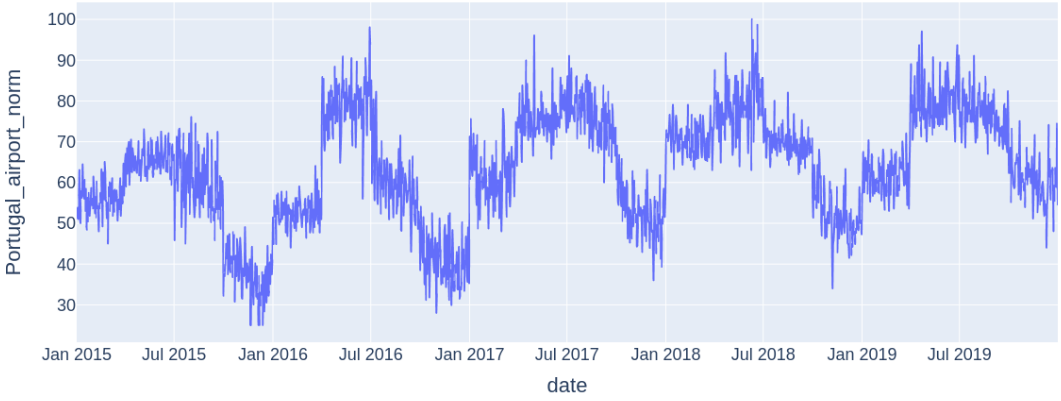
Figure A.5: Google Trends baseline prediction by model (ARIMA, Auto ARIMA, Exponential Smoothing and Regression Model).



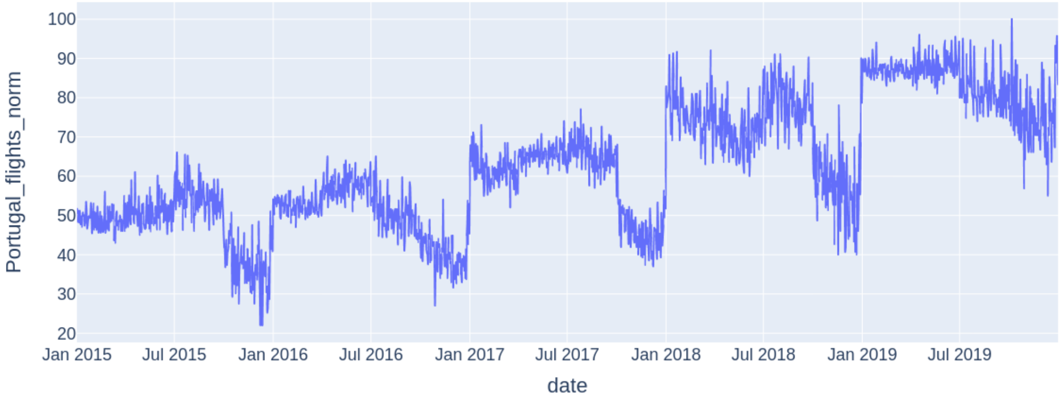
(i) Random Forest.

Figure A.5: Google Trends baseline prediction by model (Random Forest).

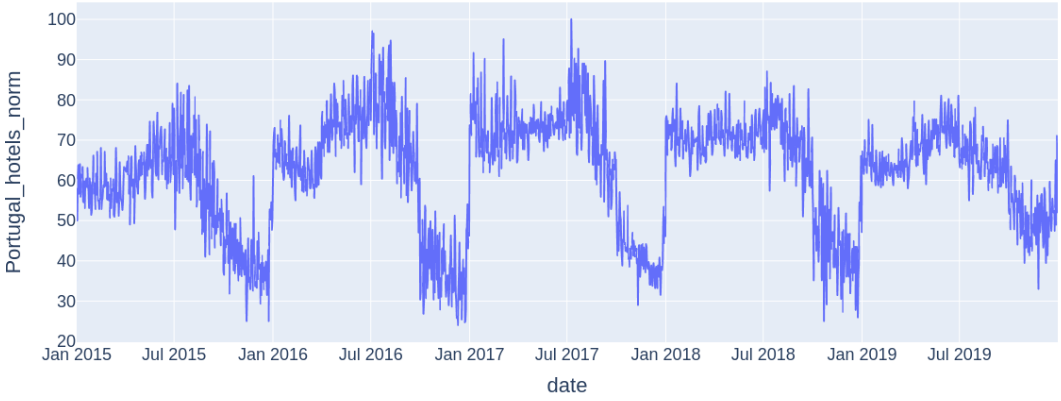
## A.4 Google Trends Data Visualization



(a) Portugal Airport.

