SHORT-TERM FORECASTING OF TOURIST ARRIVALS IN MALTA WITH GOOGLE TRENDS DATA

Lynn Cumbo

Dissertation submitted in partial fulfilment of the award of the degree of Master of Science in Economics at the University of Malta

September 2022

ABSTRACT

The aim of this dissertation is to assess whether a model with Google Trends search query data is able to provide more accurate forecasts of tourist arrivals in Malta than a model without such data. Google Trends search query data is collected following a careful methodological process to select queries that are related to travelling to Malta. A Seasonal Autoregressive Integrated Moving Average model is employed as a benchmark model. Principal Component Analysis is conducted on the Google Trends data collected and the principal components are added as predictors to the benchmark model to obtain the competing Google Trends model.

Pseudo-out-of-sample one-step ahead forecast simulations are carried out from both models estimated on the period January 2004 – December 2016 and their forecasting performance is tested on the sample of observations from January 2017 – December 2019 using measures of the mean error. The results reveal that the model with Google Trends data generates better forecasts as the forecast accuracy metrics show a lower error for the forecasts on the testing sample. The robustness of these results is checked via the Clark & West (2007) test for nested models and another pseudo-out-of-sample one-step ahead forecast simulation from a model with European Gross Domestic Product and the Real Effective Exchange as income and price variables, respectively, which are typical explanatory factors of tourism demand. The forecasts from the Google Trends model are robust against these checks which allowed for a demonstration of a practical use of this model in a nowcasting exercise. Tourist arrivals for July 2022 and August 2022 were nowcasted from the model with the Google Trends data which included a dummy variable for April, May and June 2020 to control for the period in which the airport was closed as a measure against the spread of Covid-19.

These findings elicit important implications for private sector businesses in the tourism industry, public sector institutions such as the Malta Tourism Authority, national institutions concerned with projections of the Maltese macroeconomy like the Central Bank of Malta and various other sectors in the economy which are impacted by the indirect and induced effects of tourist activity.

Keywords: Forecasting, Nowcasting, Google Trends, Tourism

JEL Codes: C22, C53, C82, L83**,** Z32

Dedicated to my first economics teacher,

Mrs Rita Scicluna,

Thank you for believing that I would be able to
become an economist from just thirteen years of age.

ACKNOWLEDGEMENTS

I would like to show my sincere gratitude towards my supervisor Dr Ian P. Cassar for his invaluable advice and time which made the completion of this dissertation possible; it has not only been an academic learning experience but also an intellectual one. I extend this appreciation to all the lecturers at the Department of Economics.

My genuine thanks goes to Reuben Ellul who was the person that proposed the idea of this dissertation and encouraged me to pursue it. I am also indebted to Ian Borg at the Central Bank of Malta for providing valuable insights and suggestions from his expertise in the field.

I am eternally grateful for my parents; Andrew and Graziella, who always supported me since the start of my academic journey, my sister; Yana, who inspires me to give the best of my efforts in everything, my close friends and lastly, my boyfriend; Kurt, thank you for your support, encouragement and most importantly, for the time spent helping me with coding in *R* and patiently teaching me in the process.

TABLE OF CONTENTS

ABSTRACT.....ii

ACKNOWLEDGEMENTS.....iv

TABLE OF CONTENTS.....v

LIST OF TABLES.....ix

LIST OF FIGURES.....x

LIST OF ABBREVIATIONS.....xi

CHAPTER 1 INTRODUCTION.....1

1.1 INTRODUCTION.....2

1.2 RESEARCH QUESTION AND CONTRIBUTION.....3

1.3 DISSERTATION OUTLINE.....3

CHAPTER 2 LITERATURE REVIEW.....5

2.1 FORECASTING TOURISM DEMAND.....7

2.1.1 Modelling Tourism Demand.....7

2.1.2 Empirical Approaches to Forecasting Tourism Demand.....8

2.1.3 Forecasting Performance of Tourism Demand Models.....10

2.2 FORECASTING TOURISM DEMAND WITH GOOGLE TRENDS SEARCH QUERY DATA.....10

2.2.1 Google Trends Data as a Predictor.....10

2.2.2 Evaluation of Forecasting Performance.....12

2.2.3 Significance of Google Trends Data in Forecasting and Nowcasting Studies.....12

CHAPTER 3 METHODOLOGY & DATA.....14

3.1 FORECASTING THEORETICAL FRAMEWORK.....16

3.1.1 Forecasting from a Time Series Model.....16

3.1.2 Forecast Errors and Prediction Intervals 18
3.1.3 Evaluation of Forecasting Performance 20
3.1.4 Approaches to Pseudo-Out-of-Sample Forecasts 21
3.2 THE AUTOREGRESSIVE INTEGRATED MOVING AVERAGE METHODOLOGY 21
3.2.1 The Box-Jenkins Approach 21
3.2.2 The ARIMA Model 24
3.2.3 The Seasonal ARIMA Model 25
3.3 PROCEDURES IMPLEMENTED FOR ASSESSING FORECASTING PERFORMANCE 26
3.3.1 Model Selection 26
3.3.2 Pseudo-Out-of-Sample Forecasting Simulation 28
3.3.3 Robustness Checks 29
3.4 THE DATASET 31
3.4.1 The Target Variable: Tourist Arrivals 31
3.4.2 The Predictor Variable: Google Trends Data 32

CHAPTER 4 RESULTS & DISCUSSION 40
4.1 GOOGLE TRENDS DATA 42
4.1.1 Correlation with Arrivals Series 42
4.1.2 Principal Component Analysis 44
4.2 MODEL SELECTION AND EVALUATION 50
4.2.1 Stationarity 50
4.2.2 The Benchmark Model 51
4.2.3 The Google Trends Model 53
4.3 FORECASTING SIMULATION 54
4.3.1 Overall Forecasting Performance 54
4.3.2 Relative Performance of Benchmark and Google Trends Model 55

4.4 ROBUSTNESS CHECKS 57
4.4.1 Clark & West (2007) Test 57
4.4.2 Pseudo-Out-of-Sample Forecasting Simulation from a Model with GDP and REER 58
4.5 NOWCASTING EXPERIMENT 60
4.5.1 Nowcasting Methodology 61
4.5.2 Nowcasting Results 62
4.6 DISCUSSION 64
4.6.1 Summary of Results 64
4.6.2 Nowcasts 66
4.6.3 Implications for Private and Public Sector Institutions 66
4.6.4 Comparison of Findings with Other Studies 67

CHAPTER 5 CONCLUSION 69
5.1 MAIN FINDINGS 70
5.2 LIMITATIONS 70
5.3 SUGGESTIONS FOR FURTHER RESEARCH 71

REFERENCES 73

APPENDIX A 73
Appendix A.1: Google Trends Data and Other Economic Variables 74
The Labour Market 74
Consumption 75
The Exchange Rate 75
APPENDIX B 76
Appendix B.1: Principal Component Analysis 77
The Optimisation Problem 77

Summary 78
Appendix B.2: Variables Employed for Robustness Check 79
Real Gross Domestic Product 79
Real Effective Exchange Rate 80
Appendix B.3: Recursive Window Forecasting Approach 82
Appendix B.4: The Dickey-Fuller (1979) Test and the Ljung-Box (1978) Q-statistic 83
Appendix B.5: Search Queries Retrieved from Top and Rising Filters in Google Trends 85
Appendix B.6: Eliminated Search Queries 91
APPENDIX C 97
Appendix C.1: Results from Correlation Analysis 98
Appendix C.2: List of 33 Queries in the Final Google Trends Dataset 103
Appendix C.3: Stationarity Results from ADF Test on Tourist Arrivals and First Principal Components 104
Appendix C.4: Diagnostic Test Results on the $SARIMA(1,1,9)(1,0,1)$ 106
Appendix C.5: Adjustment to Training Sample in the Estimation of $SARIMA(1,1,0)(0,1,1)$ Model 107
Appendix C.6: Stationarity Results from ADF Test on GDP and the REER 108
Appendix C.7: Results from Correlation Analysis, PCA, Stationarity Tests and Model Estimation for the Nowcasting Experiment 111

LIST OF TABLES

Table 3.1: Specification of Seasonal and Non-Seasonal Terms. 27
Table 3.2: 11 Seed Queries Entered into Google Trends Based on 6 Aspects of Trip Planning. 34
Table 3.3: Specification of Geographical Location, Time Range, Category and Type of Search Filters in Google Trends. 34

Table 4.1: Loadings of the First Principal Component for the Dataset Lagged by 5 Months. 47
Table 4.2: Loadings of the First Principal Component for the Dataset Lagged by 4 Months. 47
Table 4.3: Loadings of the First Principal Component for the Dataset Lagged by 1 Month...48
Table 4.4: Ljung-Box Test Results on SARIMA(1,1,0)(0,1,1) Estimated on Training Sample. 52
Table 4.5: SARIMA(1,1,0)(0,1,1) Model Estimation Results on Training Sample. 53
Table 4.6: Google Trends Model Estimation Results on Training Sample. 54
Table 4.7: Overall Forecast Accuracy for One-Step Ahead Forecasts from the Benchmark Model and from the Google Trends Model. 55
Table 4.8: Results from Clark & West (2007) Test. 57
Table 4.9: SARIMA (1,1,0)(0,1,1) Model with GDP and the REER as Explanatory Variables Estimation Results on Training Sample. 58
Table 4.10: Overall Forecast Accuracy for One-Step Ahead Forecasts from the Explanatory Variables Model and from the Google Trends Model. 59
Table 4.11: Nowcasted Tourist Arrival Values. 64
Table 4.12: Forecast Accuracy for Each Year of the Testing Sample. 65

LIST OF FIGURES

Figure 3.1: Monthly Tourist Arrivals (Jan 2004 – Dec 2019). 32
Figure 3.2: Summary of Google Trends Data Collection and Reduction Process (left pane) and Number of Queries after each Step (right pane). 36
Figure 3.3: Google Trends Search Query Data (Jan 2004 – Dec 2019). 37

Figure 4.1: Summary of Google Trends Data Collection and Reduction Process (Left Pane) and Number of Queries after each Step (Right Pane). 43
Figure 4.2: Scree Plot of 17 Principal Components from the PCA on Dataset Lagged by 5 Months. 44
Figure 4.3: Scree Plot of 5 Principal Components from the PCA on Dataset Lagged by 4 Months. 45
Figure 4.4: Scree Plot of 11 Principal Components from the PCA on Dataset Lagged by 1 Month. 45
Figure 4.5: Time Series Plot of the First Principal Component of the Lagged by 5 Months Dataset 49
Figure 4.6: Time series Plot of the First Principal Component of the Lagged by 4 Months Dataset 49
Figure 4.7: Time Series Plot of the First Principal Component of the Lagged by 1 Month Dataset 50
Figure 4.8: Residual Plot, ACF and PACF for SARIMA(1,1,0)(0,1,1) Estimated on Training Sample 52
Figure 4.9: Google Trends Model RMSE Relative to Benchmark Model RMSE for each Month in the Testing Sample. 56
Figure 4.10: Explanatory Variables Model RMSE Relative to Google Trends Model RMSE for One-Step Ahead Forecasts (red dotted line is equal to 1). 60
Figure 4.11: Nowcasts of Tourist Arrivals for July and August 2022 (black line) with 90% (darkest red), 95% and 99% (lightest red) Confidence Intervals. 63

LIST OF ABBREVIATIONS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
CPI	Consumer Price Index
ECB	European Central Bank
GDP	Gross Domestic Product
GVA	Gross Value Added
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MIA	Malta International Airport
MIDAS	Mixed Data Sampling
MLE	Maximum Likelihood Estimation
MSPE	Mean Squared Percentage Error
MTA	Malta Tourism Authority
NSO	National Statistics Office
OLS	Ordinary Least Squares
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
REER	Real Effective Exchange Rate

RMSE	Root Mean Squared Error
SAR	Seasonal Autoregressive
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Explanatory Variables
SIC	Schwarz Information Criterion
UK	United Kingdom
USA	United States of America
VAR	Vector Autoregressive

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The tourism sector in Malta has grown and evolved to be a fundamental pillar of the Maltese economy since the 1950s (Attard, 2019). The value of the economic contribution of tourism to the Maltese economy differs according to the modelling framework used to provide this estimate, however, according to Cassar et al. (2016) tourism accounts for around 5.7% of total Gross Value Added (GVA) when considering direct effects only, 12% of total GVA with the consideration of indirect effects and approximately 17% of total GVA including induced effects.

In view of the importance of tourism for Malta's economic activity, inbound tourism statistics, especially the number of tourist arrivals, are observed by several institutions as an indicator of demand for tourism in Malta. In this regard, accurate and timely forecasts of inbound tourists are a requirement for the various sectors in the economy concerned with the activities of tourists. The availability of short-term forecasts of tourist arrivals is also relevant to macroeconomic projections of the Maltese economy due to the contribution of the industry to Malta's travel exports.

Google Trends data has recently been applied in providing forecasts of important macroeconomic variables due to its real-time availability. Google Trends is a free product offered by Google, the world's most popular search engine (Statista, 2022), that can be accessed by anyone, anytime and anywhere. It provides a sample of actual search queries made by Google users, which is made anonymous, categorised into topics and aggregated (Google, n.d.). Search queries made to Google are reflective of user interest in a subject matter, thus Google Trends is a way to quantify interest and has grown to be a useful tool in explaining and predicting humans' wants, needs and intentions. The data is a potential aid in the forecasts of tourist arrivals since a lot of holiday planning and booking is done online through the internet according to the last wave of the *Preferences of Europeans towards Tourism* survey (European Commission, 2016). Therefore, searches for accommodation, flights and activities in Malta indicate interest in travelling to Malta and thus can be used as an input to forecast tourist arrivals.

The importance of this sector in the Maltese economy coupled with the emergence of such data provides the motivation to assess whether a model of tourist arrivals with the inclusion of Google Trends data is able to generate accurate and timely forecasts.

1.2 RESEARCH QUESTION AND CONTRIBUTION

The research question that this dissertation aims to answer is:

"Does Google Trends data improve forecasts of tourist arrivals in Malta?"

The attempt to answer this question is done by performing pseudo-out-of-sample one-step ahead forecasts from a benchmark model without such data and from a competing model that incorporates this data. The models used are Seasonal Autoregressive Integrated Moving Average models and a process to select relevant queries from Google Trends is applied. The data is included in the model by means of Principal Component Analysis on the large number of selected search queries. The difference between the monthly forecast of tourist arrivals and the actual value of tourist arrivals during that month indicates which model is able to generate the more accurate forecasts. This is measured by metrics of forecasting performance which use the value of the forecast error in this evaluation.

Several studies have attempted to answer the same research question in the context of other countries and cities, namely Antolini & Grassini (2019) for Italy, Artola et al. (2015) for Spain, Havranek & Zeynalov (2021) for Prague and Park et al. (2016) for South Korea amongst others referenced throughout this dissertation, however, the potential of Google Trends data to forecast inbound tourists has not been applied to Malta. Therefore, the contribution of this research to the literature is to include Malta in the series of studies on this topic.

1.3 DISSERTATION OUTLINE

The structure of this dissertation is organised as follows: ***Chapter 2*** is a review of the literature on the standard methods used to forecast tourism demand and also the approach taken by other studies to answer a similar research question. ***Chapter 3*** contains an extensive review of the methods that shall be employed, an explanation of two robustness checks that shall be carried out and a description of the dataset together with an overview of the process applied to collect the Google Trends data. The nature of this dissertation is highly technical and statistically oriented; therefore, this chapter is the substance of the research. The results from the analysis of the Google Trends data, the forecast simulations and the robustness checks are presented

and discussed in ***Chapter 4*** together with a nowcast experiment that stemmed from the results attained. ***Chapter 5*** concludes with the limitations encountered during the course of the study and some suggestions for further research.

CHAPTER 2
LITERATURE REVIEW

LITERATURE REVIEW

This dissertation focuses on a niche field of economic literature which is statistically oriented, therefore, the literature review is a concise overview of the approaches to forecasting tourism demand that shall inform the methodology of this research. ***Section 2.1*** is a review of how tourism demand is usually modelled and forecasted without the use of Google Trends data while ***Section 2.2*** covers the methods adopted by studies that examined a similar research question for different countries together with their findings.

Appendix A.1 provides a brief review of the literature on the application of Google Trends data to forecast other aspects of the economy which contained some relevant insights for this dissertation.

2.1 FORECASTING TOURISM DEMAND

2.1.1 Modelling Tourism Demand

In empirical studies, tourism demand is defined as "the willingness and ability of consumers to buy different amounts of a tourism product at different prices during any one time period" (Dwyer et al., 2020, p.17). Tourism demand can be measured by the number of tourist arrivals, tourist expenditure, length of stay, nights spent and variations of these in per capita terms or disaggregated in terms of purpose of travel or country of origin, according to comprehensive reviews of the literature by Li et al. (2005), Lim (1997), Song et al. (2012) and Song & Li (2008).

Song et al. (2010) compare econometric models and their forecasting performance based on tourist arrivals and tourist expenditure, in aggregate and per capita form, as the dependent variable. Aggregate tourist expenditure and arrivals are more accurately predicted than per capita measures suggesting that models for forecasting purposes should be specified in aggregate form.

Systematic literature reviews and mainstream literature on models of tourism demand show that the following are the frequently used explanatory factors in tourism demand equations with tourist arrivals as the dependent variable: population, income, own price of touring, substitute prices, transport costs, exchange rates, tastes, marketing, expectations, habit persistence, lags of the dependent variable and qualitative factors (Athanasopoulos et al., 2011; Dwyer et al., 2020; Li et al., 2005; Lim, 1997, 1999; Song et al., 2012; Song & Li, 2008 and Song & Witt, 2000).

The inclusion of all these factors in a tourism demand model is not the case in practice because of multicollinearity.[1] (Li et al., 2005; Lim, 1997, 1999 and Smeral, 1988). Furthermore, the use of explanatory variables in forecasting tourism demand with time series methods is usually not required since independent variables are mostly included in econometric analyses whereby the

[1] Multicollinearity refers to the statistical issue wherein two or more explanatory variables are "approximately linearly related" in a regression model (Gujarati & Porter, 2010, p.245).

purpose is to identify the relationship of such variables with tourism demand (Song & Li, 2008).

2.1.2 Empirical Approaches to Forecasting Tourism Demand

The modelling and forecasting of tourism demand can adopt a qualitative or quantitative approach. Since this dissertation employs a quantitative method using time series models coupled with econometric analysis, this section provides an overview of the quantitative methods employed in the literature.

Quantitative methods are sub-categorised into non-causal time series models and causal econometric models. Time series models explain tourism demand using past realisations and a random disturbance term together with regard for historical trends and seasonal patterns in order to derive forecasts based on these features of the series (Song & Li, 2008). On the other hand, according to a methodological review of tourism demand forecasting by Goh & Law (2011), "the objective of an econometric tourism demand model is to explain and forecast the future movement of tourism demand based on the quantitative relationship between quantity demanded and its determinants" (p. 297-298).

2.1.2.1 Non-Causal Time Series Methods

The Box-Jenkins (1976) autoregressive integrated moving average (ARIMA) model is the most frequently used time series model in modelling and forecasting tourism demand. As is further explained in the subsequent chapter, an ARIMA model involves the combination of an autoregressive (AR) and moving average (MA) model and the transformation of the series into a stationary process through differencing, hence being *integrated* (Aljandali & Tatahi, 2018 and Verbeek, 2004). Seasonal ARIMA (SARIMA) models have gained popularity as the seasonal variation in tourism demand should not be ignored in the modelling process. SARIMA models are an extension to the ARIMA model such that they contain seasonal AR, MA and differencing components in conjunction with the non-seasonal part of the ARIMA model (Song & Li, 2008).

Other time series models employed in the literature include naïve models (random walk models) and exponential smoothing models. The naïve model shows that forecasted values are equal to the latest available value of the variable (Song & Li, 2008) while exponential smoothing models involve the attachment of weights to more recent observations of the series that exponentially decrease with earlier observations in order to obtain forecasts from a weighted average method (Hyndman & Athanasopoulos, 2018). These models together with simple AR models are often used as benchmark models in the literature (Song & Li, 2008).

2.1.2.2 Causal Econometric Models

Econometric models employed in past studies of tourism demand made use of Ordinary Least Squares (OLS) regressions which incorporate explanatory variables and enable causal analysis. However, several models gave rise to spurious regressions as the stationarity properties of the dependent and independent variables were often ignored and hence, the need for more sophisticated econometric models was called for (Song & Li, 2008). These models typically contain elements from time series models which represent important features of tourism demand (Goh & Law, 2011 and Morley, 2009). In this regard, the most popular econometric models in the literature are the Autoregressive Distributed Lag model, the Error Correction Model and the Vector Autoregressive (VAR) model (Li et al., 2006 and Song & Li, 2008).

Notwithstanding the causal inferences that can be drawn from these models, the major disadvantage for forecasting purposes lies in the fact that forecasts of the explanatory variables are necessary prior to generating forecasts for tourism demand (Athanasopoulos et al., 2011 and Song & Li, 2008). The disadvantage of pure time series models is that they lack the inclusion of variables in line with economic theory.

In this regard, the emergence of the combination of time series approaches with econometric models has allowed for the inclusion of independent factors in accordance with economic theory without forgoing the properties of time series models. (S)AR(I)MAX models are an example of this combination, where X refers to explanatory variables added to ARMA, ARIMA, SARIMA or SARMA models, and they have proved to counteract the inconsistent performance of pure time series models and better accommodate quarterly and monthly data (Li et al., 2005 and Song & Li, 2008). However, there is still no consensus on which method

or model specification possesses the best forecasting abilities (Li et al., 2005; Song et al., 2008 and Song & Li, 2008).

2.1.3 Forecasting Performance of Tourism Demand Models

Evaluation of the forecasting performance of tourism demand models is mainly done on the examination of forecast error magnitudes measured by the Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). These are calculated by measuring the difference between the forecasted value and the actual realisation.

Several factors affect the forecasting performance of tourism demand models such as the forecast horizon, the data frequency and the competitor models. Generally, forecasts become more inaccurate as the forecasting horizon lengthens. SARIMA models fitted to monthly data typically perform well for the purpose of forecasting while causal econometric models forecast annual and quarterly data better. Superior forecasting performance is also attributed to which type of models are being compared; time series models perform better in comparison with static regression models; however, these non-causal time series models are outperformed by more sophisticated causal models, such as VARs, particularly when forecasting annual tourism demand. (Li et al., 2005).

2.2 FORECASTING TOURISM DEMAND WITH GOOGLE TRENDS SEARCH QUERY DATA

2.2.1 Google Trends Data as a Predictor

Construction of the variable representing Google Trends data involves a process of choosing among several search queries that are related to planning a holiday in a specific country and including the selected search terms in the model. Since the forecasting performance of the model depends on the selection and inclusion process of the Google Trends data as an independent variable, some researchers have devised steps for this to be included in an effective way.

Pan et al. (2006) find that online search behaviour related to tourism demand takes place in stages according to the travel planning process. They divided search data into six categories: destinations, hotels, restaurants, transportation, attractions and activities. This categorisation of search queries was adopted by García Rodríguez (2017) in the selection phase of the Google search terms. "Seed queries" are queries associated with these aspects of tourism that are initially entered into Google Trends so that related queries are obtained (Yang et al., 2015). Höpken et al. (2019) enter seed queries in the Keyword Planner available by Google which is a tool that allows businesses to discover keywords that are related to the advertisements placed on Google in order to retrieve related queries.

The second phase of constructing the Google Trends variable is the incorporation of the search query series in the model. According to Li et al. (2017), in cases where there is a small number of relevant search queries, the series is directly added to the model, but multiple search queries can be aggregated into an index by Principal Component Analysis (PCA) or a method called "shift and summation" in order to retain all the information and avoid problems of multicollinearity and overfitting.

PCA is utilised by García Rodríguez (2017) and Wen et al. (2019) to reduce the dimensions of a large dataset while still retaining a faithful representation of the data. Höpken et al. (2019) and Yang et al. (2015) use the shift and summation method to construct search query indices by shifting search query series according to the most appropriate time lag determined by a correlation analysis of the queries at different lags with the arrival series and summing them into a single time series. Li et al. (2017) argue that these methods still result in information loss and not enough attention is given to dynamic correlations in the search query series and thus propose an alternative composite index through a generalised dynamic factor model. Antolini & Grassini (2019), Artola et al. (2015), and Cevik (2020) simply add the series of their chosen query to the model since only a few queries are chosen.

An additional step that some studies adopt is downloading the search series on different days and averaging the data points to obtain a single series (Antolini & Grassini, 2019; Artola et al., 2015 and Cevik, 2020). This is because the way that Google samples the data returns slightly different series for the same search query retrieved on different days as shall be explained in the next chapter.

2.2.2 Evaluation of Forecasting Performance

The literature on assessing whether Google Trends data is able to generate more accurate forecasts of tourism demand follows the methodology outlined in Choi & Varian (2012). This entails the following steps:

1. Selecting a benchmark model for the tourist arrival series
2. Augmenting the benchmark model with the Google Trends data to obtain the Google Trends model
3. Obtaining forecasts of tourist arrivals from both models
4. Evaluating the forecast performance of both models using the RMSE, MAE or MAPE

The models used in the literature are time series models considering the limitation of econometric models when it comes to forecasting, that is, the need to also forecast the exogenous variables. Time series models also allow for a clearer assessment of the contribution of Google Trends search data as a potential predictor of tourist arrivals (Höpken et al., 2019). The benchmark models utilised in the literature are AR, ARIMA or SARIMA models to which the Google Trends data is added as a predictor to obtain the competing Google Trends model which essentially is a SARIMAX model.

Pseudo out-of-sample forecast simulations using the rolling window forecasting approach, as outlined in Stock & Watson (2003) are used to evaluate the short-term forecasting performance of two or more models. The forecasting performance of the models is evaluated by comparing the magnitude of the forecast error from the benchmark model and the competing Google Trends model using the RMSE, MAE and MAPE.

2.2.3 Significance of Google Trends Data in Forecasting and Nowcasting Studies

Findings are mostly in favour of the forecasts generated by the Google Trends-augmented models. This means that models that include the Google Trends search query variable tend to improve forecasts over those generated by the benchmark models which do not include such data (Bangwayo-Skeete & Skeete, 2015; Cevik, 2020; Havranek & Zeynalov, 2021 and Park et al., 2016). Nevertheless, alternative and conflicting conclusions are still reported.

Saidi et al. (2010) conclude that their Google Trends model was not able to outperform the benchmark in forecasting guest nights especially in the short-term. However, series of a different search term pertaining specifically to air travel was able to forecast arrivals more accurately which highlights the sensitivity of the results to the chosen search query data. Conversely, Artola et al. (2015) find that Google Trends data enhanced tourist arrival forecasts only in the short-term.

With regards to the nowcasting studies, Jackman & Naitram (2015) obtain a positive outcome from their Google Trends model. Results from the competing Google Trends-augmented model were not as practical for Antolini and Grassini (2019) as although reduced nowcasting errors were obtained, the model failed to accurately predict arrivals in the peak summer months when more precise predictions are required.

CHAPTER 3
METHODOLOGY & DATA

METHODOLOGY & DATA

Section 3.1 and ***3.2*** of this chapter provide an overview of the underlying theoretical grounds of the methods employed in this dissertation which are motivated by the methodology applied in similar studies as detailed in the literature review. In particular, these sections explain the econometric and statistical workings of forecasting from a time series model and the ARIMA and SARIMA models, respectively. An explanation of PCA is in ***Appendix B.1*** as this requires details from statistical theory rather than econometric theory.

In ***Section 3.3***, the model selection process and the pseudo-out-of-sample one-step-ahead forecasting simulation which are employed in order to answer the research question are described. The robustness checks which entail a statistical test for the equality of the one-step ahead forecasts and another pseudo-out-of-sample forecasting simulation from an additional model are explained also in this section. The data used in the forecasting simulations, which consists of continuous random variables except for the Google Trends data, which is discrete, is described in ***Section 3.4***. The variables employed in the robustness check are described in ***Appendix B.2***.

3.1 FORECASTING THEORETICAL FRAMEWORK

3.1.1 Forecasting from a Time Series Model

Forecasting entails "the prediction of some future event or events" (Montgomery et al., 2015, p. 1). The forecasting problem involves the forecasting of time series data, which consists of a set of observations, y_t, recorded at a particular time t (Montgomery et al., 2015), with the aim of providing an estimate of how the series will behave in the future (Hyndman & Athanasopoulos, 2018). A primary application for which time series econometrics is employed is the development of statistical models which can be used for forecasting, interpreting and testing hypothesis for economic data (Enders, 2014).

The relationship between a dependent random variable y_t and multiple independent random variables $x_{1,t}, x_{2,t}, \ldots, x_{k,t}$ can be modelled using the following multiple linear regression model,

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \varepsilon_t, \tag{3.1}$$

where β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_k$ are the slope coefficients and ε_t is the random error term. This model can be estimated using OLS to obtain the in-sample fitted values of y_t, $\hat{y}_t$, the estimated intercept term $\hat{\beta}_0$ and slope coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$,

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \hat{\beta}_2 x_{2,t} + \ldots + \hat{\beta}_k x_{k,t}. \tag{3.2}$$

Letting $\mathcal{I}_T$ denote all the available and observed information on the independent variables, $x_{1,T}, x_{2,T} \ldots x_{k,T}$, at time T, the h-step ahead forecast for y_{T+h} is given by the expectation of y_{T+h} conditional on $\mathcal{I}_T$,

$$\mathbb{E}(y_{T+h} | \mathcal{I}_T). \tag{3.3}$$

Representing forecasts as the conditional expectation highlights the problem with obtaining forecasts from a linear regression model as in (3.2). In order to obtain (3.3), the following is required

$$\mathbb{E}(y_{T+h}|\mathcal{I}_T) = \mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1,T+h} + \hat{\beta}_2 x_{2,T+h} + \ldots + \hat{\beta}_k x_{k,T+h}\right)$$

$$\mathbb{E}(y_{T+h}|\mathcal{I}_T) = \hat{\beta}_0 + \hat{\beta}_1 \mathbb{E}(x_{1,T+h}) + \hat{\beta}_2 \mathbb{E}(x_{2,T+h}) + \ldots + \hat{\beta}_k \mathbb{E}(x_{k,T+h}). \quad (3.4)$$

The problem is that all $\mathbb{E}(x_{k,T+h})$ are unknown and can only be obtained from another set of models that are able to generate forecasts for all $x_{k,T+h}$. This is the reason why time series models are preferred for the purpose of forecasting over structural econometric models as they circumvent this issue (Brooks, 2008 and Hyndman & Athanasopoulos, 2018).

To illustrate the ease of forecasting future values of y_t from a time series model, an $AR(1)$ is used. This is given by

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t. \quad (3.5)$$

The conditional expectation of y_{T+h} given all the information available at time T is given by $\mathbb{E}_T y_{T+h}$, that is, $\mathbb{E}_T y_{T+h} = \mathbb{E}(y_{T+h}|y_T, y_{T-1}, y_{T-2}, \ldots, \varepsilon_T, \varepsilon_{T-1}, \ldots)$. Therefore, from (3.5), the forecast of y_{T+1} conditional on the information at time T, is obtained by

$$\mathbb{E}_T y_{T+1} = \alpha_0 + \alpha_1 y_T. \quad (3.6)$$

Similarly, y_{T+2} is obtained from

$$\mathbb{E}_T y_{T+2} = \alpha_0 + \alpha_1 \hat{y}_{T+1}$$

$$\mathbb{E}_T y_{T+2} = \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 y_T), \quad (3.7)$$

where $\hat{y}_{T+1}$ is the point forecast of y_{T+1} obtained from (3.6).

In general,

$$\mathbb{E}_T y_{T+h} = \alpha_0(1 + \alpha_1 + \alpha_1^2 + \ldots + \alpha_1^{h-1}) + \alpha_1^h y_T, \tag{3.8}$$

which is referred to as the forecast function. The forecast function generates the forecasts at origin T for h future periods given all the available information on current and past values of y and shows that forecasts of y_{T+h} can be generated using previously forecasted values of y_{T+h-1} (Enders, 2014).

The quality of forecasts deteriorates as they are generated for the longer-term (Enders, 2014), therefore, it is necessary to provide probability limits together with the point forecasts in order to account for the risks related to making decisions based on the provided forecasts (Box et al., 2016).

3.1.2 Forecast Errors and Prediction Intervals

The h-step ahead forecast error, $e_T(h)$, is the difference between the actual value of y_{T+h} and the forecast value, $\mathbb{E}_T y_{T+h}$ given by

$$e_T(h) = y_{T+h} - \mathbb{E}_T y_{T+h}. \tag{3.9}$$

The forecast error is the purely unforecastable proportion of y_{T+h}.[2] For the $AR(1)$ model in (3.5), the h-step ahead forecast error is given by

$$e_T(h) = \varepsilon_{T+h} + \alpha_1 \varepsilon_{T+h-1} + \alpha_1^2 \varepsilon_{T+h-2} + \alpha_1^3 \varepsilon_{T+h-3} + \ldots + \alpha_1^{h-1} \varepsilon_{T+1} \tag{3.10}$$

[2] For the 1-step ahead forecast error, $e_T(1) = y_{T+1} - \mathbb{E}_T y_{T+1} = \varepsilon_{T+1}$.

The forecasts of y_{T+h} are unbiased because the mean of (3.10) is zero, however, they are necessarily inaccurate because of the error variance (Enders, 2014). Assuming that the elements of $\{\varepsilon_T\}$ are independent with variance σ^2, the variance of (3.10) is given by

$$var[e_T(h)] = \sigma^2[1 + \alpha_1^2 + \alpha_1^4 + \alpha_1^6 + \ldots + \alpha_1^{2(h-1)}] \tag{3.11}$$

which shows that it is an increasing function of h, implying that earlier forecasts are more accurate than forecasts for the longer-term.[3]

Assuming also that the elements of $\{\varepsilon_T\}$ are normally distributed, the 95% prediction interval for the forecast of y_{T+h} is given by

$$\hat{y}_{T+h} \pm 1.96\sqrt{var[e_T(h)]}. \tag{3.12}$$

The prediction interval allows for the element of uncertainty in forecasting and is considered as being more important than the point forecast (Hall et al., 2011).[4]

The forecast error arising from the estimated parameters of the model is ignored in (3.10) because the variance in the random error is relatively larger (Verbeek, 2004).[5] Enders (2014) suggests that parsimonious models should be preferred for the purpose of forecasting and the construction of confidence intervals so that the error from the estimated coefficients reflected in the forecasted values of y_{t+h} is minimised. The uncertainty that surrounds the estimated coefficients increases the more complex a model becomes. Therefore, complex models give rise to larger forecast errors coming from the estimated coefficients and thus larger confidence intervals due to errors in the in-sample estimation.

[3] The variance for the one-step ahead forecast error is σ^2.

[4] Similarly, the 90% and 99% prediction intervals are given by $\hat{y}_{T+h} \pm 1.64\sqrt{var[e_T(h)]}$ and $\hat{y}_{T+h} \pm 2.58\sqrt{var[e_T(h)]}$, respectively.

[5] A forecast error is present in the estimated parameters because these are estimators of the true parameters, therefore, there lies another source of error. However, when this error is compared to the variance in the random error term, it is much smaller and therefore, it is ignored.

3.1.3 Evaluation of Forecasting Performance

Hyndman & Koehler (2006) categorise measures of out-of-sample forecasting performance into scale-dependent measures, measures based on percentage errors, measures based on relative errors and relative measures. The relevant classes of measures for the purpose of this dissertation are the first and the second.

Scale dependent measures encompass some of the most commonly used metrics for forecast evaluation in the literature as they are suitable for comparing forecasts generated from different models with the same scale as that of the data (Hyndman & Koehler, 2006 and Montgomery et al., 2015). These include the RMSE and the MAE given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}[e_t(1)]^2} \tag{3.13}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t(1)| \tag{3.14}$$

where $e_t(1)$ is the forecast error for one-step ahead forecasts for multiple time periods t up to n (Enders, 2014 and Hyndman & Koehler, 2008). The RMSE is popular due to its theoretical relevance but it is more sensitive to outliers than the MAE (Hyndman & Koehler, 2006). Therefore, the MAE provides another forecast accuracy measure that is not too affected by irregularities in the data. Another alternative is the MAPE given by

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{100e_t(1)}{y_t} \tag{3.15}$$

which is scale-independent but criticised for penalising positive errors more than negative errors (Hyndman & Koehler, 2006). The MAPE expresses the MAE as a percentage of the

value of the actual observation. The lower the value of the above measures, the higher is the forecast accuracy of the model.

3.1.4 Approaches to Pseudo-Out-of-Sample Forecasts

Forecasts can be one-step ahead or multi-step ahead. One-step ahead forecasts are done using actual data to forecast the period following the last actual observation. On the other hand, multi-step ahead forecasts are forecasts for two or more steps ahead which can be done iteratively by a series of one-step ahead forecasts or directly with a horizon-specific model estimated with the multi-step ahead value to be forecasted as the dependent variable (Marcellino et al., 2006).

Evaluation of the forecasting performance of a model is done by splitting the dataset into a training sample and a testing sample whereby the training sample, or training window, is used to estimate the parameters of the model while the testing sample is used to conduct the evaluation for the out-of-sample ex-post forecasts by means of the measures outlined in the previous section. The forecast exercise is "pseudo-out-of-sample" because the forecasts are generated as if there was no actual data available.

Hyndman & Athanasopoulos (2018) recommend a testing sample that comprises around 20% of the entire dataset and is at least as long as the forecast period, however, this also depends on the length of the time series at hand. The training window in the forecast exercise can be a recursive or a rolling window which are both explained in Appendix B.3.

3.2 THE AUTOREGRESSIVE INTEGRATED MOVING AVERAGE METHODOLOGY

3.2.1 The Box-Jenkins Approach

The Box-Jenkins (1976) approach consists of a three stage procedure for estimating and forecasting from a univariate time series model (Enders, 2014). These are the identification stage, the estimation stage and the diagnostic checking stage (Box et al., 2016).

3.2.1.1 Identification

The identification stage entails using the data and information on the time series to indicate candidate models for the estimation stage (Box et al., 2016). This is primarily done by producing a plot of the time series to detect irregularities in the data and examine the series for stationarity. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) also play a role in this phase as they are indicative of possible orders of autoregression and moving average (Enders, 2014). The purpose of this stage is to generate potential values of p, d and q for the general linear ARIMA model (Box et al., 2016).

Stationary models are a class of stochastic models that have been used to describe a time series process (Box et al., 2016). The assumptions underlying these models are that the series has a zero mean, a constant variance and no correlation with other realisations (Enders, 2014). Most economic time series are not stationary processes and would often need to be transformed, through logarithmic transformation and differencing. Visual inspection of the time series plot will suggest the transformations required in order to achieve a reduced mixed autoregressive-moving average process (Box et al., 2016). The Augmented Dickey-Fuller (ADF) test is most commonly used to formally test for the presence of a unit root, that is, non-stationarity and the Ljung-Box (1978) Q-statistic is used to test for the significance of a group of autocorrelations.[6]

3.2.1.2 Estimation

In the estimation stage, the parameters of the model are estimated under the assumption that efficient use of the data has been made in the previous stage so that diagnostic checking and goodness-of-fit testing in the subsequent stage are made on a well-fitted model (Box et al., 2016). A parsimonious and stationary model with a good fit should be the outcome from this step (Enders, 2014).

OLS estimation of the ARIMA parameters is not possible due to the unobserved values of the lagged error terms for the moving average component. In this regard, Maximum Likelihood Estimation (MLE) is employed. MLE entails knowledge about the conditional distribution of

[6] These are explained in Appendix B.4

the dependent variable except for a finite number of parameters which will be estimated by taking values that give the highest probability of observing the data (Verbeek, 2004).

The probability distribution of the endogenous variable is used as the likelihood function which is maximised with respect to the unknown parameters in order to obtain the maximum likelihood estimators (Verbeek, 2004). The error term ε_t is assumed to be normally distributed which although is a strong assumption, consistent estimators are still obtained with MLE when this assumption does not entirely hold. The aim is to maximise the parameters of the log-likelihood function by some optimising algorithm, such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimisation method[7] (Verbeek, 2004).

3.2.1.3 Model Selection

The goodness-of-fit of a time series model is evaluated on the basis of selection criteria, most commonly the Akaike Information Criterion (AIC) and the Schwarz (or Bayesian) Information Criterion (SIC or BIC). These are given by

$$AIC = T[ln(sum\ squared\ of\ residuals)] + 2n \tag{3.16}$$

$$SIC = T[ln(sum\ squared\ of\ residuals)] + n[ln(T)] \tag{3.17}$$

where n is the number of estimated parameters including the constant term and T is the number of usable observations (Enders, 2014).

A model is said to provide the best fit if it has the smallest AIC or SIC value provided that it is compared to other models estimated on the same sample. For each selection criteria, increasing the parameters to be estimated should decrease the sum squared of residuals, but if an additional regressor does not reduce it, then the AIC and SIC will increase. Both measures are used for model selection, however, the AIC tends to select overparameterized models while the SIC selects the most parsimonious model. If both measures select different models, the residuals of

[7] MLE is a statistically more complex estimation method compared to OLS and entails detailed statistical explanations for a more in-depth review of the method, which is unrelated to the scope of this dissertation.

the model selected by the SIC should be checked to be white noise[8] while the model selected by the AIC should have statistically significant t-statistics for all the coefficients (Enders, 2014).

3.2.1.4 Diagnostic Checking

Once the model has been estimated, it is subject to diagnostic checking. This entails checking the model's adequacy and conducting any necessary improvements and modifications revealed by tests for goodness-of-fit. A crucial procedure is the analysis of the residuals by means of visual inspection and statistical tests for autocorrelation (Box et al., 2016). The Ljung-Box (1978) Q-statistic can also be calculated for the residuals of the final model to check whether they are a white noise process.

3.2.2 The ARIMA Model

Much of the above discussion has been in terms of the ARIMA model, however, prior to the explanation of the workings of this model, it is fundamental to discuss the autoregressive and moving average components separately, and jointly as an Autoregressive Moving Average (ARMA) model.

For a stationary time series y_t, the general $AR(p)$ process where y_t is regressed on its past values, can be written as

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t, \tag{3.18}$$

where δ is the intercept term, p is the lag length of the dependent variable and ε_t is a white noise process.

[8]A white noise process is a series with uncorrelated observations and constant variance. If the series is normally distributed, then it is a Gaussian white noise (Montgomery et al., 2015).

The general $MA(q)$ process for a stationary time series y_t is given by

$$y_t = \varepsilon_t + \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2} + \ldots + \alpha_q\varepsilon_{t-q}, \tag{3.19}$$

where q is the lag length of ε_t.

Equations (3.18) and (3.19) can be combined to form the general $ARMA(p,q)$ model. Using the lag operator notation where, $L^p y_t \equiv y_{t-p}$ and $L^q\varepsilon_t \equiv \varepsilon_{t-q}$, the $ARMA(p,q)$ model is given by

$$\phi(L^p)y_t = \alpha(L^q)\varepsilon_t + \delta. \tag{3.20}$$

If y_t is a non-stationary series that requires differencing in order to be stationary, the appropriate model to represent this process is an $ARIMA(p,d,q)$ model where d represents the order of differencing required to the make the series stationary and hence *integrated*. The $ARIMA(p,d,q)$ model is given by

$$\phi(L^p)\nabla^d y_t = \alpha(L^q)\varepsilon_t + \delta, \tag{3.21}$$

where $\nabla = 1 - L$ is the difference operator and $d \geq 1$.

The choice to represent a series by an $ARIMA(p,d,q)$ model is a matter of parsimony in order to avoid long lags of an $AR(p)$ or $MA(q)$ process (Verbeek, 2004). The $ARIMA(p,d,q)$ model provides a neat way to combine the two for a nonstationary series y_t.

3.2.3 The Seasonal ARIMA Model

An extension to the $ARIMA(p,d,q)$ model in (3.21) is made to incorporate seasonality. Seasonal data requires attention because of relationships found between observations for

consecutive months in a year and also between observations of the same month in consecutive years (Box et al., 2016). The tourist arrival time series reveals a clear seasonal pattern and it would be incorrect to omit this from the modelling process. Seasonal data can be modelled additively by adding the autoregressive or moving average term at the seasonal lag to the model or multiplicatively since seasonal and non-seasonal patterns interact with each other in the ACF and PACF (Enders, 2014).

The general multiplicative seasonal $ARIMA(p,d,q)(P,D,Q)_s$ is shown by

$$\phi_p(L)\Phi_P(L^s)\nabla^d\nabla_s^D y_t = \alpha_q(L)\mathrm{A}_Q(L^s)\varepsilon_t + \delta \tag{3.22}$$

where P, D and Q are the orders of autoregression, differencing and moving average for the seasonal component, respectively, L^s is the lag operator for the seasonal component and $s = 12$ for monthly data (Box et al., 2016). There is no theoretical basis as to whether seasonality should be incorporated additively or multiplicatively (Enders, 2014), however, multiplicative models are the most common and are utilised by Box et al. (2016) for their famous airline model of monthly passenger totals represented by a multiplicative SARIMA $(0,1,1) \times (0,1,1)_{12}$ model. The Box-Jenkins methodology outlined in Section 3.2.1 applies to the SARIMA model but with consideration to the seasonal element of the series (Box et al., 2016).

3.3 PROCEDURES IMPLEMENTED FOR ASSESSING FORECASTING PERFORMANCE

3.3.1 Model Selection

The modelling procedure adheres to the Box-Jenkins (1976) methodology described in Section 3.2.1. The benchmark model is a SARIMA model and the competing Google Trends model consists of the benchmark SARIMA model with variables representing Google Trends data included by the principal components from the PCA on the data.

Selection of the SARIMA model is done using the *Arima* function in R from the *forecast* package (Hyndman, 2020) which generates a list of possible models given specified orders of p, d, q, P, D and Q together with the AIC and SIC value for each model. Table 3.1 shows the minimum and maximum orders of these terms that were specified in the function. Given these specifications, there were 1,352 SARIMA model combinations.[9]

Table 3.1: Specification of Seasonal and Non-Seasonal Terms.[10]

	Minimum	**Maximum**
Non-Seasonal Autoregressive Terms, p	0	12
Non-Seasonal Differencing, d	1	1
Non-Seasonal Moving Average Terms, q	0	12
Seasonal Autoregressive Terms, P	0	1
Seasonal Differencing, D	0	1
Seasonal Moving Average Terms, Q	0	1

The models with the lowest AIC and SIC from the 1,352 possible models are identified to be subject to diagnostic checking for confirmation of goodness-of-fit according to other criteria outlined in Section 3.2.1.4.