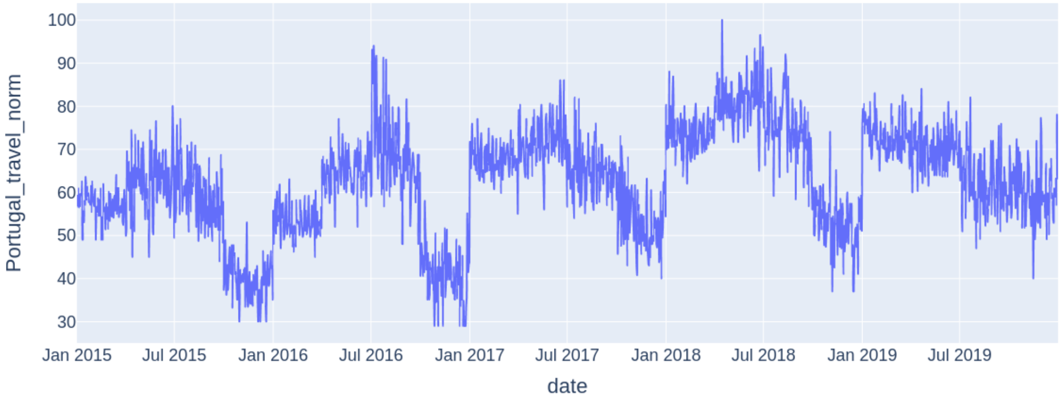


(b) Portugal Flights.



(c) Portugal Hotels.

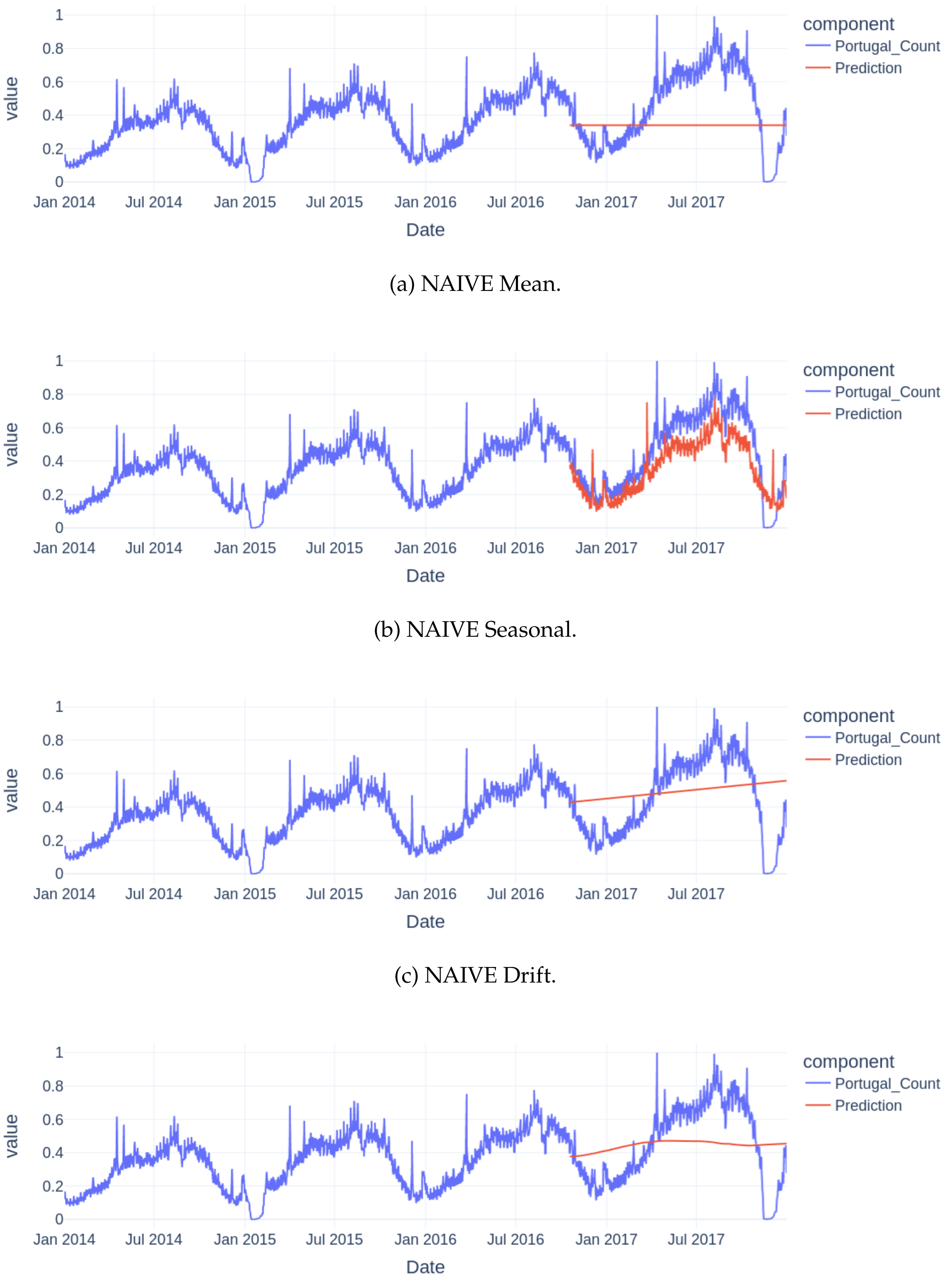
Figure A.6: Search volume on Google's engine by keyword over time.



(d) Portugal Travel.

Figure A.6: Search volume on Google's engine by keyword over time.

## A.5 SEF Baseline Prediction Visualization



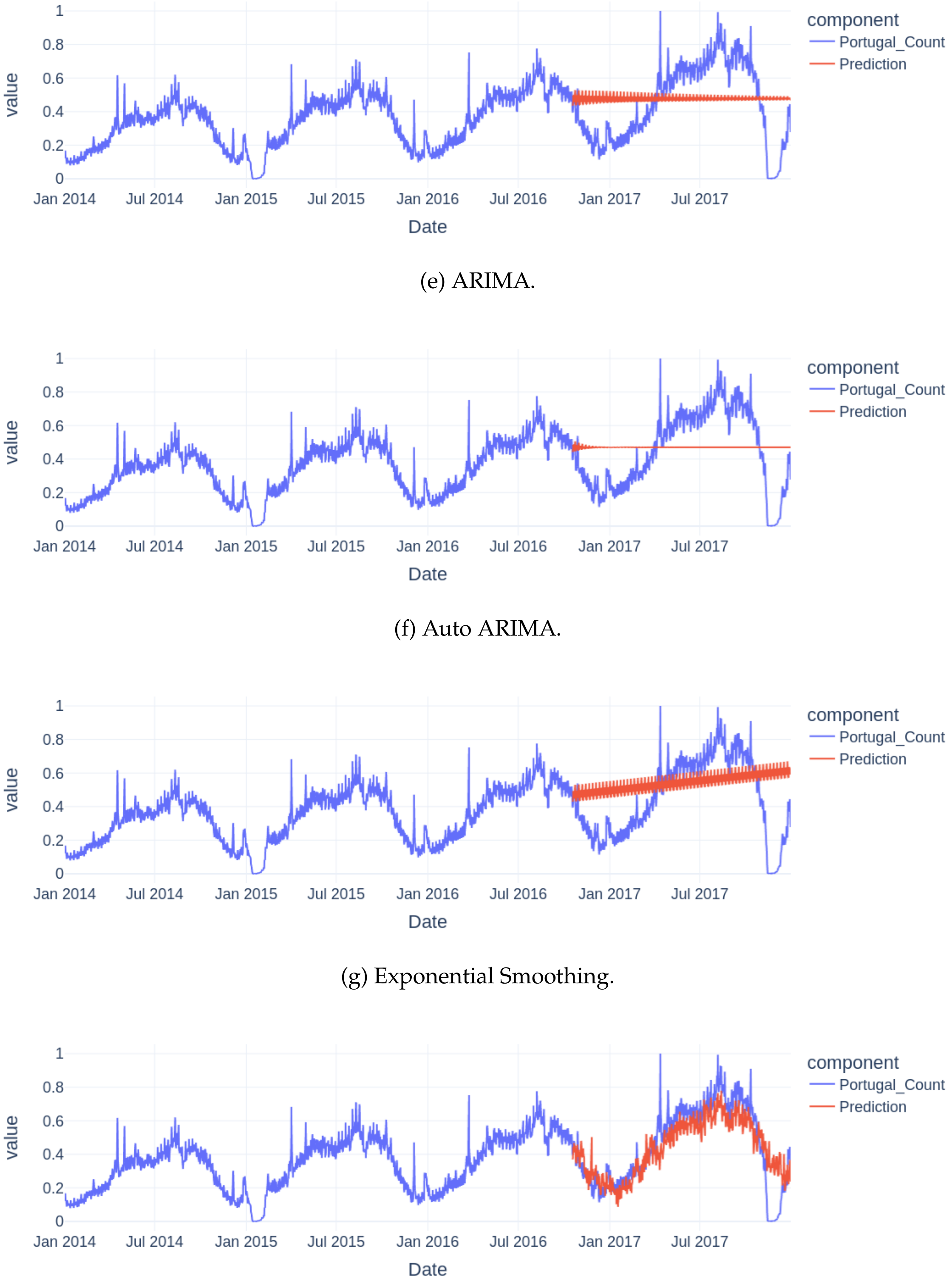
(a) NAIVE Mean.