[9] The number of possible model combinations is calculated by multiplying the number of different values P, D, Q, p, d and q can take. Therefore, including the possibility for the autoregressive, moving average and differencing orders to be 0 (except for the non-seasonal differencing order), 1,352 is obtained from $13 \times 2 \times 13 \times 2 \times 1 \times 2$.

[10] Since the tourist arrivals series is indicative of a single seasonal cycle consisting of twelve months as shown later in Figure 3.1, the maximum order for the seasonal components was specified to be one. On the other hand, the non-seasonal terms were allowed to take a maximum of twelve since the data is of a monthly frequency, except for the differencing term that was specified to be one as informed by the ADF test on the sample of data from January 2004 to December 2016 shown in Table C.3.2 in Appendix C.3.

3.3.2 Pseudo-Out-of-Sample Forecasting Simulation

Diagram 3.1 shows that the training sample on which the models are estimated on consist of data from January 2004 to December 2016 (black solid line) while the testing sample is from January 2017 to December 2019 (green solid line). This means that forecasts are made for January 2017 until December 2019 and then compared to the actual observation on the basis of the RMSE, MAE and MAPE.

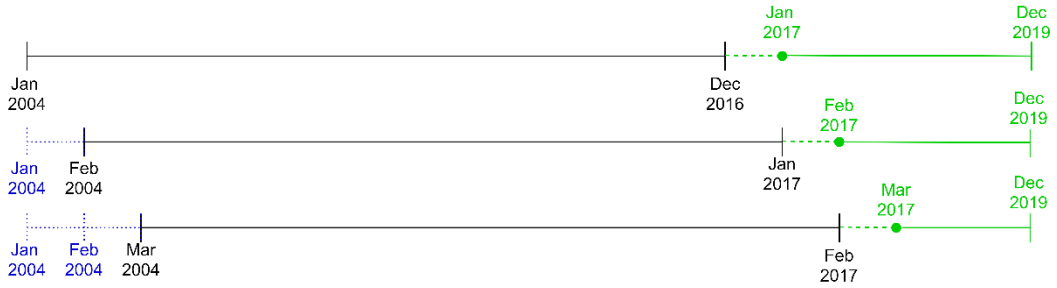
Diagram 3.1: Training and Testing Sample.

Source: Author's illustration.

The assessment of improvements made to the forecasts of tourist arrivals by Google Trends data is made on a pseudo-out-of-sample simulation of one-step ahead forecasts applying a rolling window approach. This means that each forecast is generated from a model with different coefficient values as it is estimated on a different sample of observations. Therefore, in order to generate the forecasts for the 36 months in the testing sample, 36 benchmark models and 36 Google Trends models are estimated as the estimation sample is rolled forward each time.

Diagram 3.2 illustrates how the one-step ahead forecast simulation is carried out using the rolling window approach on the sample of the data used in this dissertation. It shows that more recent observations are used to generate the one-step ahead forecasts as the estimation window is rolled forward which is suitable in the context of forecasting tourist arrivals as future values depend on recent historic values

Diagram 3.2: Rolling Window Approach for One-Step Ahead Forecasts.



Source: Author's illustration.

3.3.3 Robustness Checks

3.3.3.1 Clark & West (2007) Test

Diebold (2015) states that pseudo-out-of-sample evaluations are not enough "insurance" for claiming better forecasts from a model, therefore the robustness of the one-step ahead forecasts are checked via a statistical test to verify that the findings were not merely by chance.

Clark & West (2007) propose a procedure to test whether the forecasted values are actually generated from different models by controlling for uncertainty of the parameters when the models are nested[11]. The foundations of this process lie in the conclusions made in Clark & West (2006) whereby the Mean Squared Prediction Error (MSPE) of the nested model should be smaller than that of the competing model since the additional parameters of the competing model introduce noise into the forecasts resulting into greater errors. With this reasoning, the process described in Clark & West (2007) entails adjusting the point estimate of the difference between the MSPEs of the models for the noise attributed to forecasts from the larger model.

In this procedure, the null hypothesis is that the MSPEs are equal, thus representing equal forecasting accuracy, and the alternative hypothesis is that the larger model has a smaller MSPE

[11] Models are nested when one can be obtained from the other. The benchmark model used in this dissertation is nested within the Google Trends models because the benchmark model can be obtained from the Google Trends model by setting the coefficients of the Google Trends variables to zero.

and therefore the forecasts are generated from this model. The null hypothesis is tested by defining a series z_i,

$$z_i = (e_{1i})^2 - [(e_{2i})^2 - (f_{1i} - f_{2i})^2] \tag{3.23}$$

where e_{1i} and e_{2i} are the forecasts errors from model 1 (the nested benchmark model) and model 2 (the Google Trends model), respectively, and f_{1i} and f_{2i} are the forecasts from the same models for $i = 1, \dots, H$ number of forecasts. z_i is regressed on a constant and the null hypothesis is rejected if the resultant t-statistic for a zero coefficient of the intercept is greater than 1.645 for a one-sided test at the 5% significance level (Enders, 2008).[12]

3.3.3.2 Pseudo-Out-of-Sample Forecasting Simulation from a Model with GDP and REER

An additional robustness check following that carried out by Havranek & Zeynalov (2021) is applied involving another pseudo-out-of-sample one-step ahead forecast simulation of a model with explanatory variables of tourist arrivals as delineated by the literature. The purpose of this robustness check is to evaluate the performance of the model with Google Trends data against that of this model which encompasses variables in line with economic theory.

The literature review on tourism demand models infers that income and price variables are the factors with the greatest effect on the number of tourist arrivals in a country. Data for GDP and the Real Effective Exchange Rate (REER) is used to represent income of the sending countries and price of touring, respectively. These variables are added to the SARIMA benchmark model used in the previous forecast simulation and one-step ahead forecasts are produced which are evaluated on the RMSE, MAE and MAPE. The data used in this robustness check is described in Appendix B.2.

[12] Clark & West (2007) state that autocorrelation consistent errors should be used for serially correlated forecasts errors.

3.4 THE DATASET

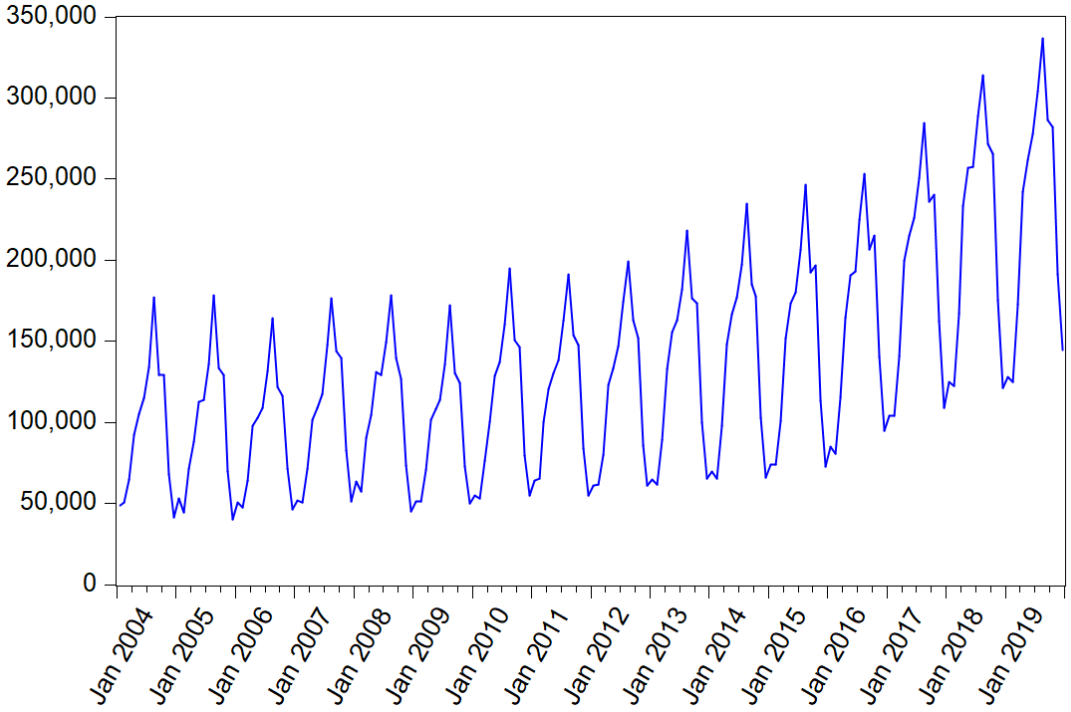
3.4.1 The Target Variable: Tourist Arrivals

The dependent variable in this study is the number of tourist arrivals which is also the target variable of the forecast problem since it is the variable being forecasted. Data for tourist arrivals in Malta from January 2004 to December 2019 was sourced from the National Statistics Office (NSO). [13] The sample starts from January 2004 because Google Trends data starts from this period and it ends in December 2019 because of the structural break in the series in the second quarter of 2020 due to travel restrictions related to the onset of the Covid-19 pandemic which could have biased the results.

In line with the Box-Jenkins (1976) methodology, a plot of the tourist arrivals series from January 2004 to December 2019 is produced in Figure 3.1 in order to visually examine the data. During this period, the time series exhibits a distinct seasonal pattern with peaks and troughs every twelve months indicating the seasonal cycle. This seasonal pattern justifies the use of a SARIMA model and is suggestive of one seasonal cycle. The series also shows an upward trend from 2013 onwards which may present a problem to the stationary properties of the series. These observations on the properties of the series were used to inform the maximum orders of p, d, q, P, D and Q in Table 3.1 for the model selection process.

[13] The NSO defines inbound tourism as "the activities of non-resident visitors travelling to Malta (i.e., outside their usual environment) and staying for not more than twelve consecutive months for personal, business or other purposes" and a tourist is "a visitor who stays at least one night in a rented and non-rented accommodation in the place/country visited". Clarification of these terms is important in order to understand what is being studied and targeted as incorrectly referring to the number of tourist arrivals as visitors to Malta also includes the number of overnight cruise passengers which is not the case in this dissertation.

Figure 3.1: Monthly Tourist Arrivals (Jan 2004 – Dec 2019).



Source: Author's illustration.

3.4.2 The Predictor Variable: Google Trends Data

3.4.2.1 Description

Google Trends is a product made available by Google which provides a sample of data that shows the interest of users in a particular subject or topic. Data samples are available as early as the hour before inquiring on Google Trends or from 2004 until approximately 36 hours[14] before inquiring. Billions of search queries are made to Google every day; therefore, the samples are adequately representative of all searches. The data is normalised to the time and geographical location of the query for ease of comparison. Normalisation is done by dividing each data point by the total searches of the location and time range and scaling the resultant

[14] The last data point is *incomplete* as it is not necessarily representative of the search made during the month when the inquiry is made as the month might have not yet ended.

numbers from 0 to 100, hence why the data is discrete, depending on the proportion of searches on the topic relative to all searches made on all topics (Google, n.d.).

The Google Trends website offers a number of features to facilitate data exploration which are used in the collection of data for this dissertation. Filters are available for the geographical location, time range, category and type of searches[15] for the search query in question. The interest over time shows a plot of the time series for the specified query and the interest by region shows the location where the search term was most popular.[16] Up to 25 queries related to a query entered are provided which can be filtered by top and rising; the difference between the two is that the former sorts queries by popularity based on their value and the latter sorts queries that saw the largest increase in searches since the last period of time (Google Trends, 2022).

3.4.2.2 Collection

The collection process of the Google Trends search data follows closely the steps outlined by Yang et al. (2015). The process encompasses four stages wherein the first two are concerned with the collection of the data; the first step entails entering *seed queries* which are broad and general search terms associated with travelling to the place in question while the second stage involves retrieving related queries (from the related queries feature in Google Trends) from these initial queries in order to expand the query library.

The third step is to calculate the Pearson correlation coefficient between each Google Trends search query series and the tourist arrival series at different lags. The rationale for finding the correlation at different lags of the Google Trends search query data is that people search for certain information about their potential trip during certain months prior to their actual arrival in the country. The final stage consists of removing queries with a correlation coefficient lower than a certain threshold and lagging the series according to the lag number that returned the

[15] Type of searches includes web search, image search, news search, Google shopping and Youtube search.

[16] Google Trends states that a region with a high value represents a higher proportion of all queries not a higher absolute number of searches made for the query: “a tiny country where 80% of the queries are for "bananas" will get twice the score of a giant country where only 40% of the queries are for "bananas" (Google Trends, n.d.). This is important to specify as Malta’s size may result in this distortion in the interest by region. However, this does not affect the quality of the data used in this dissertation.

highest correlation with the arrival series. As in García Rodríguez (2017), PCA is then performed after the completion of these steps in order to obtain the Google Trends variable to be included in the Google Trends model. The results from the third and fourth steps are presented in the next chapter.

This process is applied by García Rodríguez (2017) in the context of Mallorca as the subject destination. Since Mallorca and Malta share similar country characteristics, the seed queries entered by the author are applied in context of the Maltese Islands. Table 3.2 shows the eleven seed queries initially entered based on the following aspects related to trip planning identified by Pan et al. (2006): activities, attractions, destination, hotels, restaurants and transportation. The geographical location, time range, category and type of search filters in Google Trends are set as shown in Table 3.3.

Table 3.2: 11 Seed Queries Entered into Google Trends Based on 6 Aspects of Trip Planning.

Activities	Attractions	Destination	Hotels	Restaurants	Transportation
Malta Events	Malta Party	Malta Tourism	Malta Hotel	Malta Restaurant	Malta Flights
Malta Beach		Malta Travel			Malta Airport
Malta Weather					
Malta Activities					

Table 3.3: Specification of Geographical Location, Time Range, Category and Type of Search Filters in Google Trends.

Geographical Location	Worldwide
Time Range	01/01/2004 – 31/12/2019
Category	All categories
Type of Search	Web Search

The related queries feature in Google Trends was used to retrieve queries related to the eleven queries above using both the *top* and *rising* filter described in Section 3.4.2.1. The related queries obtained from the *rising* filter are shown in Table B.5.1 and those obtained from the *top* filter are shown in Table B.5.2 in Appendix B.5 which amount to the 517 related queries obtained in total. The next section explains the steps taken to reduce the number of queries.

3.4.2.3 Reduction

The process of cleaning the large Google Trends search query library obtained from the previous step which consisted of 517 queries is outlined below. Reducing the number of queries is a crucial step as not all are necessarily related to tourism which might distort the forecast path. Furthermore, the compression of the dataset was done for ease of the PCA as otherwise, not all query data would have been relatively equally represented.

Prior to the application of the third step in the process of query collection and reduction as outlined above, examination of the 517 search queries reveals that not all relate to tourism in Malta and that some are specific searches which did not contain enough data points to bear any predictive power. Therefore, queries with the following characteristics are eliminated:

- Queries that do not contain the word *Malta*, *Gozo*, *Comino* or a place in Malta or Gozo
- Queries related to Covid-19
- Queries completely unrelated to travelling or tourism
- Queries that refer to particular places or companies
- Queries that are very specific (time, weather, etc.)

Tables B.6.1 and B.6.2 in Appendix B.6 show the queries eliminated according to these characteristics after which 270 remain including the eleven seed queries. Duplicate queries which featured in the related queries obtained from both the *top* and *rising* filters were also removed including those which are the same but with a different word order as Google Trends recognises these as being identical. Following this, the number of queries is 128.

The data for these 128 queries was downloaded with the use of the *gtrendsR* package (Massicotte & Eddelbuettel, 2022) in R in which the criteria in Table 3.3 was entered and the 128 queries were specified. After the data for the queries was downloaded and visually examined, those queries that contained at least one zero-value data point were removed in order to retain queries that hold the most information for prediction purposes The number of queries after this step was 59. This process up to this step is summarised in Figure 3.2.

Figure 3.2: Summary of Google Trends Data Collection and Reduction Process (left pane) and Number of Queries after each Step (right pane).

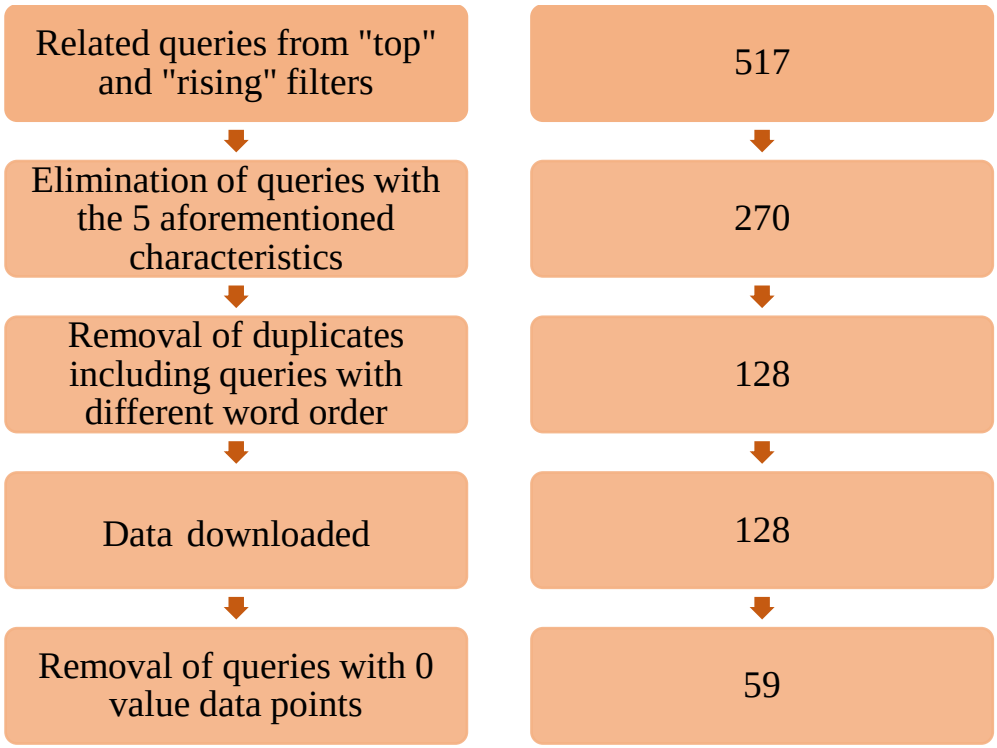


Figure 3.3 shows the time series plots for the resultant 59 search queries. The majority of the series exhibit the same seasonal pattern as the arrivals series in Figure 3.1 thereby justifying the relevance of the research question in exploring any potential predictive power of such data. It is also clear that some series lack the distinct seasonal pattern shown by others. Examples of these are “malta independent”, “times of malta”, “malta today” and “malta park” of which the first three may be referring to online local news portals and the last relates to the local online marketplace.

Figure 3.3: Google Trends Search Query Data (Jan 2004 – Dec 2019).

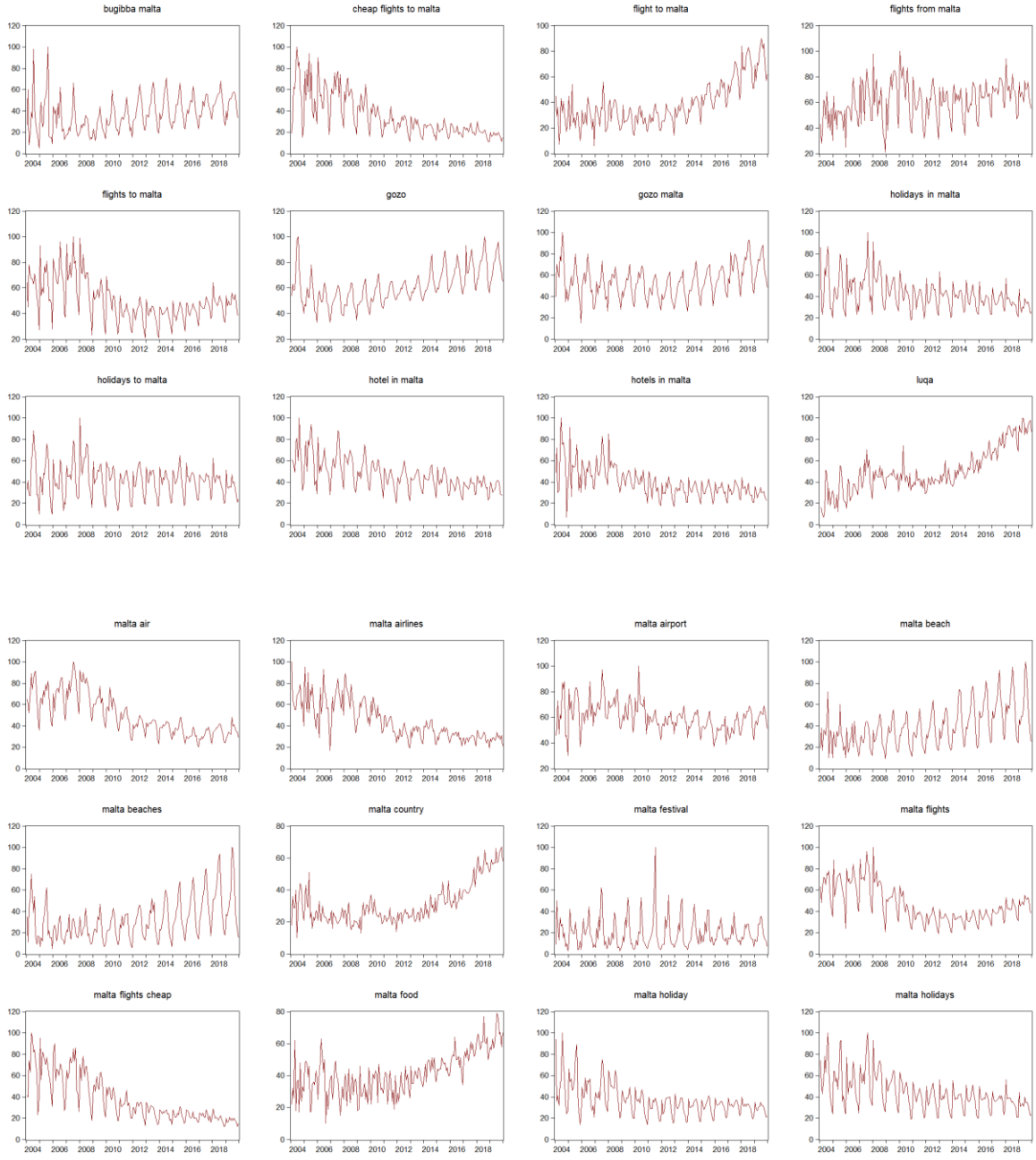


Figure 3.3: Google Trends Search Query Data (Jan 2004 – Dec 2019).

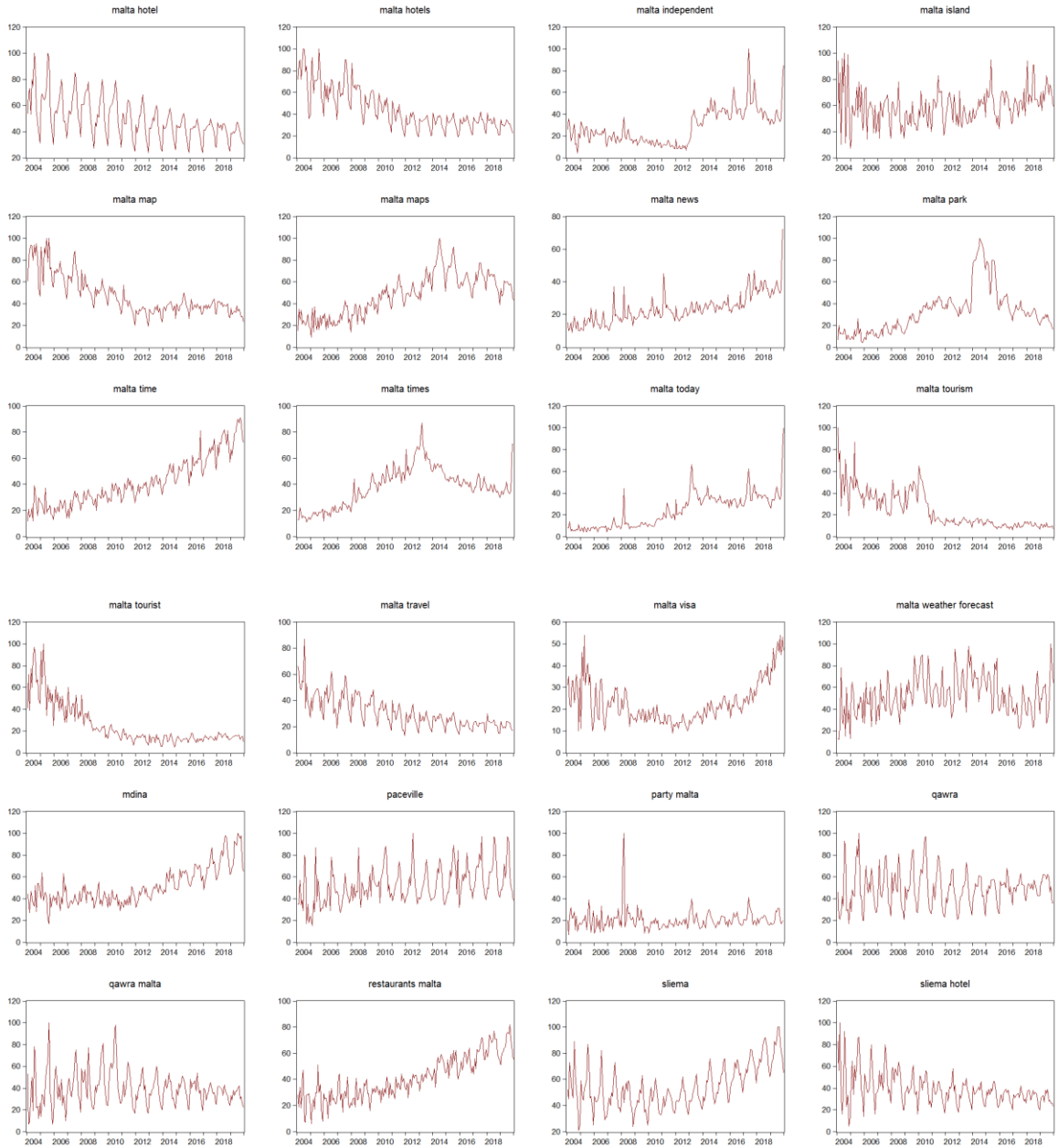
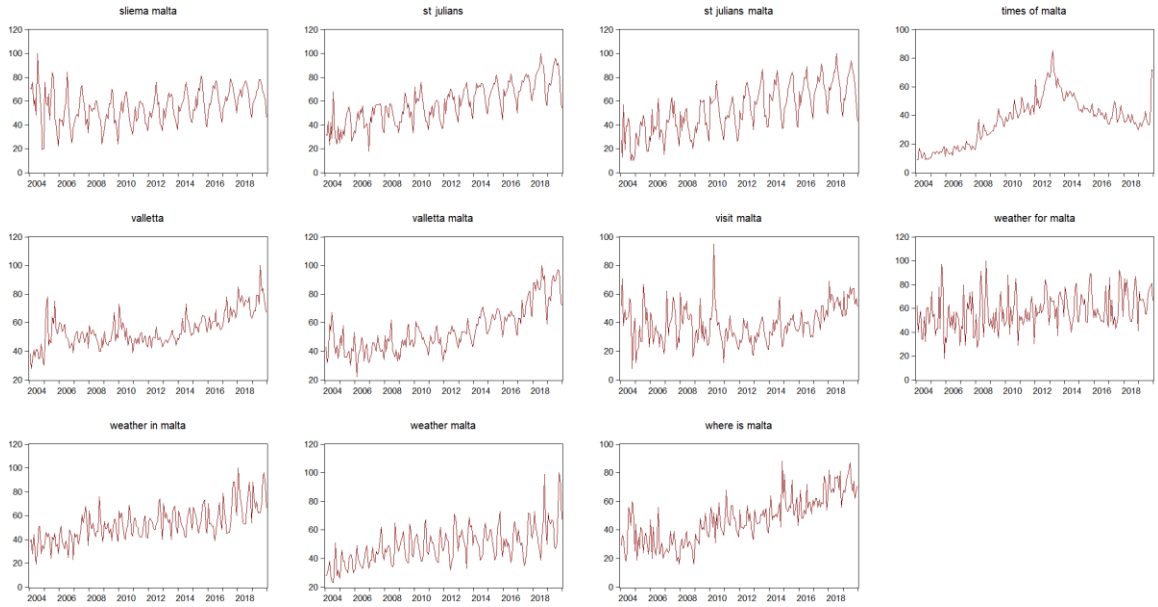


Figure 3.3: Google Trends Search Query Data (Jan 2004 – Dec 2019).



Source: Author's illustration.

CHAPTER 4
RESULTS & DISCUSSION

RESULTS & DISCUSSION

The results from the methods described in the previous chapter and applied in order to provide the answer to the research question of this dissertation are presented in this chapter. In ***Section 4.1*** the outcomes of the correlation analysis and principal component analysis conducted on the Google Trends data are discussed. The selection of the benchmark model and the Google Trends model together with an econometric evaluation is outlined in ***Section 4.2***. The results from the one-step ahead forecasting simulations which are the key findings that show whether Google Trends data provides better forecasts of tourist arrivals are presented in ***Section 4.3***. ***Section 4.4*** shows the outcomes from the two robustness checks carried out on the results obtained in the preceding section. The findings in Section 4.3 and 4.4 allowed for a nowcasting experiment which is in ***Section 4.5*** of this chapter and depicts a practical application of the findings of this dissertation.

Section 4.6 is the part of this chapter related to the discussion of the results. A summary of the results obtained is provided together with an evaluation of the nowcasts from the nowcasting experiment. The findings of this dissertation are discussed in terms of their policy implications and they are also compared with the findings of other studies mentioned in the literature review chapter. ***Appendix C*** contains supplementary material that is referenced to throughout this chapter.

4.1 GOOGLE TRENDS DATA

4.1.1 Correlation with Arrivals Series

In line with the search query collection process adopted by Yang et al. (2015) outlined in the previous chapter, the Pearson correlation coefficient between the tourist arrivals series and the 59 search queries with lags up to seven was calculated, shown in Table C.1.1 in Appendix C.1. As highlighted in Figure 3.3 in the previous chapter, not all queries exhibit the same seasonal pattern as the arrivals series, therefore the correlation analysis is an aid for identifying which query series are potentially related to tourism in Malta albeit a high *correlation* may not necessarily be *causing* the variation in tourist arrivals.

The results from the correlation analysis in Appendix C.1 show that several queries related to accommodation, transportation and holidays in Malta are correlated the most with the arrival series at a four- or five-month lag. This means that tourists search for their flights and accommodation four to five months prior to their stay. Similarly, queries related to itinerary planning, such as "mdina", "paceville", "gozo" and "restaurant malta", are correlated the highest with the arrivals series at the one-month lag. This shows that tourists plan their activities one month before coming to Malta.

These results also reveal that queries without a distinct seasonal pattern have a correlation below the average of the highest absolute correlation that is, less than 0.31. In view of this, the Google Trends dataset was further reduced by eliminating queries with a highest absolute correlation value below 0.30. Elimination of queries under this threshold means that the final number of queries to be used is 33 listed in Appendix C.2.[17]

The correlation results were used to create three separate Google Trends datasets according to the lag of the highest correlation. This means that the series of the queries with the highest correlation at the 5^{th}, 4^{th} and 1^{st} lag were lagged accordingly and grouped according to their lag number. Therefore, the three datasets consisted of a dataset with queries lagged by five periods,

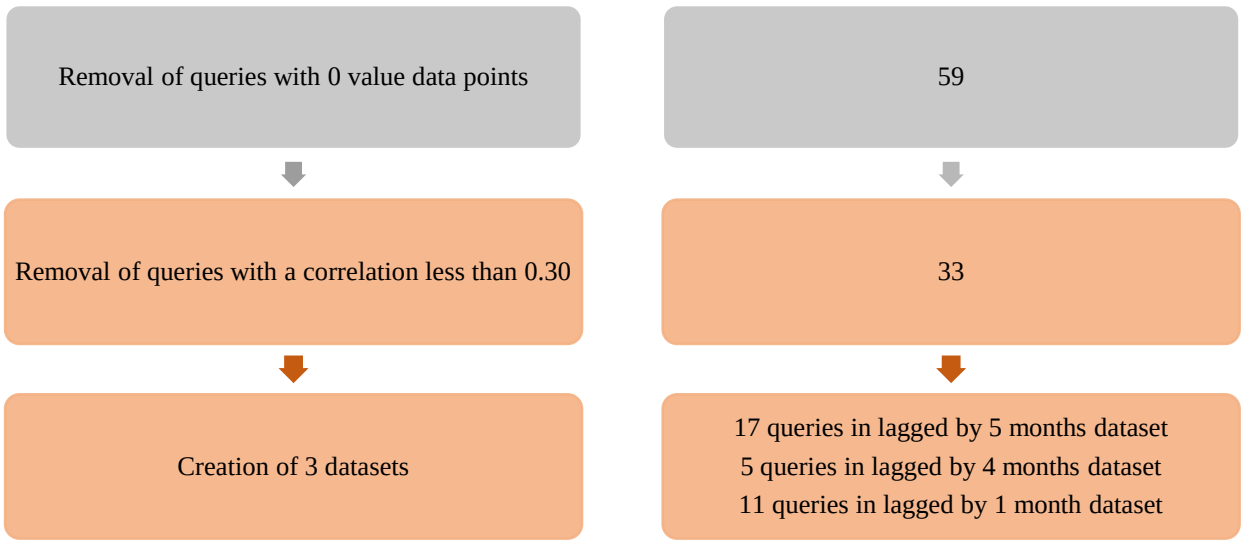
[17] The number of queries that resulted after the elimination of queries under the 0.30 highest absolute correlation threshold was in fact 34. However, only one query was left that was correlated the highest with the arrivals series at the two month lag. This query was also eliminated as it could not be included in the PCA or in the model as another explanatory variable as it would be fully represented as opposed to the other queries that are represented according to their loading value in the principal component.

another dataset with queries lagged by four periods and a third with queries lagged by one period.

The rationale for this step is because the PCA conducted on all 33 variables generated low loading values and therefore, not all queries were being sufficiently captured. Moreover, since the queries have different lag relations to the arrival series, PCA on one dataset with all the queries would not be considering these relations. A solution to this problem could have been to first, alter the query series with the appropriate lag and then, conduct the PCA on a single dataset with all query series. However, the initial issue of unrepresentativeness of all queries in the principal components would still be present. As a result, the creation of three datasets overcame the initial problem while also considering the lag relation.

Figure 4.1 is a continuation of Figure 3.2 in the previous chapter which summarises the steps performed above on the Google Trends dataset that resulted into the ultimate number of keywords that shall be used in the Google Trends model.

Figure 4.1: Summary of Google Trends Data Collection and Reduction Process (Left Pane) and Number of Queries after each Step (Right Pane).

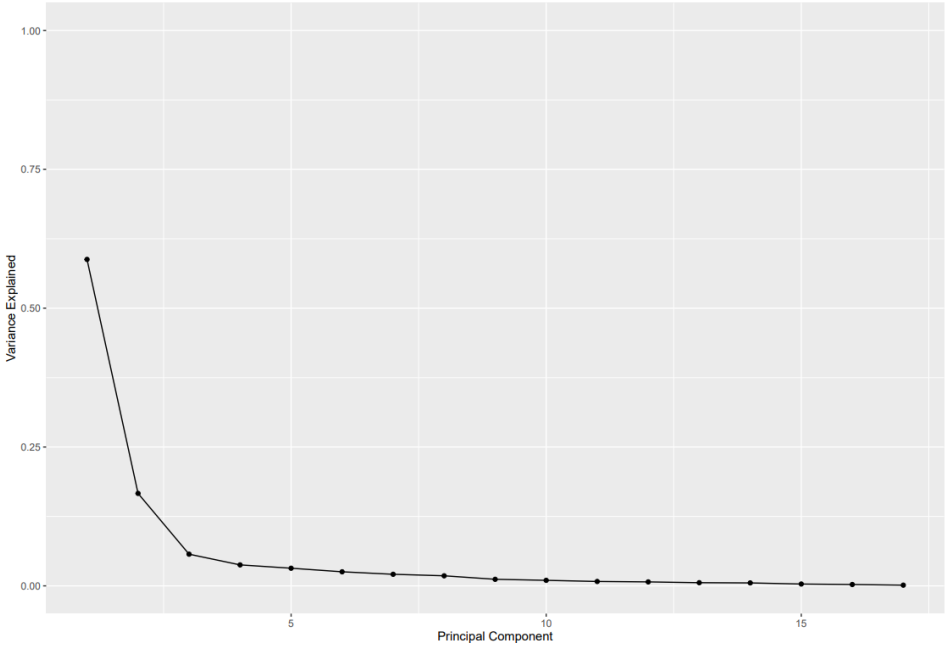


4.1.2 Principal Component Analysis

The resulting proportion of variance explained by the principal components from each PCA are shown in Figures 4.2, 4.3 and 4.4. The cumulative variance explained by the principal components is equal to 1, that is, 100% of the variance of the data. The length of the x-axis pertains to the number of principal components from each dataset; for instance, the dataset that is lagged by five periods contains seventeen variables, therefore the PCA produces seventeen principal components.

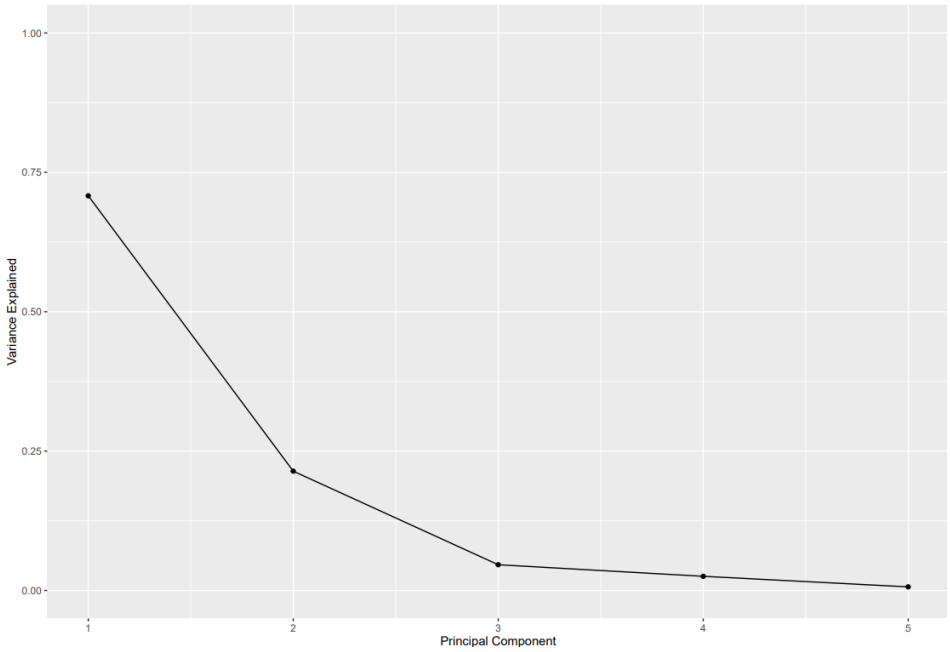
The scree plots depict that the first principal component of each dataset explains the most variance in the data which decreases monotonically with each principal component as explained in Appendix B.1. Figure 4.2 shows that the first principal component of the PCA conducted on the dataset lagged by five months explains around 59% of the variance in the data and the first principal components from the PCA conducted on the dataset lagged by four months and one month both explain around 71% of the variance of the data as shown in Figures 4.3 and 4.4, respectively.

Figure 4.2: Scree Plot of 17 Principal Components from the PCA on Dataset Lagged by 5 Months.



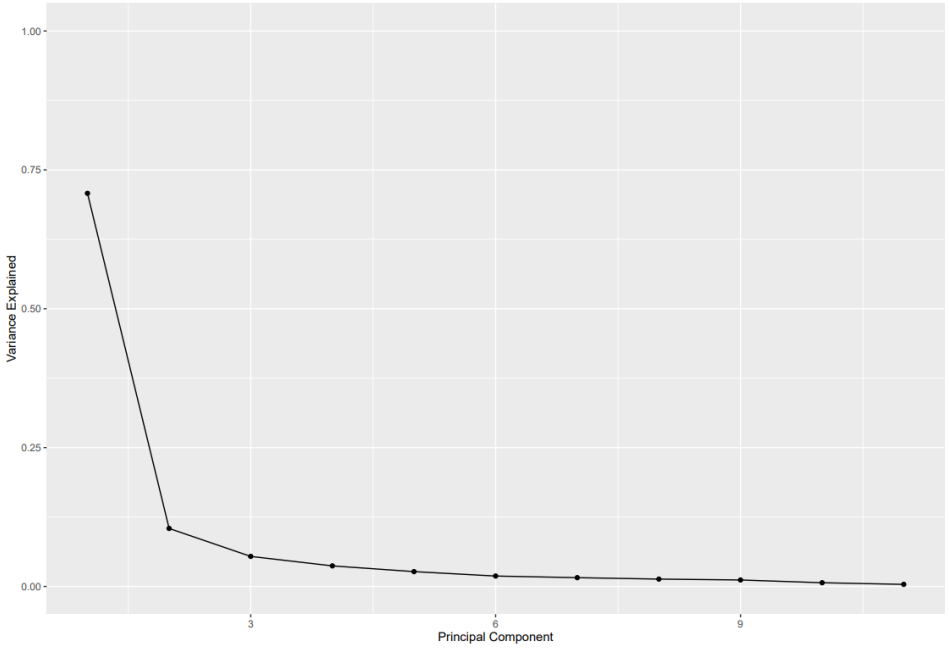
Source: Author's calculations.

Figure 4.3: Scree Plot of 5 Principal Components from the PCA on Dataset Lagged by 4 Months.



Source: Author's calculations.

Figure 4.4: Scree Plot of 11 Principal Components from the PCA on Dataset Lagged by 1 Month.



Source: Author's calculations.

While there is no definite rule that dictates how many principal components should be considered (Rachev et al., 2007), the variance of the data explained by each first principal component is large enough to represent the Google Trends data in the model competing against the benchmark model. Therefore, from the results of the PCA, the Google Trends model shall include three explanatory variables which are the first principal component from each PCA on the three datasets.

Tables 4.1, 4.2 and 4.3 show the loadings of the first principal component from each dataset in order to highlight what each first principal component is actually representing. As explained in Appendix B.1, the loadings are the eigenvectors of the variance-covariance matrix of each dataset. The value of the loadings reflects the correlation of each query with the principal component, whether positive or negative. The larger the absolute value of the loading, the stronger is the effect of that variable in the principal component.

The loadings of the first principal component in Table 4.1 show that the hotel, flights and holiday queries are the largest contributors compared to the rest of the queries. Therefore, this principal component is the variable representing the very initial stage of travel planning. The queries of the four- and one-month lagged dataset are approximately equally represented in the first principal component as implied by the loading value as shown in Tables 4.2 and 4.3.

Table 4.1: Loadings of the First Principal Component for the Dataset Lagged by 5 Months.

Query	Loading
bugibba malta	0.15
flights to malta	0.27
gozo malta	0.19
holidays to malta	0.26
hotel in malta	0.29
hotels in malta	0.27
malta air	0.25
malta festival	0.16
malta holiday	0.29
malta holidays	0.30
malta hotel	0.30
malta hotels	0.28
malta map	0.24
malta weather forecast	-0.10
qawra	0.21
sliema hotel	0.26
sliema malta	0.18

Table 4.2: Loadings of the First Principal Component for the Dataset Lagged by 4 Months.

Query	Loading
cheap flights to malta	0.47
malta flights cheap	0.49
malta travel	0.47
weather in malta	-0.38
weather malta	-0.42

Table 4.3: Loadings of the First Principal Component for the Dataset Lagged by 1 Month.

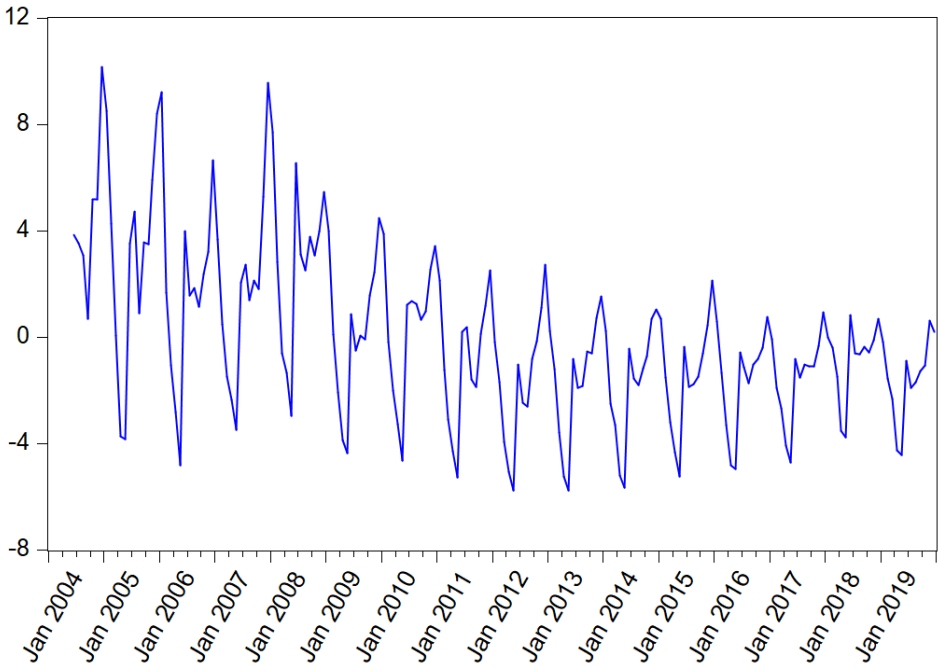
Query	Loading
flight to malta	-0.32
flights from malta	-0.19
luqa	-0.32
malta news	-0.29
malta time	-0.34
mdina	-0.31
paceville	-0.23
restaurants malta	-0.31
st julians	-0.33
st julians malta	-0.32
valletta malta	-0.33

.