(b) NAIVE Seasonal.

(c) NAIVE Drift.

(d) NAIVE Moving Average.

Figure A.7: SEF baseline predictions by model (NAIVE Mean, NAIVE Seasonal, NAIVE Drift and NAIVE Moving Average).



(e) ARIMA.

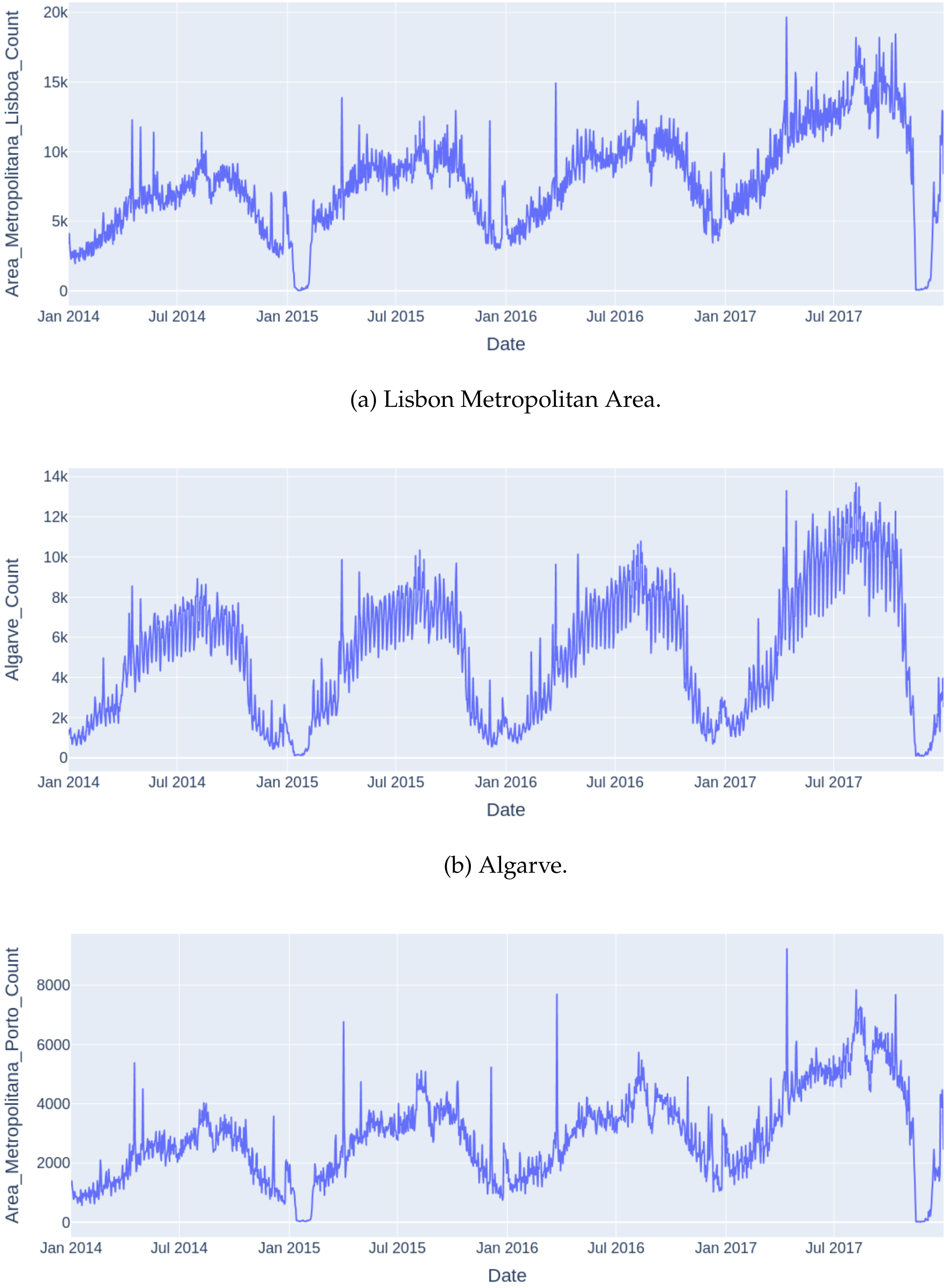
(f) Auto ARIMA.

(g) Exponential Smoothing.

(h) Regression Model.

Figure A.7: SEF baseline predictions by model (ARIMA, Auto ARIMA, Exponential Smoothing and Regression Model).