The first principal component from each dataset is plotted in Figures 4.5, 4.6 and 4.7. These variables exhibit a seasonal pattern in line with the arrival series implying also that the seasonal element of the search query series as depicted in Figure 3.3 in the previous chapter, has been preserved. The plots are also suggestive of a stationarity issue due to the presence of a trend and a non-constant mean. The stationary properties of the principal components are discussed in the next section.[18]

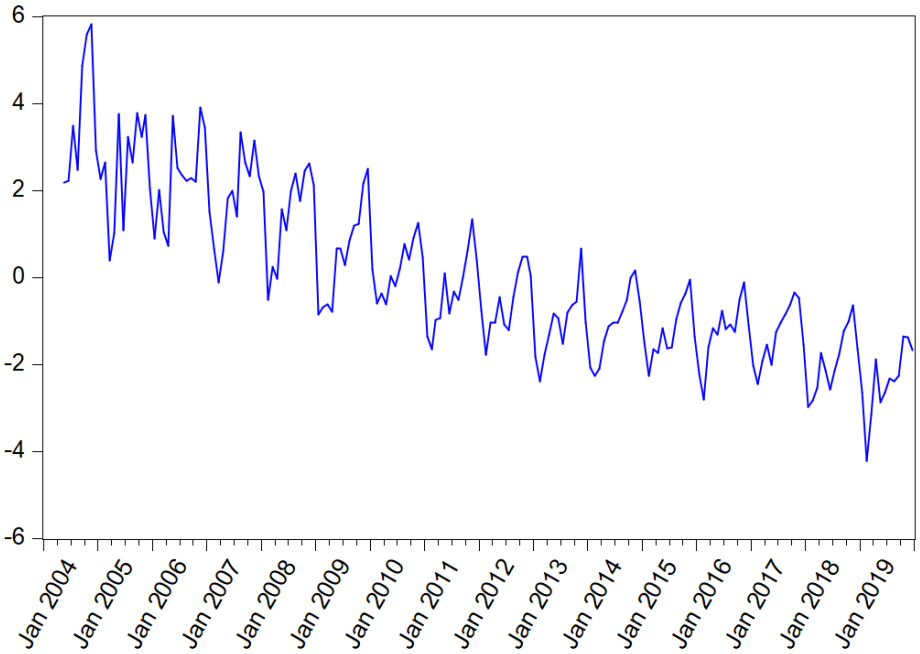
[18] The Google Trends series were not transformed into stationary processes before the PCA but the principal components were tested for stationarity prior to including them in the Google Trends model. This reasoning follows from García Rodríguez (2017) and also because the Google Trends data exhibited different orders of integration which would result into different transformations. Hamilton & Xi (2022) propose a procedure for stationary and nonstationary variables undergoing PCA, however, being very recent research, this has not been applied in other studies and was thus not applied in this dissertation.

Figure 4.5: Time Series Plot of the First Principal Component of the Lagged by 5 Months Dataset.



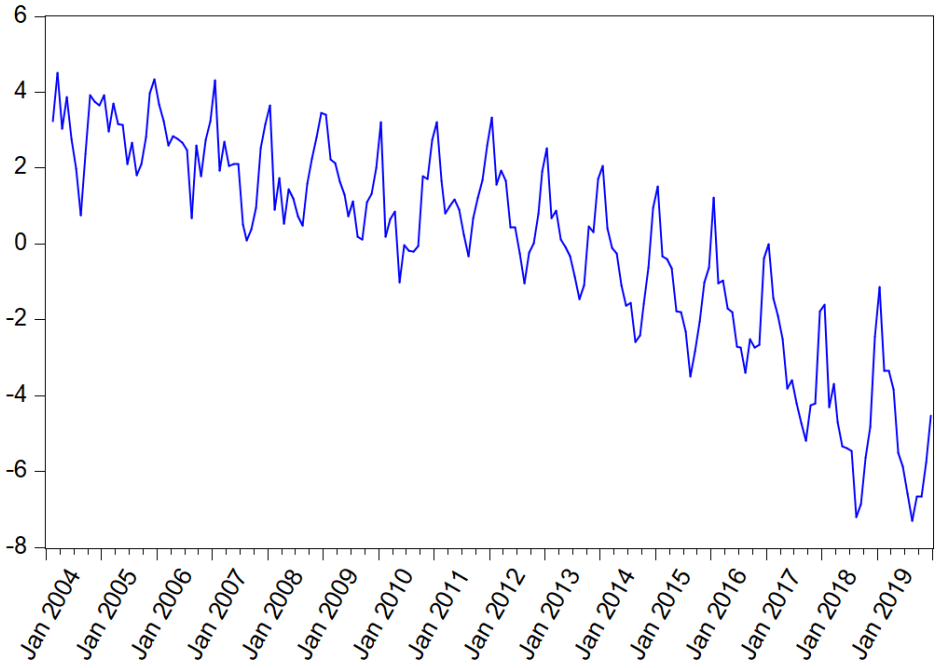
Author's calculations.

Figure 4.6: Time series Plot of the First Principal Component of the Lagged by 4 Months Dataset.



Source: Author's calculations.

Figure 4.7: Time Series Plot of the First Principal Component of the Lagged by 1 Month Dataset.



Source: Author's calculations.

4.2 MODEL SELECTION AND EVALUATION

4.2.1 Stationarity

Prior to the model selection process, the arrivals series and the principal components series were tested for stationarity using the ADF test on the training sample, that is, from January 2004 to December 2016.

Panels A of Table C.3.1 and C.3.2 in Appendix C.3 show that the arrivals series is not stationary in levels and in first differences, however it is stationary in first difference when log transformed as shown in Panel B of Table C.3.2. The results in Panels A of Tables C.3.3, C.3.4 and C.3.5 in Appendix C.3 also show that the first principal components are not stationary in levels, but Panels B of the same tables show that they are stationary in first difference.

Therefore, the benchmark model and the Google Trends model are modelled and estimated with the dependent variable first differenced in log form and the principal components in the

Google Trends model are also in first differences. These are not log transformed because of the presence of negative values.

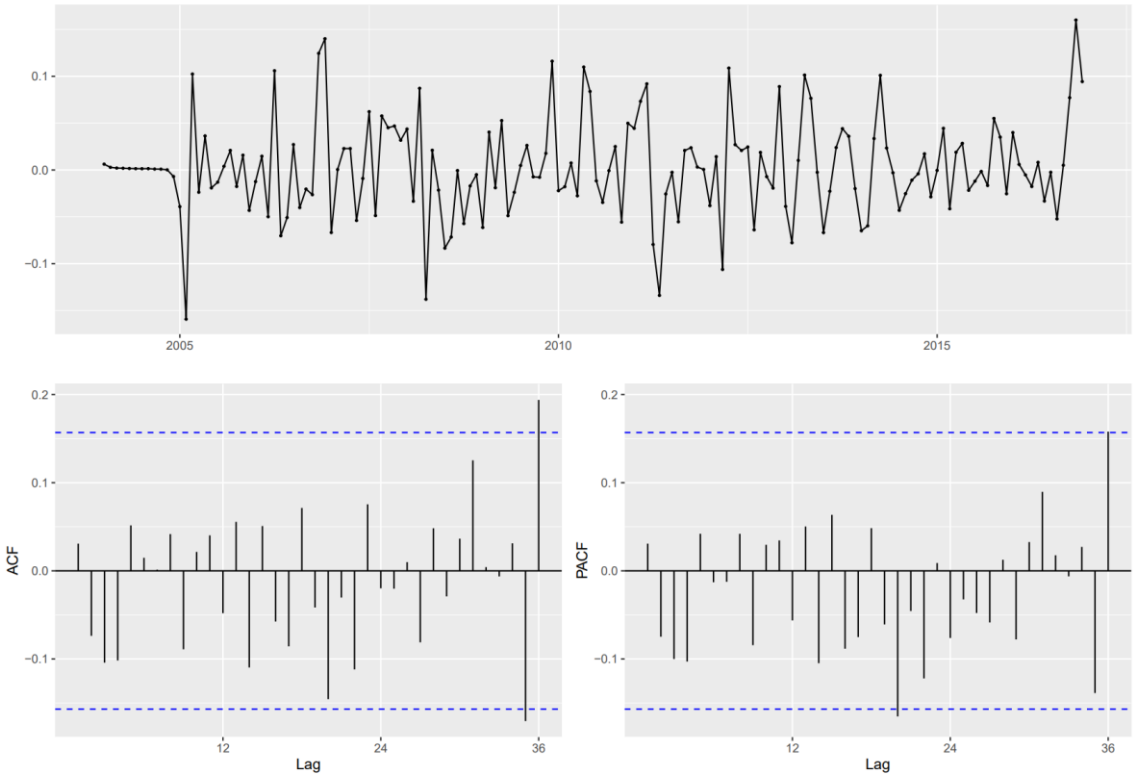
4.2.2 The Benchmark Model

Selection of the benchmark model from the 1,352 possible combinations of SARIMA models was done based on the AIC and SIC. The AIC suggested a $SARIMA(1,1,9)(1,0,1)$ while the SIC suggested a $SARIMA(1,1,0)(0,1,1)$. Since both goodness-of-fit measures referred to different models both were estimated on the sample of observations from January 2004 to December 2016, and subject to diagnostic checks to determine the econometrically correct model. The autocorrelation and residual tests on the $SARIMA(1,1,0)(0,1,1)$ revealed that this model held the best goodness-of-fit properties over the $SARIMA(1,1,9)(1,0,1)$.[19]

Figure 4.8 shows the ACF and PACF of the residuals of the $SARIMA(1,1,0)(0,1,1)$ model estimated on the training sample. A few slightly statistically significant spikes are present in the both the ACF and PACF at later lags, since they cross the dashed line, however the Ljung-Box test results in Table 4.4 indicate that the null hypothesis, that the residuals are independently distributed, is accepted, since the probability value is larger than 0.05 and thus, there is no serial correlation among the residuals. The Ljung-Box test results for the $SARIMA(1,1,9)(1,0,1)$ estimated on the same sample given in Appendix C.4 show that the null hypothesis is accepted as the probability value is less than 0.05 thereby validating the choice of the $SARIMA(1,1,0)(0,1,1)$ as the optimal benchmark model.

[19] The diagnostic test results on the $SARIMA(1,1,9)(1,0,1)$ are given in Appendix C.4.

Figure 4.8: Residual Plot, ACF and PACF for SARIMA(1,1,0)(0,1,1) Estimated on Training Sample.



Source: Author's calculations.

Table 4.4: Ljung-Box Test Results on SARIMA(1,1,0)(0,1,1) Estimated on Training Sample.

Q* value	Degrees of freedom	p-value	Model degrees of freedom	Total lags used
40.067	34	0.219	2	36

In view of these results, the $SARIMA(1,1,0)(0,1,1)$ was chosen as the benchmark model which is also the more parsimonious model from the two. The results from the estimated model are shown in Table 4.5, where $AR1$ refers to the non-seasonal autoregressive term and $SMA1$ is the seasonal moving average term.

Table 4.5: SARIMA(1,1,0)(0,1,1) Model Estimation Results on Training Sample.

	AR1	SMA1
Coefficient	-0.4709***	-0.7015***
Standard error	0.0756	0.0727
p-value	0.000	0.000

*** denotes statistical significance at the 0.01 level.

It is important to highlight that due to seasonal and non-seasonal differencing in the $SARIMA(1,1,0)(0,1,1)$ model, when the model is being estimated, it is actually being estimated from February 2005 instead of January 2004. This is because twelve observations are discarded due to one order of seasonal differencing and another observation is lost because of non-seasonal differencing. This is explained in Appendix C.5 in the context of the rolling window forecasting approach; however, it has no other implications on the results.

Furthermore, these results are only relevant for the model used to forecast January 2017 because this model estimated on the entire training sample. With a rolling window approach, the forecast for February 2017 and for the other months in the testing sample is generated from a model estimated on a different window of observations as explained in Section 3.3.2. Therefore, the coefficients of the $AR1$ and $SMA1$ terms change with each forecast. It is assumed that the goodness-of-fit properties hold for each model estimated to generate each forecast.

4.2.3 The Google Trends Model

The three first principal components from each dataset representing the Google Trends data were added to the benchmark model in order to construct the Google Trends model. The results from the model estimated on the training sample are shown in Table 4.6, where D denotes that the variables are first differenced.

Table 4.6: Google Trends Model Estimation Results on Training Sample.

	AR1	SMA1	D(PC1 Lag 5)	D(PC1 Lag 4)	D(PC1 Lag 1)
Coefficient	-0.4369***	-0.6641***	0.0012	0.0040	0.0031
Standard error	0.0783	0.0767	0.0029	0.0047	0.0055
p-value	0.000	0.000	0.681	0.392	0.576

*** denotes statistical significance at the 0.01 level.

The results in Table 4.6 show that the first principal components from each lagged dataset representing the Google Trends data are statistically insignificant at the 10% level and that the coefficients are very close to zero. The statistical insignificance of the variables does not necessarily infer the inability of this model to produce better forecasts than the benchmark model. The aim is to assess whether the out-of-sample performance of the Google Trends model outperforms that of the benchmark model and thus in-sample results are not the outcomes of interest.

As highlighted for the benchmark model, these results pertain to the model estimated on the sample starting from January 2004 to December 2016. The coefficients for the autoregressive term, the seasonal moving average term and the principal components change each time the estimation window is rolled forward to generate the next forecast.

4.3 FORECASTING SIMULATION

4.3.1 Overall Forecasting Performance

As described in the previous chapter, the forecasting performance of the benchmark model and the Google Trends model is evaluated using the RMSE, MAE and MAPE. These are obtained from the forecast error which is the difference between the forecasted value of tourist arrivals for each month generated by both models and the actual value of tourist arrivals in that month. These errors are calculated as shown in equations (3.13), (3.14) and (3.15) in order to obtain a measure of the overall forecasting performance of the model on the testing sample. The results are summarised in Table 4.7.

Table 4.7: Overall Forecast Accuracy for One-Step Ahead Forecasts from the Benchmark Model and from the Google Trends Model.

Forecast Accuracy Metric	Benchmark Model	Google Trends Model
RMSE	9040.19	8717.21
MAE	7495.80	7218.51
MAPE	3.72	3.59

Despite the statistical insignificance of the Google Trends variables and their coefficient value being close to zero, the Google Trends model's overall forecasting performance on the testing set of 36 months is better than that of the benchmark model because the different measures of the forecast error are all less than that of the benchmark model.

This implies that forecasts of tourist arrivals are more accurate from the Google Trends model than from the benchmark model and indicates that Google Trends data does indeed provide more accurate forecasts of tourist arrivals for the short-term. The forecast error for each month forecasted is examined in the next section for a more in-depth analysis of the forecasting performance of the Google Trends model throughout the period January 2017 to December 2019.

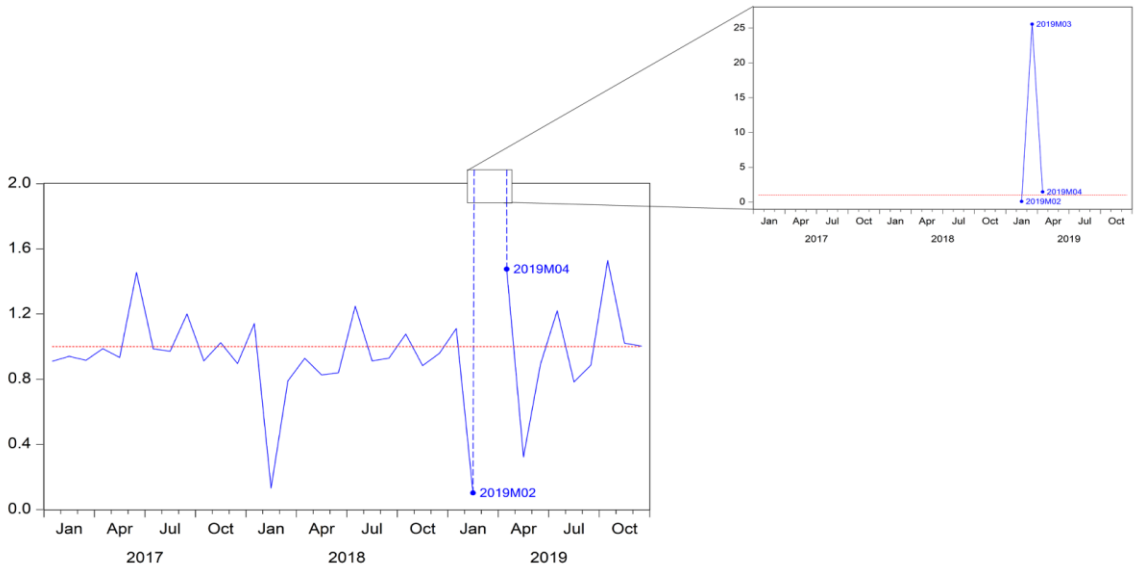
4.3.2 Relative Performance of Benchmark and Google Trends Model

The resultant RMSEs from the forecasts of the Google Trends model relative to those of the benchmark model for each month is plotted in Figure 4.9 where the horizonal red line is equal to one.[20] A value below one indicates that the Google Trends model is performing better than the benchmark model while values greater than one show worse performance. The relative

[20] The month-by-month RMSE is chosen because it is equal to the month-by-month MAE for one-step ahead forecasts. This is because in the calculation of the RMSE for each monthly forecast, the error is squared, divided by 1 because the errors are for *1*-step ahead forecasts and then square rooted which results into the absolute value of the error since it was initially squared. Furthermore, the MAPE shows the same information as the RMSE, therefore, the choice to show the RMSE only does not affect the presentation of results.

error shows that the Google Trends model does not consistently outperform the benchmark model in all months as sometimes it is larger than one.

Figure 4.9: Google Trends Model RMSE Relative to Benchmark Model RMSE for each Month in the Testing Sample.



Source: Author's calculations.

The Google Trends model outperforms the benchmark model in the first months of the first year of the testing sample. It is then outperformed by the benchmark model in June but regains accuracy in the months of July and August. The Google Trends model provides more accurate forecasts for October and December during the last four months of 2017.

For the year of 2018, the Google Trends model performs the best from February to June and for the months of August, September, November and December. The performance of this model during this year is similar to that of the first year as it is only beaten by the benchmark model in three months.

From this plot, the Google Trends model is shown to perform the worst during 2019 as it only manages to outperform the benchmark model during the months of February, May, June, August and September. For 2019, the benchmark model provides better forecasts for more than half of the year as it exhibits a lower forecast error for the rest of the months.

A very significant spike is recorded in March 2019 as the performance of Google Trends model is 25 times worse than that of the benchmark model. The forecast from the benchmark model during this month is **173,087** arrivals while that from the Google Trends model is **175,948.** The actual number of tourist arrivals during March 2019 is **172,971** which is very close to the figure forecasted by the benchmark model which results into a RMSE (and MAE) for the forecast of this month to be 116. On the other hand, the RMSE of the Google Trends model is 2977 which is larger than that from the benchmark model and thus results into a larger error in relative terms.

4.4 ROBUSTNESS CHECKS

4.4.1 Clark & West (2007) Test

One way to check the robustness of the results obtained from the one-step ahead forecast simulation is to perform the procedure by Clark & West (2007) explained in the previous chapter. The results from the regression of z_i on a constant term are presented in Table 4.8.

Table 4.8: Results from Clark & West (2007) Test.

	Estimate	Standard error	t-statistic
Constant term	7,958,929	3,983,961	1.998

The result of interest is the value of the t-statistic. Since it is larger than 1.645, which is the value for a one-sided test at the 5% level of significance, the null hypothesis that the MSPEs of the benchmark model and the Google Trends model are equal to each other, is rejected. Therefore, since the alternative hypothesis is accepted, it can be concluded that the MSPE from the Google Trends model is smaller, and the forecasts are generated by this model and not the benchmark model.

In the words of Diebold (2015), this result provides better "insurance" than the pseudo-out-of-sample simulations that the forecasts from the Google Trends model are better than those from the benchmark model without Google Trends data.

4.4.2 Pseudo-Out-of-Sample Forecasting Simulation from a Model with GDP and REER

In line with the literature on tourism demand models, a third model was formulated with real GDP and the REER as predictors. These variables were tested for stationarity using the ADF test and the results are given in Appendix C.6. The results in Panel B of Table C.6.2 show that the GDP series is stationary at the 5% level when log transformed and first differenced. The results from the ADF test on the REER series in Table C.6.3 and C.6.4 show that it is stationary in first differences both in levels and when log transformed. Therefore, these variables are included in the model log transformed and in first differences.

These variables are added to the $SARIMA(1,1,0)(0,1,1)$ benchmark model for comparison with the Google Trends model. Table 4.9 shows the results of the model estimated on the training sample which only apply for the forecast of January 2017 as highlighted in Section 4.2.2 and 4.2.3. The explanatory variables are statistically insignificant at the 10% level, which shows that they do not explain the variation in tourist arrivals. However, as in the case of the Google Trends model, this is not indicative of the predictive power of these variables.

Table 4.9: SARIMA (1,1,0)(0,1,1) Model with GDP and the REER as Explanatory Variables Estimation Results on Training Sample.

	AR1	SMA1	D(log(GDP))	D(log(REER))
Coefficient	-0.4615***	-0.6922***	1.5637	0.0589
Standard error	0.0761	0.0745	1.6318	0.4248
p-value	0.000	0.000	0.338	0.890

*** denotes statistical significance at the 0.01 level.

Table 4.10 shows the overall forecasting performance of this model in comparison with that of the Google Trends model. These results show that although the RMSE, MAE and MAPE of

each model are very close to each other, the Google Trends model still outperforms the model with other explanatory variables of tourism demand as suggested by economic theory.

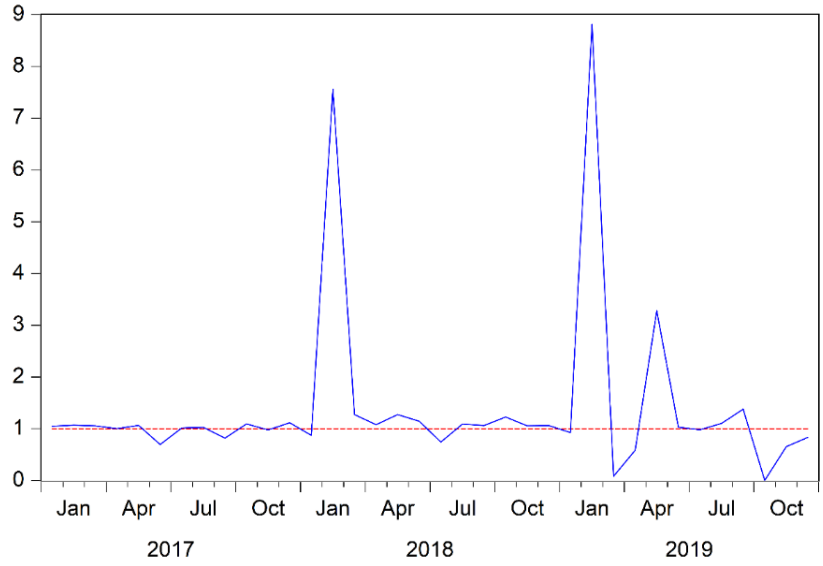
Table 4.10: Overall Forecast Accuracy for One-Step Ahead Forecasts from the Explanatory Variables Model and from the Google Trends Model.

Forecast Accuracy Metric[21]	Google Trends Model	Explanatory Variables Model
RMSE	8717.208	8836.573
MAE	7218.514	7293.525
MAPE	3.588	3.595

The relative RMSE in Figure 4.10 is calculated by dividing each month's RMSE of the model with explanatory variable by that from the Google Trends model since this model is the "benchmark model" in this case. In most instances, the value of the relative error is greater than one signifying the worse performance of the explanatory variables model in comparison with that of the Google Trends model.