## A.6 SEF Data Visualization

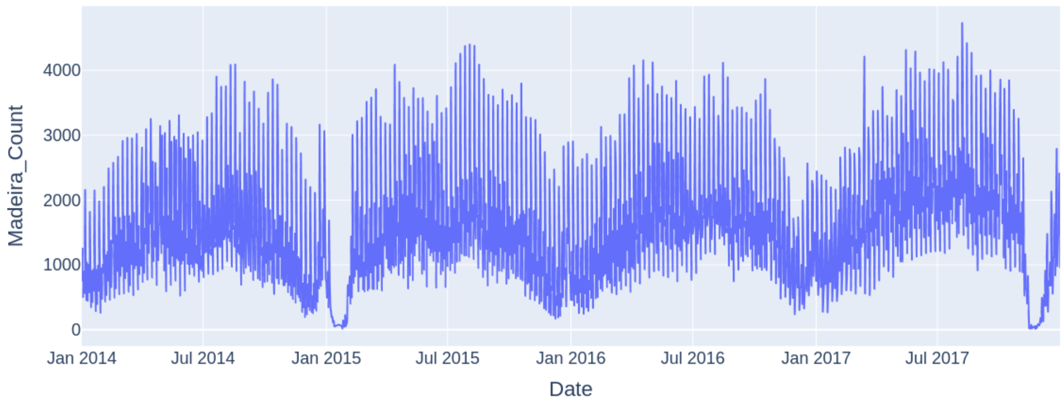


(a) Lisbon Metropolitan Area.

(b) Algarve.

(c) Porto Metropolitan Area.

Figure A.8: SEF data. Number of accommodations over time in each NUTS III Region presented.



(d) Madeira.

Figure A.8: SEF data. Number of accommodations over time in each NUTS III Region presented.