[21] As explained in Section 3.1.3, these are measures of the forecast error that are used to compare and evaluate the out-of-sample forecasting performance of a model. The RMSE takes the square root of the average of sum of squared errors, the MAE takes the average of the sum of the absolute value of the errors and the MAPE takes the MAE as a percentage of the actual value of the forecasted observation. The forecast errors in question are the differences between the actual value of tourist arrivals and the forecasted value over the testing sample which is from January 2017 to December 2019.

Figure 4.10: Explanatory Variables Model RMSE Relative to Google Trends Model RMSE for One-Step Ahead Forecasts (red dotted line is equal to 1).



Source: Author's calculations.

The most notable spikes occur in February 2018 whereby the model with explanatory variables performs seven times worse than the Google Trends model, and in February 2019 where the performance of the same model is almost nine times worse. The reason for these spikes is similar as to why the Google Trends model performed 25 times worse than the benchmark model in Section 4.3.2. In this case, the forecast of tourist arrivals from the Google Trends model for February 2018 and 2019 was much closer to the actual value of inbound tourists than the forecast from the model with explanatory variables thereby resulting into a worse performance of the latter model in relative terms.

4.5 NOWCASTING EXPERIMENT

Since the Google Trends model appears to show superior forecasting performance and is also robust to the Clark & West (2007) test and the model with GDP and the REER as explanatory variables, these results permit further exploration of Google Trends data by conducting a nowcasting experiment.

Nowcasting is defined as forecasting the present, the immediate future or the very recent past (Bańbura et al., 2013) and has become the subject of several working papers of central banks

(see Eraslan & Schröder, 2019 and Jardet & Meunier, 2022) including the European Central Bank (see Ashwin et al., 2021 and Marozzi, 2021) and the Central Bank of Malta (see Ellul & Ruisi, 2022). The need for real-time forecasting arises from the lack of availability of high frequency data and publication lags by national statistics bureaus. Publication lags are problematic for national policymakers as the latest state of the economy remains unknown until statistics are made available.

Google Trends data is available almost instantaneously and certainly with a lead over the publication of the tourist arrivals statistics by the NSO therefore this is why it can be used as a predictor for inbound tourism. At the time of writing in mid-September 2022, inbound tourism statistics for July 2022 were released on the 5th of September by the NSO, that is, with a six week lag since the end of July. Therefore, assuming that it is the beginning of September and July tourism statistics are yet to be released, the number of tourist arrivals in July and August can be nowcasted using tourist arrivals data up to June 2022 and Google Trends data that is available up to August 2022 in the absence of tourism statistics for July and August.

This experiment nowcasts tourist arrivals for July and August as if it were the beginning of September and July figures were not yet made available, therefore it is a demonstration of forecasting the very recent past. The experiment differs from the simulation carried out in the previous section because the purpose of it is to merely demonstrate a practical application of the Google Trends model after having identified it as the model with superior forecasting abilities. Considering the data affected by the pandemic, the methodology employed to generate the nowcasts is an oversimplification of what could be done in practice to control for the period affected by travel restrictions.

4.5.1 Nowcasting Methodology

The methodology used to generate the nowcasts entails the generation of two-step ahead forecasts from the $SARIMA(1,1,0)(0,1,1)$ model estimated on tourist arrival data from January 2004 to June 2022 obtained from the NSO, and Google Trends data from January 2004 to August 2022. The two-step ahead forecast involves a one-step ahead forecast to obtain the forecast for July 2022 and the use of this forecasted observation to forecast August 2022 since it is assumed that no actual data for July 2022 is available at the beginning of September prior to the NSO's news release on the 5th.

Since the tourist arrival data is log transformed, the data points for April, May and June 2020 which are zero because of the closure of the Malta International Airport (MIA) due to the onset of the pandemic, are changed to 0.001. Furthermore, because the pandemic causes a break in the inbound tourists time series, a dummy variable that is equal to one during April, May and June 2020 and equal to zero otherwise is included. This is a simple way to control for the effect of the pandemic on the data while still being able to show the Google Trends model at work. It is left as a suggestion for further research in the concluding remarks of this dissertation to deal with the *Covid-19 data* in a more econometrically adequate way.

Google Trends data for the 128 queries which are left after duplicates were removed after the collection process as explained in Section 3.4.2.3 are re-downloaded. The rest of the process outlined in Figure 4.1 is applied to this data. Due to the way that Google samples the data as described in Section 3.4.2.1, the newly downloaded data is slightly different, therefore, the queries with zero value data points that were removed are different and 25 queries remain on which the correlation analysis is carried out on.

The results from the correlation analysis are presented in Appendix C.7 and from these, two datasets which contain queries lagged by four months and queries lagged by one period are formed.[22] The PCA is conducted on these datasets and following the same reasoning applied in Section 4.1.2, the first principal component from each PCA is included in the model. The results from the PCA, stationarity tests and the model estimated on the sample from January 2004 to June 2022 are also in Appendix C.7 as the focus of this experiment is to show the nowcasts generated.

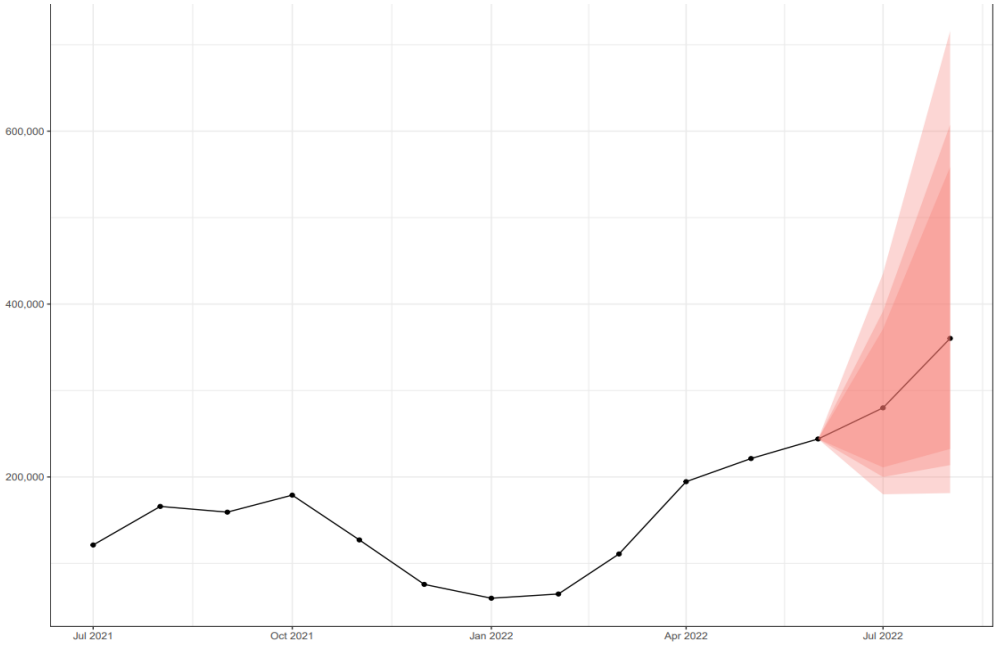
4.5.2 Nowcasting Results

The results from the nowcasting experiment are presented in the fan chart in Figure 4.11. The fan chart is a plot of actual tourist arrival data from July 2021 to June 2022 together with the point nowcasts for July and August 2022 and confidence interval bands. The black solid line

[22] This is because after removing queries with a highest correlation less than the 0.30 threshold as before, two queries were left that were highly correlated at the 5th lag and one query was left that was highly correlation at the 7th lag, therefore PCA could not be conducted on two or one variables. The 0.30 threshold was chosen because removing queries with a highest absolute correlation less than the average of this dataset, which was 0.37, would have resulted into a few queries which does not allow for PCA.

in the red part of Figure 4.11 shows the point nowcasts for July and August 2022 while the red bands depict the upper and lower bounds of the 90%, 95% and 99% confidence intervals, from darkest to lightest.

Figure 4.11: Nowcasts of Tourist Arrivals for July and August 2022 (black line) with 90% (darkest red), 95% and 99% (lightest red) Confidence Intervals.



Source: Author's calculations.

The results from the nowcasts as demonstrated in Figure 4.11 correspond to the figures in Table 4.11. The point estimates indicate a total of total of 279,947 arrivals in July 2022 and 360,291 tourists in August 2022. July figures are 92% of 2019 levels while those of August imply a more than 100% recovery on August 2019 levels as the forecasted value is 15% higher than that in 2019. The implications of these results are discussed in Section 4.6.2.

Table 4.11: Nowcasted Tourist Arrival Values.

	Lower			Point Forecast	Upper		
	90%	95%	99%		90%	95%	99%
July 2022	211,124	200,015	179,963	279,947	371,204	391,829	435,479
August 2022	232,441	213,721	181,376	360,291	558,463	607,377	715,692

4.6 DISCUSSION

4.6.1 Summary of Results

The aim of this dissertation was to test the predictive power of Google Trends search data in forecasting tourist arrivals in Malta. The advantage of Google Trends data over other data that is used to forecast tourist arrivals is that it is available with a lead over inbound tourism statistics and accessible to anyone and thus can be utilised as a predictor. The research question was answered by comparing the forecasting performance of a model with this data with that of a benchmark model without such data.

The benchmark model consists of a $SARIMA(1,1,0)(0,1,1)$ chosen according to the SIC and in line with the appropriate residual testing. The competing Google Trends model entailed the same benchmark model but augmented with Google Trends search data represented by the first principal component from three datasets that contained queries lagged by five, four and one month separately. One-step ahead forecasts were obtained from each model and their overall performance is summarised in Table 4.7.

The results show a superior forecasting performance of the Google Trends model on the testing sample relative to that of the benchmark model since the value of each accuracy metric for the Google Trends model is less than that of the benchmark model. In this regard, it can be said that Google Trends search data generates better short-term forecasts of tourist arrivals.

Figure 4.9 demonstrated the month-by-month RMSE of the Google Trends model relative to that of the benchmark model. In this figure, the Google Trends model is better at forecasting a particular month if the value of the relative error is less than one, otherwise, it means that it is outperformed by the benchmark model. By means of this demonstration, the Google Trends model's inconsistency in forecasting is highlighted as it does not beat the benchmark model in all of the months of the forecast period despite being the model with the better overall forecasting ability.

In Table 4.12 the RMSE, MAE, and MAPE for each year of the testing sample are shown. Even though the Google Trends model is inconsistent when its monthly performance is examined, the errors for each year of the testing sample show that the Google Trends model outperforms the benchmark model in each year.

Table 4.12: Forecast Accuracy for Each Year of the Testing Sample.

	Benchmark Model			**Google Trends Model**		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**
RMSE	10182.27	9320.32	7391.09	10108.76	8634.68	7157.13
MAE	8524.93	7752.69	6209.80	8502.20	7182.61	5970.74
MAPE	4.55	3.40	3.21	4.51	3.13	3.12

The robustness of the one-step ahead forecast results were tested via the Clark and West (2007) test and against the forecasts from a third model. The Clark & West (2007) procedure confirmed, at the 5% level of significance, that the models are not of equal predictive accuracy and that the forecasts are more likely to be generated from the Google Trends model.

The second robustness check consisted of a forecasting simulation from a model with two independent variables that are typically used to explain tourism demand as delineated in the second chapter: GDP and the REER. The Google Trends model acted as the benchmark model while the model with explanatory variables was the competing model in this exercise. The results from this check show that the forecast accuracy metrics are very close in magnitude, however, the Google Trends model still revealed the better forecasting performance as summarised in Table 4.10.

4.6.2 Nowcasts

The nowcasting experiment was carried out as if it were the beginning of September and data for tourist arrivals was available up to June 2022. This was used in order to nowcast the figures for July 2022 and August 2022 with Google Trends data up to August 2022. The nowcast for July 2022 can be evaluated against the actual statistics published by the NSO on the 5th of September 2022.

The nowcasted figure for tourists in July 2022 is **279,947** while the actual number of tourist arrivals published by the NSO is **273,646** (NSO, 2022).[23] This shows that the nowcast from the Google Trends model is overestimated by only 2.3% of the actual figure and would thus have been a good estimate of tourist arrivals to bridge the publication lag.

On the other hand, the nowcast for August 2022 might be overly optimistic on account of being larger than the pre-pandemic level recorded in August 2019 which was 336,547. Since data for August 2022 shall be released on the 6th of October according to the NSO's release calendar, which is after the submission of this dissertation, this nowcast cannot be evaluated against actual data. However, the given that the forecast indicates a higher level of tourist arrivals than that recorded in August 2019 before the pandemic, it is highly likely that tourist arrivals for August 2022 are being overestimated.

4.6.3 Implications for Private and Public Sector Institutions

Short-term forecasts of tourist arrivals are useful to private and public sector players in the tourism industry such as hotels and the Malta Tourism Authority (MTA). Such forecasts would enable these entities to make efficient and effective plans according to the number of tourists expected to come to Malta. The hospitality industry in the private sector can benefit from short-term forecasts of tourist arrivals as an indicator of the demand for their services. For instance, hotels can set prices in accordance with the demand indicated by the forecasts of arrivals for the month.

[23] This figure relates to News Release 163/2022 published by the NSO which might differ from the figure in future news releases due to possible data revisions.

The impact of Google Trends data in forecasting tourist arrivals is also an indicator that much of holiday planning is done by searching on the internet. The implication of this with regards to the public sector is that the MTA should direct its attention to online searching behaviour in order to capture the interests of potential tourists and thus digitally market Malta's attractions accordingly.

The improvement to the forecasts made by Google Trends data also shows that this data could be considered in macroeconomic projections of the Maltese economy as an input in the forecasts of tourism exports. Furthermore, the recent past can be nowcasted with the use of this data and used as unofficial data in projections until actual data is made available. National institutions that are concerned with such projections are the Central Bank of Malta and the Economic Policy Department in the Ministry for Finance.

4.6.4 Comparison of Findings with Other Studies

This study has been the first to explore the potential of Google Trends search data in predicting a macroeconomic variable of the Maltese economy. In this regard, since it is not possible to compare the findings of this study to previous work, reference is made to studies that also assessed the ability of Google Trends data to forecast tourist arrivals in other countries.

The findings of this dissertation are comparable to those of Antolini & Grassini (2019) and Artola et al. (2015) since an almost identical methodology is employed to test the predictive power of Google Trends data. Artola et al. (2015) find that Google Trends search data improves forecasts of tourist arrivals in Spain for the short-term but record a worsened performance of the Google Trends model later in the forecast period. The former assess the ability of the data to nowcast inbound tourism in Italy two-months ahead of the last estimation period and conclude that Google Trends search data did not significantly contribute to the nowcasting exercise. Moreover, the Google Trends model provided worse results for the summer period during which more accurate forecasts are needed.

The application of PCA in García Rodríguez's (2017) similar work on the Balearic Islands proved to be successful in improving forecasts of tourist arrivals over the simple baseline model. This shows that incorporating a high volume of data by means of this method is an alternative way of exploiting the potential of Google Trends data instead of limiting the

explanatory variables to a few keywords and thus encountering significant loss of data. Park et al. (2016) also included several keyword data in the form of an index and report positive results from the out-of-sample performance of the Google Trends models when forecasting arrivals to South Korea.

The consensus from studies in this field is that Google Trends data is able to provide better forecasts of inbound tourism for the short-term. The findings of this dissertation are mostly in line with those of Antolini & Grassini (2019) and Artola et al. (2015) in that the usefulness of Google Trends data lies in forecasting arrivals for the short-term. However, it cannot be said that a worsened performance is recorded at the end of the forecast period since a lower error for each year of the training sample is obtained from the forecasts of the Google Trends model as shown in Table 4.12. Similar also to Havranek & Zeynalov (2021), the results provide evidence for superior forecasting abilities of Google Trends search data over GDP and the REER which are other explanatory variables of tourism demand. Nevertheless, this does not imply that Google Trends data is not without limitations and therefore, there lies room for improvements as shall be considered in the next chapter.

CHAPTER 5
CONCLUSION

5.1 MAIN FINDINGS

This study has been the first to assess the potential of Google Trends data in forecasting a macroeconomic variable in the Maltese context. In particular, the research question that this dissertation sought to answer was:

"Does Google Trends data improve forecasts of tourist arrivals in Malta?".

Econometric time series methods were used to provide an answer and it can be concluded that Google Trends data does indeed provide better forecasts of inbound tourists and such forecasts are statistically significant and also robust to forecasts from a structural econometric model with GDP and the REER as predictors. In view of these results, it was possible to conduct a nowcasting experiment to demonstrate a practical application of the model with Google Trends data. The nowcast of tourist arrivals for July 2022 was compared to the actual figure that was published by the NSO and this showed that it was very close to the number of arrivals in Malta during July 2022 thereby justifying the use of this model in public and private sector settings.

The implications of these results are relevant for the tourism industry in Malta, public policymakers such as the Malta Tourism Authority, national institutions that are concerned with projections of the macroeconomy like the Central Bank of Malta and several sectors of the economy that are impacted by the indirect and induced positive and negative externalities of tourism.

5.2 LIMITATIONS

Despite the positive results attained, this study is not without its limitations. One major limitation is in regard to the length restrictions of the dissertation. The forecasting potential of Google Trends data can be tested in so many ways that a lot of possible tests for the forecasting ability of the data had to be forgone.

Another shortcoming is that of the Google Trends data. As explained in the third chapter, the data varies from day to day due to the way that Google samples the data. Therefore, the data that was downloaded on a particular day for the use in this dissertation, is not the same as the data for the same queries downloaded on another day. This is highlighted in several studies that

question the use of Google Trends data such as Cebrián & Domenech (2022) and Dergiades (2018) however, they still do not invalidate the use of such data. The problem with this feature of the data is that it interferes with the replicability of the study since the exact same results cannot be achieved. Nevertheless, the changes in the data are not extremes; this limitation pertains to data points not being the same by a few numbers but the seasonality features are still preserved. In addition to this, the collection of the Google Trends data was done rather subjectively especially in the initial stage whereby certain queries are eliminated, therefore this could have biased the results.

Furthermore, the effect of the Covid-19 pandemic on the time series in the second quarter of 2020 presented a limitation since the study could not be carried out using all data available. This drawback was also encountered for the nowcasting experiment which led to the use of dummy variables for a simple way to control for the break in the series.

5.3 SUGGESTIONS FOR FURTHER RESEARCH

This dissertation can be seen as a foundation for further research on the potential of Google Trends data to forecast tourism demand in Malta. A general recommendation is to apply different methodological approaches, model specifications and also variations in the collection of Google Trends data for more evidence, or rejection, of this data as a tool for forecasts of tourist arrivals.

A different approach to the forecasting methodology can be applied to examine the performance of Google Trends data in providing longer-term forecasts. Direct multi-step ahead forecasts as opposed to the one-step ahead forecasts performed in this study can be simulated to test for this. The model specifications used in future research should not be limited to time series models. MIDAS regressions that combine low frequency data with higher frequency data can explore the potential of daily or weekly Google Trends data in forecasting tourist arrivals at a lower frequency.

Different search queries can be tested to assess any changes in results. Furthermore, principal component analysis was used as a method for the inclusion of a relatively large amount of data compared with the inclusion of just one or two series of query data. There are other methods

proposed in the literature which are worth exploring as different approaches to incorporating the data such as to see how results may differ from those of this study.

Finally, recently there have been national concerns that the *quality* and not the *quantity* of tourists should be targeted in order to ensure significant contributions from tourists to Malta's economy. Therefore, the identification of search queries related to the expenditure patterns of tourists which can be used to forecast tourist expenditure is another scope for research. The application of forecasting with Google Trends data can also be extended to other areas of the macroeconomy like those discussed in Appendix A.1.

REFERENCES

Aljandali, A., & Tatahi, M. (2018). *Economic and Financial Modelling with EViews A Guide for Students and Professionals*. Springer International Publishing AG, part of Springer Nature.

Antolini, F., & Grassini, L. (2019). Foreign arrivals nowcasting in Italy with Google Trends data. *Quality and Quantity, 53*(5), 2385-2401. https://doi.org/10.1007/s11135-018-0748-z.

Anttonen, J. (2018). *Nowcasting the Unemployment Rate in the EU with Seasonal BVAR and Google Search Data*. Helsinki: The Research Institute of the Finnish Economy (ETLA).

Artola, C., Pinto, F., & de Pedraza, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*(1), 103-116. https://doi.org/10.1108/IJM-12-2014-0259.

Ashwin, J., Kalamara, E. & Saiz, L. (2021). Nowcasting euro area GDP with news sentiment: a tale of two crises. *European Central Bank Working Paper Series No 2616/November 2021*. http://dx.doi.org/10.2139/ssrn.3971974.

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly, 55*(2), 107-120. https://ssrn.com/abstract=1480251 or http://dx.doi.org/10.2139/ssrn.1480251.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822-844. https://doi.org/10.1016/j.ijforecast.2010.04.009.

Attard, S. (2019). The evolution of Malta's Tourism Sector. *Xjenza, 7*(1), 37-48. https://doi.org/10.7423/XJENZA.2019.1.04.

Bańbura, M., Giannone, D., Modugno, M & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow. *Handbook of Economic Forecasting, 2*(A), 195-237. https://doi.org/10.1016/B978-0-444-53683-9.00004-9.

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46,* 454-464. https://doi.org/10.1016/j.tourman.2014.07.014.

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

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time Series Analysis Forecasting and Control* (5th ed.). John Wiley & Sons Inc.

Brooks, C. (2008). *Introductory Econometrics for Finance* (2nd ed.). Cambridge University Press.

Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, *37*(3), 303-315. https://doi.org/10.1002/for.2500.

Jardet, C. & Meunier, B. (2022). Nowcasting World GDP Growth with High-Frequency Data. *Journal of Forecasting, 41*(6), 1181-1200. https://doi.org/10.1002/for.2858.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting, 32(4),* 289-298. https://doi.org/10.1002/for.1252.

Cassar, I. P., Vella, K., & Buttigieg, S. (2016). Understanding the Economic Contribution of Tourism in Malta: A Literature Review. *Mediterranean Journal of Social Sciences, 7*(6), 49-60. https://doi.org/10.5901/mjss.2016.v7n6p49.

Cebrián, E. & Domenech, J. (2022). Is Google Trends a Quality Data Source? *Applied Economics Letters.* 1-5. https://doi.org/10.1080/13504851.2021.2023088.

Cevik, S. (2020). Where Should We Go? Internet Searches and Tourist Arrivals. International Journal of Finance & Economics, 1-10. https://doi.org/10.1002/ijfe.2358.

Choi, H., & Varian, H. (2009). *Predicting Initial Claims for Unemployment Benefits* Google.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *The Economic Record, 88*(s1), 2-9. http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x.

Clark, T. E. & West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics, 135*(1-2), 155-186. https://doi.org/10.1016/j.jeconom.2005.07.014.

Clark, T. E. & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*(1), 291-311. https://doi.org/10.1016/j.jeconom.2006.05.023.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting, 33*(4), 801-816. https://doi.org/10.1016/j.ijforecast.2017.03.004.

Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management, 66*, 108-120. https://doi.org/10.1016/j.tourman.2017.10.014.

Dickey, D. A. & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association, 74*, 427-431. https://doi.org/10.2307/2286348.

Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1198/073500102753410444.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics, 33*(1), 1-9. https://doi.org/10.1080/07350015.2014.983236.

Dwyer, L., Forsyth, P., & Dwyer, W. (2020). *Tourism Economics and Policy* (2nd ed.) Channel View Publications.

Ellul, R & Ruisi, G. (2022). Nowcasting the Maltese Economy with a Dynamic Factor Model. *Central Bank of Malta Working Paper WP/02/2022*.

Enders, W. (2014). *Applied Econometric Time Series* (4th ed.). John Wiley & Sons Inc.

Eraslan S. & Schröder, M. (2019). Nowcasting GDP with a larger factor model space. *Deutsche Bundesbank Discussion paper 41/2019.* http://dx.doi.org/10.2139/ssrn.3507664.

European Commission. (2016). Preferences of Europeans towards tourism. Retrieved from: https://europa.eu/eurobarometer/surveys/detail/2065.

Eurostat. (2021). Effective exchange rate indices (ert_eff). Retrieved from: https://ec.europa.eu/eurostat/cache/metadata/en/ert_eff_esms.htm.

García Rodríguez, Ó. (2017). Forecasting tourism arrivals with an online search engine data: A study of the Balearic Islands. *Journal of Tourism and Cultural Heritage, 15*(4), 943-958. https://doi.org/10.25145/j.pasos.2017.15.064.

Goh, C., & Law, R. (2011). The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature. *Journal of Travel & Tourism Marketing, 28*(3), 296-317. https://doi.org/10.1080/10548408.2011.562856.

Google. (2022). FAQ about Google Trends data. Retrieved from: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052.

Google. (n.d.). Basics of Google Trends. Retrieved from: https://newsinitiative.withgoogle.com/training/lesson/5748139575214080?image=trends&tool=Google%20Trends.

Gujarati, D. N., & Porter, D. C. (2010). *Essentials of Econometrics* (4th ed.). McGraw-Hill/Irwin.

Hall, R. C., Griffiths, W. E. & Lim, G. C. (2019). *Principles of Econometrics* (4th ed.). John Wiley & Sons, Inc.

Hamilton, J. D., & Xi, J. (2022). Principal Component Analysis for Nonstationary Series. Retrieved from: https://econweb.ucsd.edu/~jhamilto/HX.pdf.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, 13*(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4.

Havranek, T., & Zeynalov, A. (2021). Forecasting tourist arrivals: Google Trends meets mixed-frequency data. *Tourism Economics, 27*(1), 129-148. https://doi.org/10.1177/1354816619879584.

Holland, M. S. (2019). *Principal Components Analysis (PCA).* Retrieved from: http://strata.uga.edu/8370/handouts/pcaTutorial.pdf.

Höpken, W., Eberle, T., Fuchs, M., & Lexhagen, M. (2019). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology & Tourism, 21*, 45-62. https://doi.org/10.1007/s40558-018-0129-4.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components. *Journal of Educational Psychology, 24*(6), 417-441. https://doi.org/10.1037/h0071325.

Hyndman, R. (2020). forecast package – R Documentation. Retrieved from: https://www.rdocumentation.org/packages/forecast/versions/8.12.

Hyndman, R. J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice* (2nd ed.). OTexts: Melbourne, Australia. OTexts.com/fpp2.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Inoué, A., Jin, L., & Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics, 196*(1), 55-67. https://doi.org/10.1016/j.jeconom.2016.03.006.

Ito, T., & Takeda, F. (2022). Do sentiment indices always improve the prediction accuracy of exchange rates? *Journal of Forecasting, 41*(4), 840-852. https://doi.org/10.1002/for.2836.

Ito, T., Masuda, M., Naito, A., & Takeda, F. (2021). Application of Google Trends-based sentiment index in exchange rate prediction. *Journal of Forecasting, 40*(7), 1154-1178. https://doi.org/10.1002/for.2762.

Jackman, M., & Naitram, S. (2015). *Research note:* Nowcasting tourist arrivals in Barbados - just Google it! *Tourism Economics, 21*(6), 1309-1313. https://doi.org/10.5367/te.2014.04.

Li, G., Song, H., & Witt, S. F. (2006). Forecasting Tourism Demand Using Econometric Models. *Tourism management dynamics: Trends, Management and Tools* (pp. 219-228). Oxford: Butterworth-Heinemann. https://doi.org/10.1016/B978-0-7506-6378-6.50033-0.

Li, G., Song, H., & Witt. S. F. (2005). Recent Developments in Econometric Modelling and Forecasting. *Journal of Travel Research, 44*(1), 82-99. https://doi.org/10.1177/0047287505276594.

Li, X., Pan, B., Law, R., & Xiankai, H. (2017). Forecasting Tourism Demand with Composite Search Index. *Tourism Management, 59*, 57-66. https://doi.org/10.1016/j.tourman.2016.07.005.

Lim, C. (1997). Review of International Tourism Demand Models. *Annals of Tourism Research, 24*(4), 835-849. https://doi.org/10.1016/S0160-7383(97)00049-2.

Lim, C. (1999). A Meta-Analytic Review of International Tourism Demand. *Journal of Travel Research, 37*(3), 273-284. https://doi.org/10.1177/004728759903700309.

Ljung, G. M. & Box, G. E. P. (1978). On a measure of lack of fit in time series models, *Biometrika*, *65*(2), 297-303. https://doi.org/10.2307/2335207.

Malta International Airport. (2017). Traffic Results – August 2017. *Company Announcement 263/2017.* Retrieved from: https://www.maltairport.com/wp-content/uploads/2016/03/Traffic-Results-August-2017.pdf.

Malta International Airport. (2018). Traffic Results – August 2018. *Company Announcement 287/2018.* Retrieved from: https://www.maltairport.com/wp-content/uploads/2018/09/Traffic-Results-August-2018.pdf.

Malta International Airport. (2022). August Traffic Results. *Company Announcement 372/2022.* Retrieved from: https://www.maltairport.com/wp-content/uploads/2022/09/Traffic-August-2022.pdf.

Marcellino, M., Stock, H. J., & Watson, W. M. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics, 135*(1-2), 499-526. https://doi.org/10.1016/j.jeconom.2005.07.020.

Marozzi, A. (2021). The ECB's tracker: nowcasting the press conferences of the ECB. *European Central Bank Working Paper Series No 2609/October 2021.*

Massicotte, P., & Eddelbuettel, D. (2022). Package 'gtrendsR'. Retrieved from: https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf.

McKinley, S. & Levine, M. (n.d.). Cubic Spline Interpolation. Retrieved from: https://www.rajgunesh.com/resources/downloads/numerical/cubicsplineinterpol.pdf.

Montgomery, C. D., Jennings, L. C., & Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting* (2nd ed.). John Wiley & Sons Inc.

Morley, C. L. (2009). Dynamics in the specification of tourism demand models. *Tourism Economics, 15*(1), 23-39. https://doi.org/10.5367/000000009787536654.

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the U.S. unemployment rate using Google Trends. *Finance Research Letters, 30,* 103-109. https://doi.org/10.1016/j.frl.2019.04.005.

National Statistics Office. (2022). Inbound Tourism: July 2022. *News Release 163/2022*. Retrieved from: https://nso.gov.mt/en/News_Releases/Documents/2022/09/News2022_163.pdf.

Pan, B., Litvin, S. W., & Goldman, H. (2006). Real Users, Real Trips, and Real Queries: An Analysis of Destination Search on a Search Engine. *Annual Conference of Travel and Tourism Research Association, Ireland.*

Park, S., Lee, J., & Song, W. (2016). Short-term forecasting of Japanese tourist inflows to South Korea using Google trends data. *Journal of Travel & Tourism Marketing*, *34*(3), 357-368. https://doi.org/10.1080/10548408.2016.1170651.

Rachev, T.S., Mittnik, S., Fabozzi, J.F., Focardi, M.S., & Jašić, T. (2007). *Financial Econometrics From Basics to Advanced Modelling Techniques.* John Wiley & Sons. Inc.

Saidi, N., Scacciavillani, F., & Ali, F. (2010). *Forecasting Tourism in Dubai* Dubai International Financial Centre, Economic Note No. 8.

Smeral, E. (1988). Tourism Demand, Economic Theory and Econometrics: An Integrated Approach. *Journal of Travel Research, 26(4), 38-43.* https://doi.org/10.1177/004728758802600407.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting - A review of recent research. *Tourism Management, 29*(2), 203-220. https://doi.org/10.1016/j.tourman.2007.07.016.

Song, H., & Witt, S. F. (2000). *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches* (1st ed.) Elsevier Science Ltd.

Song, H., Dwyer, L., Li, G., & Cao, Z. (2012). Tourism Economics Research: A Review and Assessment. *Annals of Tourism Research, 39*(3), 1653-1682. https://doi.org/10.1016/j.annals.2012.05.023.

Song, H., Li, G., Witt F., S., & Fei, B. (2010). Tourism demand modelling and forecasting: how should demand be measured? *Tourism Economics, 16*(1), 63-81. https://doi.org/10.5367/000000010790872213.

Song, H., Smeral, E., Li, G., & Chen, J. L. (2008). *Tourism Forecasting: Accuracy of Alternative Econometric Models Revisited*. Vienna: Austrian Institute of Economic Research (WIFO).

Statista. (2022). Global search engine market share 2022. Retrieved from: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/.

Stock, J.H., & Watson, M.W. (2003). Forecasting Output and Inflation: The Role of Asset Prices. *Journal of Economic Literature, 41*(3), 788-829. https://doi.org/ 10.1257/002205103322436197.

Verbeek, M. (2004). *A Guide to Modern Econometrics* (2nd ed.). John Wiley & Sons Ltd.

Vosen, S., & Schmidt, T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting, 30*(6), 565-578. https://doi.org/10.1002/for.1213.

Wei., W. S. W. (2019). *Multivariate Time Series Analysis* (1st ed.). John Wiley & Sons Ltd.

Wen, L., Liu, C., & Song, H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. *Tourism Economics, 23*(3), 309-329. https://doi.org/10.1177/1354816618768317.

Yang, X., Pan, B., Evan, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386-397. https://doi.org/10.1016/j.tourman.2014.07.019.

APPENDIX A

Appendix A.1: Google Trends Data and Other Economic Variables

Choi & Varian (2009, 2012) were among the first to demonstrate the capacity of Google Trends data to predict the present or the very near future. Apart from applying Google Trends to predict tourist arrivals, they showed that this data source can be used for forecasting unemployment, consumption through retail and automobile sales and the housing market. Since then, researchers have followed suit and studied the application of this data in nowcasting and forecasting important variables of the economy.

The Labour Market

Google Trends data has been linked to the conditions of the labour market, especially the unemployment situation in an economy. The usefulness of this application is justified by the lag in the publication of related statistics by national offices. Thus, search queries related to job seeking and benefit registering may indicate a status of unemployment and can therefore indicate the current unemployment rate.

Askitas & Zimmermann (2009) and Choi & Varian (2009) find improvements in predicting unemployment in the USA and Germany, respectively, using Google Trends data. D'Amuri & Marcucci (2017) and Nagao et al. (2019) also do the same for the USA but arrive at different conclusions since the former find that Google data improves longer-term forecasts while the latter claim that the data does not provide significant contributions thereby demonstrating the limitations of Google Trends search data.

Anttonen (2018) uses Google Trends search query data in a seasonal Bayesian VAR to nowcast unemployment in the European Union and reports "modest improvements" (p. 5) in forecasts of three or more periods ahead. A variation of the model with PCA on the search data did not result into better forecasts than those from the model without such data.

Consumption

Vosen & Schmidt (2011) test the predictive power of Google Trends data in forecasting private consumption in the USA as opposed to that of the Michigan University's Consumer Sentiment Index and the Conference Board Consumer Confidence Index which are survey-based consumption indicators. They find that in-sample and out-of-sample performance of the Google-augmented model significantly dominates that of other competing models. In nowcasting consumption behaviour in Chile, Carrière-Swallow & Labbé (2013) examine the application of Google Trends data to car sales. Results are also in favour of models incorporating such data indicating that the potential of Google Trends is not only prevalent in the context of advanced economies, but also emerging ones such as that of Chile.

The Exchange Rate

Forecasts of different currencies vis-à-vis the US dollar are improved with the inclusion of Google search data (Bulut, 2018 and Ito et al., 2021). Specifically, the former finds that the true direction of changes of the nominal exchange rates is better predicted with the search data. In addition to this, Ito & Takeda (2022) find that although models with the Google sentiment index provide increased accuracy in the short-term prediction of the exchange rate, they must be updated with new search terms on a weekly basis.

APPENDIX B

Appendix B.1: Principal Component Analysis

Introduced by Hotelling (1933), PCA is a dimensionality reduction technique that is applied to a large number of variables in order to prevent overfitting problems and multicollinearity in a regression model. The operations of PCA are outlined following closely the explanation in Holland (2019), Rachev et al. (2007) and Wei (2019).

The Optimisation Problem

In a set of n time series X_i where $i = 1, \dots, n$, a linear combination of these series Y_i exists, identified by an n-vector of weights $\mathbf{w}_i$. The linear combinations Y_i are known as the principal components.

The first principal component, Y_1, is given by

$$Y_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1n}X_n \tag{B.1}$$

where $w_{11}, w_{12}, \dots, w_{1n}$ are the weights and their sum of squares is 1. In general, the i^{th} principal component is given by

$$Y_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{in}X_n, \tag{B.2}$$

where $\boldsymbol{w}_i = [w_{i1}, \dots, w_{in}]^{\mathrm{T}}$ are the weights for the i^{th} principal component. The first principal component has the largest possible variance of the dataset and the second principal component is obtained in the same way such that it accounts for the next largest possible variance.

In matrix notation, the transformation of the variables to principal components is given by

$$\mathbb{Y} = \mathbb{X}\mathbb{W} \tag{B.3}$$

where $\mathbb{Y}$ is the matrix of principal components, $\mathbb{X}$ is the data matrix and $\mathbb{W} = [\mathbf{w}_1, \dots, \mathbf{w}_n]^{\mathrm{T}}$ is the matrix of eigenvectors. These eigenvectors are obtained from the variance-covariance matrix of the original data, σ, and the elements of an eigenvector are the weights w_{ij} which are also known as the loadings. The elements in the diagonal of the variance-covariance matrix of the principal components, are the eigenvalues which depict the variance explained by each principal component and decrease monotonically from the first to the last principal component.

PCA results into non-unique solutions since the variance of each principal component Y_i may be changed by multiplying the weight with a constant. Therefore, the weights, w_i are restricted to be of unit lengths. In this regard, the first principal component is the linear combination $\boldsymbol{w}_1{}'\boldsymbol{X}$ that maximises the variance of $\boldsymbol{w}_1{}'\boldsymbol{X}$ subject to $\boldsymbol{w}_1{}'\boldsymbol{w}_1 = 1$ and the second principal component is the linear combination $\boldsymbol{w}_2{}'\boldsymbol{X}$ that maximises the variance of $\boldsymbol{w}_2{}'\boldsymbol{X}$ subject to $\boldsymbol{w}_2{}'\boldsymbol{w}_2 = 1$ and the covariance $(\boldsymbol{w}_1{}'\boldsymbol{X}, \boldsymbol{w}_2{}'\boldsymbol{X}) = 0$ which is the orthogonality condition. This means that in general, the i^{th} principal component is the linear combination $\boldsymbol{w}_i{}'\boldsymbol{X}$ that maximises the variance of $\boldsymbol{w}_i{}'\boldsymbol{X}$ subject to $\boldsymbol{w}_i{}'\boldsymbol{w}_i = 1$ and the covariance $(\boldsymbol{w}_i{}'\boldsymbol{X}, \boldsymbol{w}_n{}'\boldsymbol{X}) = 0$ for $n < i$.

Summary

Put simply, PCA involves the identification of the eigenvalues and eigenvectors of the variance-covariance matrix or the correlation matrix in order to obtain a linear combination of a set of n time series X_i. The eigenvectors are used as the weights in the n orthogonal linear combinations which are called the principal components. The variance of each principal component is equal to its respective eigenvector. This means that the first principal component, which is the linear combination of the variables associated with the first eigenvalue, has the maximum possible variance and on the other hand, the last principal component has the least possible variance.

The n principal components obtained from the PCA performed on the variance-covariance matrix or the correlation matrix are linear combinations of the original time series, $X = (X_1, \dots, X_N)'$. These are obtained from the multiplication of $\mathbb{X}$ by the matrix of eigenvectors $\mathbb{W}$.

The dimensionality reduction from PCA is a way to preserve a small number of components from a large dataset. Not all components require consideration as choosing those with the

largest variance is sufficient to explain a significant proportion of the total variance of X. Together, the components explain 100% of the total variance.

Appendix B.2: Variables Employed for Robustness Check

Real Gross Domestic Product

Quarterly GDP data for the period January 2004 to December 2019 was sourced from Eurostat for 34 European countries.[24] The data is neither seasonally adjusted nor calendar adjusted and is in chain linked volumes with 2010 being the base year. Although this data does not represent the income of all countries from which Malta receives tourists, it is still a suitable approximation since the majority of tourists in Malta are from European countries.[25] An increase in GDP reflects an increase in income therefore, tourist arrivals are expected to increase as demand for travel increases while a decrease in GDP results into less arrivals.

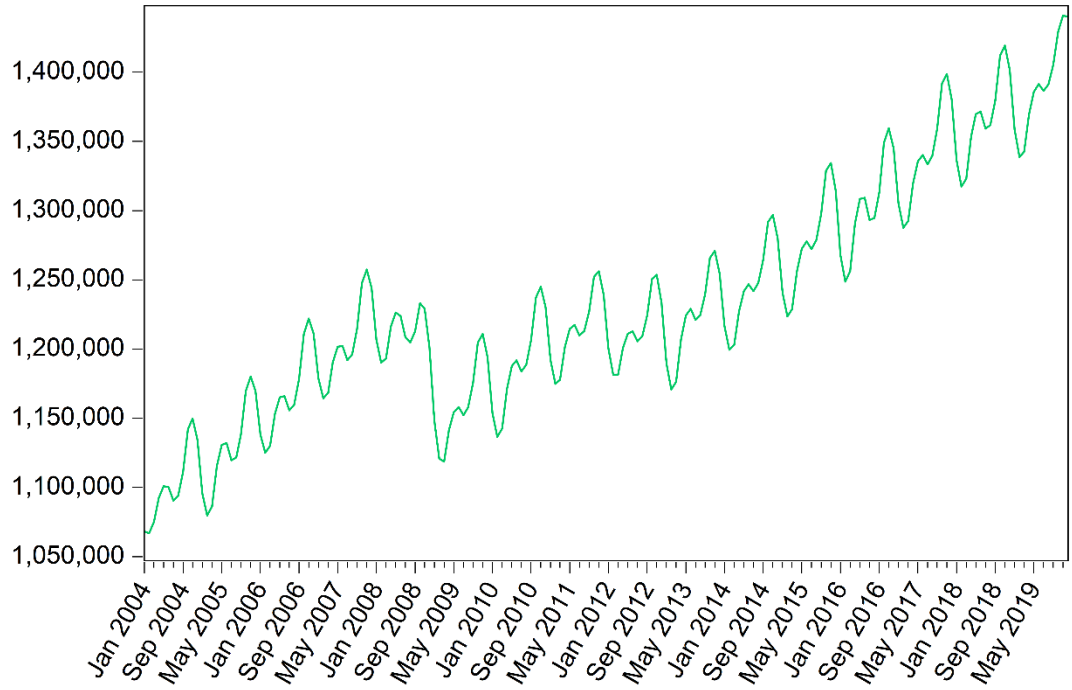
In order to obtain monthly GDP figures, the cubic spline interpolation method was used on the quarterly data. Cubic spline interpolation entails the fitting of cubic polynomials through the low frequency (quarterly) data wherein the polynomial coefficients are used to interpolate the high frequency (monthly) data. The spline deliberately passes through all the data points such that it is smooth and there are no discontinuities (see McKinley & Levine, n.d.).

Figure B.2.1 demonstrates the monthly interpolated GDP figures from January 2004 to December 2019. The plot shows a cyclical pattern with a break in the series from the last quarter of 2008 till the first quarter of 2009 due to the period of the Financial Crisis. Thereafter, the series exhibits an upward trend which suggests the presence of a unit root.

[24] GDP figures are for Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, Iceland, Norway, Switzerland, the UK, North Macedonia, Serbia, Turkey and Bosnia & Herzegovina.

[25] NSO inbound tourism data shows that on average between 2010 and 2020, tourists from Belgium, Denmark, France, Germany, Hungary, Ireland, Italy, Netherlands, Poland, Spain, Sweden, Switzerland and the UK made up 79.6% of the total annual arrivals during those years.

Figure B.2.1: Monthly GDP Interpolated from Quarterly Data Using Cubic Spline Interpolation (Jan 2004 – Dec 2019).



Source: Author's illustration.

Real Effective Exchange Rate

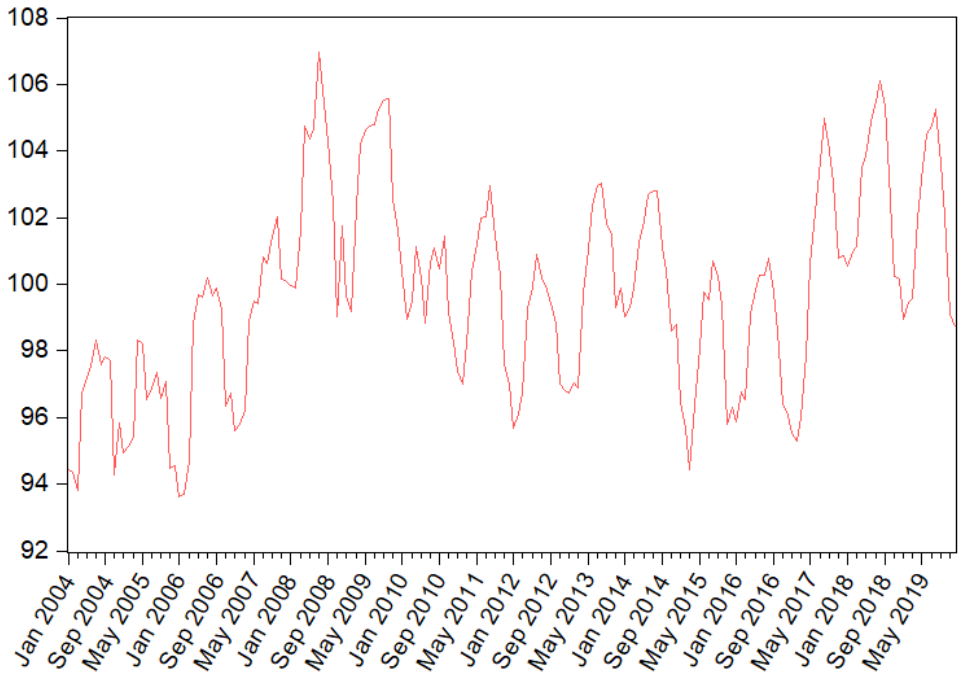
Monthly data for Malta's REER in relation to 36 trading partners[26] between January 2004 and December 2019 was obtained from Eurostat. The data is in the form of an index where 2010 is the base year and is thus equal to 100. An increase in the index signifies an appreciation in the currency which means that Maltese goods and services are more expensive relative to those of the trading partners. This means that demand for tourism in Malta should fall and hence the number of tourist arrivals declines. The converse applies for a depreciation in the currency implying a negative relationship between the real effective exchange rate and the number of inbound tourists.

[26] These are Australia, Austria, Belgium, Bulgaria, Canada, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, the UK and the USA.

Effective exchange rate data provides a comparable measure of a country's price and cost competitiveness in relation to exchange rate movements and cost and price trends. The real effective exchange rate is the nominal effective exchange rate deflated by the CPI and it is used to indicate a country's price and cost competitiveness comparative to competitors in the international markets (Eurostat, 2021). The review of the literature on tourism demand models revealed that the exchange rate is commonly used as a price variable.

The series displays troughs and peaks throughout the sample period as shown in Figure B.2.2 which shows the volatile nature of the exchange rate. Since the adoption of the euro in January 2008, the series is indicative of a change in the average value of the real effective exchange rate thereby also posing a potential issue to stationarity.

Figure B.2.2: Monthly Real Effective Exchange Rate (Jan 2004 – Dec 2019).



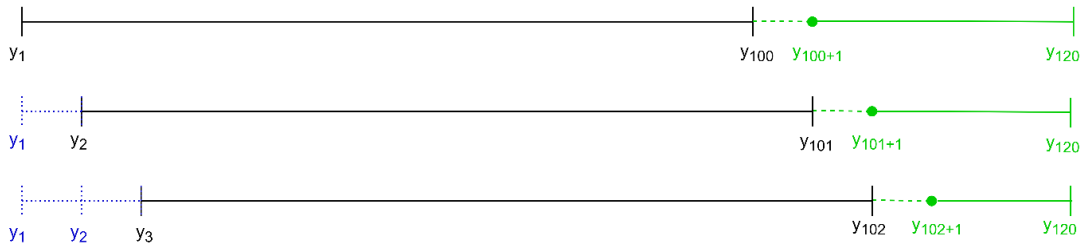
Source: Author's illustration.

Appendix B.3: Recursive Window Forecasting Approach

Diagram B.3.1 illustrates the rolling window approach on a hypothetical example of 120 observations of a time series y whereby the first 100 are used as the training sample (black solid line) and the last twenty comprise the testing sample (green solid line).

The forecast for y_{100+1} is obtained from the entire training sample, however, the forecast for y_{101+1} is obtained by dropping y_1 (blue dotted line) and adding the actual observation of y_{101} to the estimation window. The same is done for the forecast of y_{102+1} until the end of the forecast period thereby keeping the estimation window size constant with each one-step ahead forecast generated (Brooks, 2008 and Inoué et al. 2016). This approach to the pseudo-out-of-sample forecast simulation is suitable for the forecast problem in this dissertation as more recent observations are relevant for forecasting tourist arrivals in the short-term.

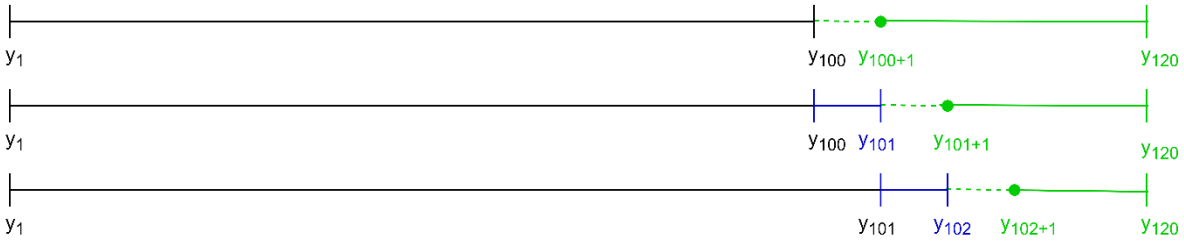
Diagram B.3.1: Rolling Window Forecasting Approach.



Source: Author's illustration.

Following the same hypothetical example, Diagram B.3.2 shows the recursive window forecasting approach. The first model is estimated on the first observation y_1 until y_{100} (black solid line) in order to obtain the one-step ahead forecast of y_{100+1} (green dashed line). The next one-step ahead forecast, y_{101+1} is obtained by estimating the model starting from the same first observation y_1 and adding an actual observation, y_{101} (blue solid line) to the estimation window. This means that the training sample is lengthened with each one-step ahead forecast produced until the entire testing sample up to observation y_{120} (green solid line) is forecasted.

Diagram B.3.2: Recursive Window Forecasting Approach.



Source: Author's illustration.

Appendix B.4: The Dickey-Fuller (1979) Test and the Ljung-Box (1978) Q-statistic

The Dickey-Fuller (1979) test is performed by the estimation of three possible regressions:

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t \tag{B.4}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \varepsilon_t \tag{B.5}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \alpha_2 t + \varepsilon_t \tag{B.6}$$

Equation (B. 4) is a random walk model, (B. 5) includes an intercept term and (B. 6) includes both an intercept and a linear time trend. All models are estimated using OLS. The null hypothesis is that $\gamma = 0$ and if accepted, it means that the series contains a unit root and is thus nonstationary. The resulting t-statistic is compared with the suitable value in the Dickey-Fuller table in order to reject or accept the null hypothesis.

The ADF test modifies the equations of the original Dickey-Fuller (1979) test by including a finite $AR(p)$ process of the Δy_t term. Selection of the lag length is done by the general-to-specific modelling procedure to obtain statistically significant autoregressive terms or using information criterion such as the AIC and SIC. The parameter of interest is γ and accepting or rejecting the null hypothesis that $\gamma = 0$ is done by comparing the resulting t-statistic with the suitable value in the Dickey-Fuller table (Enders, 2014).

The Ljung-Box (1978) Q-statistic is given by

$$Q = T(T+2)\sum_{k=1}^{s}\frac{r_k^2}{(T-k)} \tag{B. 7}$$

where T is the number of observations for y_t and r_k is a groups of autocorrelations. The null hypothesis is that $r_k = 0$. Q asymptotically follows a χ^2 distribution with s degrees of freedom. If the value of Q exceeds the critical value from this distribution, then at least one value of r_k is statistically different from zero at the relevant level of significance (Enders, 2014).

Appendix B.5: Search Queries Retrieved from Top and Rising Filters in Google Trends

Table B.5.1a: Queries Retrieved from Rising Filter in Google Trends.

11 seed queries and related queries by rising filter: queries that saw the largest increase in searches since the last period of time.

Malta Tourism	Malta Travel	Malta Hotel	Malta Restaurant	Malta Flights	Malta Airport	Malta Party	Malta Events	Malta Beach	Malta Weather	Malta Activities
malta tourism covid	travel republic	excelsior hotel	best restaurant malta	skyscanner	malta airport bus	boat party malta	valletta events	on the beach	windfinder malta	kids activities
malta covid	travel republic malta	excelsior hotel malta	restaurant sliema	flights to malta from gatwick	st julians malta	boat party	valletta malta	on the beach malta	windfinder	malta things to do
malta country	malta travel covid	tripadvisor	sliema malta	jet2	luton airport	hotel malta	valletta	best beach in malta	weather in malta in march	malta weather
malta island	malta covid	tripadvisor malta	sliema	jet2 flights	malta airport to valletta	party shop	malta events 2019	on the beach holidays	malta weather in march	things to do in malta
sicily tourism	ryanair	grand hotel excelsior malta	st julians restaurant	cheap flights to malta from london	malta airport to gozo	beach party malta	malta today	beach garden malta	malta weather in december	gozo
sicily	ryanair malta	grand hotel excelsior	indian restaurant malta	malta to catania flights	skyscanner	party time malta	events today	beach garden	ryanair	gozo malta
malta times	travel to malta covid	hotel bayview malta	best restaurant in malta	malta weather october	windfinder malta	party time	malta events 2017	hotels in malta	malta weather august	activities for kids

Table B.5.1b: Queries Retrieved from Rising Filter in Google Trends.

malta news	travel republic holidays	be hotel malta	gozo malta	flights to malta from luton	windfinder	party venues malta	malta party	hotel beach garden malta	time in malta	malta holidays
where is malta	rocs travel	radisson blu malta	gozo	things to do in malta	malta park	pool party malta	gozo events	beach garden hotel	maltapark	team building activities
malta time	rocs travel malta	trivago	malta weather	ibiza	malta airport to st julians	malta weather	malta events 2016	sliema malta	malta independent	malta today
malta beach	rocs	seashells hotel malta	best restaurants malta	santorini	cyprus	st julians malta	christmas malta	sliema beach	st julians malta	malta flights
malta tourism minister	rocs malta	the george hotel malta	malta food	airbnb malta	bugibba malta	paceville malta	times of malta	sliema beach malta	14 day weather forecast malta	outdoor activities
ryanair malta	can i travel to malta	hotel argento malta	tripadvisor	malta beaches	malta airport shuttle	paceville	fresh events malta	sliema	weather in malta in june	valletta malta
ryanair	malta travel requirements	trip advisor	tripadvisor malta	malta weather december	southend airport	europe	fresh events	beaches in malta	thomas cook	valletta
malta maps	malta coronavirus	booking.com malta	valletta restaurants	wizz air	malta airport lounge	malta events	fresh	marina hotel corinthia beach resort	skyscanner	malta hotels
times of malta	malta coronavirus travel	booking.com	best restaurants in malta	travel republic	airports in malta	party food	fresh malta	marina hotel corinthia beach resort malta	times of malta news	activities for children

Table B.5.1c: Queries Retrieved from Rising Filter in Google Trends.

malta beaches	tripadvisor	skyscanner	fish restaurant malta	thomson holidays	malta beaches	party supplies	which beach malta	malta weather 14 days
croatia tourism	skyscanner	times of malta	malte	st julians malta	things to do in malta	sliema malta	which beach	thomson
croatia	hays travel	beach garden hotel malta	times of malta	flights to malta from liverpool	emirates	party holidays	sunny beach	accuweather malta
tripadvisor malta	tui travel	cavalieri art hotel malta	marsaxlokk	malta weather april	munich airport	malta holidays	cheap holidays	weather malta ny
sliema malta	tui	travel republic	barracuda restaurant malta	flights to malta from cardiff	flights to malta from uk	malta news	cheap beach holidays	malta weather hourly
malta airlines	malta weather october	seashells resort at suncrest	barracuda malta	malta weather march	jet2	labour	tripadvisor	benidorm weather
malta park	times of malta	on the beach	barracuda restaurant	on the beach	malta public transport		tenerife	cape verde
malta food	things to do in malta	marina hotel corinthia beach resort	barracuda	malta weather november	taxi malta airport		malta temperature	weather in majorca
malta economy	malta beaches	radisson blu resort malta	hilton malta	weather in malta in october	ryanair		blue lagoon malta	malta weather site

Table B.5.2a: Queries Retrieved from Top Filter in Google Trends.

11 seed queries and related queries by top filter: queries sorted by popularity based on their value.

Malta Tourism	Malta Travel	Malta Hotel	Malta Restaurant	Malta Flights	Malta Airport	Malta Party	Malta Events	Malta Beach	Malta Weather	Malta Activities
tourism in malta	travel to malta	hotel in malta	restaurants malta	flights to malta	airport malta weather	malta labour party	events in malta	hotel malta	weather in malta	activities in malta
malta tourism authority	malta holidays	malta hotels	restaurants	flights from malta	malta weather	labour party	malta weather	beach in malta	malta weather forecast	kids activities
malta weather	holidays	hotels	valletta restaurant	malta flights cheap	airport weather	labour	valletta events	on the beach	weather forecast	malta things to do
malta map	malta weather	palace hotel malta	valletta	cheap flights	airport in malta	boat party malta	valletta malta	on the beach malta	weather of malta	malta weather
malta tourism map	travel republic	malta holidays	valletta malta	cheap flights to malta	malta airport arrivals	boat party	valletta	malta beach holidays	malta airport weather	gozo
malta travel	travel republic malta	hotel sliema malta	best restaurant malta	air malta	malta arrivals	hotel malta	malta events 2019	holidays	malta airport	gozo malta
malta hotel	malta flights	sliema	restaurants in malta	air malta flights	air malta	party shop malta	malta today	malta holidays	malta holidays	things to do in malta
malta tourist	flights	sliema hotel	sliema	ryanair	malta flights	party shop	events today	beach holidays	holidays	activities for kids
malta holidays	malta covid	sliema malta	restaurant sliema	ryanair malta	flights to malta	beach party malta	malta festival	malta weather	malta times	malta holidays
malta hotels	malta covid travel	holidays	sliema malta	ryanair flights	malta airport hotel	party time malta	malta events 2017	beach club malta	times	team building activities

Table B.5.2b: Queries Retrieved from Top Filter in Google Trends.

malta airport	flights to malta	weather malta	st julians restaurant	ryanair malta flights	hotel malta	party time	malta party	hotels malta	weather for malta	malta today
europe	malta air	grand hotel malta	st julians malta	malta holidays	ryanair	pool party malta	times	malta beach hotels	malta weather october	valletta malta
malta flights	malta travel restrictions	corinthia hotel malta	indian restaurant	malta airport	ryanair malta	party venues malta	malta events 2016	malta holiday	times of malta	valletta
valletta malta	malta holiday	corinthia hotel	indian restaurant malta	holidays	malta international airport weather	malta weather	gozo events	best beach in malta	bbc weather	outdoor activities
valletta	holidays to malta	corinthia malta	chinese restaurant	ryanair flights to malta	departures malta airport	st julians	christmas malta	malta beaches	bbc weather malta	malta flights
italy tourism	weather in malta	park hotel malta	chinese restaurant malta	london malta flights	flights from malta	paceville malta	times of malta	weather in malta	malta april weather	malta hotels
italy	malta airport	qawra malta	best restaurant in malta	cheap flights from malta	holidays	paceville	fresh	golden beach	the weather in malta	activities for children
malta italy	hotels malta	qawra	gozo malta	malta weather	malta holidays	st julians malta	fresh malta	malta golden beach	malta news	what to do in malta
visit malta	holidays in malta	hotels in malta	gozo	holidays to malta	malta airport bus	nationalist party malta	fresh events malta	holidays in malta	malta weather december	
malta covid	ryanair	dolmen hotel malta	malta weather	malta to london flights	luqa airport	nationalist party	fresh events	holidays to malta	malta november weather	

Table B.5.2c: Queries Retrieved from Top Filter in Google Trends.

flights to malta	ryanair malta	dolmen hotel	best restaurants malta	london to malta	malta luqa airport	times of malta	on the beach holidays	malta today
malta visa	travel insurance	valletta malta	malta food	easyjet	luqa malta	europe	beach garden malta	malta weather today
air malta	travel insurance malta	valletta	tripadvisor	easyjet flights	luqa	malta events	beach garden	malta weather march
gozo malta	travel to malta from uk	hilton malta hotel	tripadvisor malta	easyjet malta	malta times	party food	malta flights	holidays in malta
malta country	malta map	st julians malta	mdina	flight to malta	malta airport map	malta sliema	hotels in malta	malta weather february

Appendix B.6: Eliminated Search Queries

- Queries that do not contain the word *Malta*, *Gozo*, *Comino* or a locality in Malta or Gozo
- Queries related to Covid-19
- Queries unrelated to travelling or tourism
- Queries that refer to particular places or companies
- Queries that are very specific

Table B.6.1a: Queries Eliminated According to the 5 Characteristics.

11 seed queries and related queries by rising filter: queries that saw the largest increase in searches since the last period of time.

Malta Tourism	Malta Travel	Malta Hotel	Malta Restaurant	Malta Flights	Malta Airport	Malta Party	Malta Events	Malta Beach	Malta Weather	Malta Activities
malta tourism covid	travel republic	excelsior hotel	best restaurant malta	skyscanner	malta airport bus	boat party malta	valletta events	on the beach	windfinder malta	kids activities
malta covid	travel republic malta	excelsior hotel malta	restaurant sliema	flights to malta from gatwick	st julians malta	boat party	valletta malta	on the beach malta	windfinder	malta things to do
malta country	malta travel covid	tripadvisor	sliema malta	jet2	luton airport	hotel malta	valletta	best beach in malta	weather in malta in march	malta weather
malta island	malta covid	tripadvisor malta	sliema	jet2 flights	malta airport to valletta	party shop	malta events 2019	on the beach holidays	malta weather in march	things to do in malta
sicily tourism	ryanair	grand hotel excelsior malta	st julians restaurant	cheap flights to malta from london	malta airport to gozo	beach party malta	malta today	beach garden malta	malta weather in december	gozo
sicily	ryanair malta	grand hotel excelsior	indian restaurant malta	malta to catania flights	skyscanner	party time malta	events today	beach garden	ryanair	gozo malta

Table B.6.1b: Queries Eliminated According to the 5 Characteristics.

malta times	travel to malta covid	hotel bayview malta	best restaurant in malta	malta weather october	windfinder malta	party time	malta events 2017	hotels in malta	malta weather august	activities for kids
malta news	travel republic holidays	be hotel malta	gozo malta	flights to malta from luton	windfinder	party venues malta	malta party	hotel beach garden malta	time in malta	malta holidays
where is malta	rocs travel	radisson blu malta	gozo	things to do in malta	malta park	pool party malta	gozo events	beach garden hotel	maltapark	team building activities
malta time	rocs travel malta	trivago	malta weather	ibiza	malta airport to st julians	malta weather	malta events 2016	sliema malta	malta independent	malta today
malta beach	rocs	seashells hotel malta	best restaurants malta	santorini	cyprus	st julians malta	christmas malta	sliema beach	st julians malta	malta flights
malta tourism minister	rocs malta	the george hotel malta	malta food	airbnb malta	bugibba malta	paceville malta	times of malta	sliema beach malta	14 day weather forecast malta	outdoor activities
ryanair malta	can i travel to malta	hotel argento malta	tripadvisor	malta beaches	malta airport shuttle	paceville	fresh events malta	sliema	weather in malta in june	valletta malta
ryanair	malta travel requirements	trip advisor	tripadvisor malta	malta weather december	southend airport	europe	fresh events	beaches in malta	thomas cook	valletta
malta maps	malta coronavirus	booking.com malta	valletta restaurants	wizz air	malta airport lounge	malta events	fresh	marina hotel corinthia beach resort	skyscanner	malta hotels
times of malta	malta coronavirus travel	booking.com	best restaurants in malta	travel republic	airports in malta	party food	fresh malta	marina hotel corinthia beach resort malta	times of malta news	activities for children
malta beaches	tripadvisor	skyscanner	fish restaurant malta	thomson holidays	malta beaches	party supplies		which beach malta	malta weather 14 days	
croatia tourism	skyscanner	times of malta	malte	st julians malta	things to do in malta	sliema malta		which beach	thomson	

Table B.6.1c: Queries Eliminated According to the 5 Characteristics.

croatia	hays travel	beach garden hotel malta	times of malta	flights to malta from liverpool	emirates	party holidays		sunny beach	accuweather malta
tripadvisor malta	tui travel	cavalieri art hotel malta	marsaxlokk	malta weather april	munich airport	malta holidays		cheap holidays	weather malta ny
sliema malta	tui	travel republic	barracuda restaurant malta	flights to malta from cardiff	flights to malta from uk	malta news		cheap beach holidays	malta weather hourly
malta airlines	malta weather october	seashells resort at suncrest	barracuda malta	malta weather march	jet2	labour		tripadvisor	benidorm weather
malta park	times of malta	on the beach	barracuda restaurant	on the beach	malta public transport			tenerife	cape verde
malta food	things to do in malta	marina hotel corinthia beach resort	barracuda	malta weather november	taxi malta airport			malta temperature	weather in majorca
malta economy	malta beaches	radisson blu resort malta	hilton malta	weather in malta in october	ryanair			blue lagoon malta	malta weather site

Table B.6.2a: Queries Eliminated According to the 5 Characteristics.

11 seed queries and related queries by **top** filter: queries sorted by popularity based on their value.

Malta Tourism	Malta Travel	Malta Hotel	Malta Restaurant	Malta Flights	Malta Airport	Malta Party	Malta Events	Malta Beach	Malta Weather	Malta Activities
tourism in malta	travel to malta	hotel in malta	restaurants malta	flights to malta	airport malta weather	malta labour party	events in malta	hotel malta	weather in malta	activities in malta
malta tourism authority	malta holidays	malta hotels	restaurants	flights from malta	malta weather	labour party	malta weather	beach in malta	malta weather forecast	kids activities
malta weather	holidays	hotels	valletta restaurant	malta flights cheap	airport weather	labour	valletta events	on the beach	weather forecast	malta things to do
malta map	malta weather	palace hotel malta	valletta	cheap flights	airport in malta	boat party malta	valletta malta	on the beach malta	weather of malta	malta weather
malta tourism map	travel republic	malta holidays	valletta malta	cheap flights to malta	malta airport arrivals	boat party	valletta	malta beach holidays	malta airport weather	gozo
malta travel	travel republic malta	hotel sliema malta	best restaurant malta	air malta	malta arrivals	hotel malta	malta events 2019	holidays	malta airport	gozo malta
malta hotel	malta flights	sliema	restaurants in malta	air malta flights	air malta	party shop malta	malta today	malta holidays	malta holidays	things to do in malta
malta tourist	flights	sliema hotel	sliema	ryanair	malta flights	party shop	events today	beach holidays	holidays	activities for kids
malta holidays	malta covid	sliema malta	restaurant sliema	ryanair malta	flights to malta	beach party malta	malta festival	malta weather	malta times	malta holidays
malta hotels	malta covid travel	holidays	sliema malta	ryanair flights	malta airport hotel	party time malta	malta events 2017	beach club malta	times	team building activities

Table B.6.2b: Queries Eliminated According to the 5 Characteristics.

malta airport	flights to malta	weather malta	st julians restaurant	ryanair malta flights	hotel malta	party time	malta party	hotels malta	weather for malta	malta today
europe	malta air	grand hotel malta	st julians malta	malta holidays	ryanair	pool party malta	times	malta beach hotels	malta weather october	valletta malta
malta flights	malta travel restrictions	corinthia hotel malta	indian restaurant	malta airport	ryanair malta	party venues malta	malta events 2016	malta holiday	times of malta	valletta
valletta malta	malta holiday	corinthia hotel	indian restaurant malta	holidays	malta international airport weather	malta weather	gozo events	best beach in malta	bbc weather	outdoor activities
valletta	holidays to malta	corinthia malta	chinese restaurant	ryanair flights to malta	departures malta airport	st julians	christmas malta	malta beaches	bbc weather malta	malta flights
italy tourism	weather in malta	park hotel malta	chinese restaurant malta	london malta flights	flights from malta	paceville malta	times of malta	weather in malta	malta april weather	malta hotels
italy	malta airport	qawra malta	best restaurant in malta	cheap flights from malta	holidays	paceville	fresh	golden beach	the weather in malta	activities for children
malta italy	hotels malta	qawra	gozo malta	malta weather	malta holidays	st julians malta	fresh malta	malta golden beach	malta news	what to do in malta
visit malta	holidays in malta	hotels in malta	gozo	holidays to malta	malta airport bus	nationalist party malta	fresh events malta	holidays in malta	malta weather december	
malta covid	ryanair	dolmen hotel malta	malta weather	malta to london flights	luqa airport	nationalist party	fresh events	holidays to malta	malta november weather	
flights to malta	ryanair malta	dolmen hotel	best restaurants malta	london to malta	malta luqa airport	times of malta		on the beach holidays	malta today	

able B.6.2c: Queries Eliminated According to the 5 Characteristics.

malta visa	travel insurance	valletta malta	malta food	easyjet	luqa malta	europe		beach garden malta	malta weather today
air malta	travel insurance malta	valletta	tripadvisor	easyjet flights	luqa	malta events		beach garden	malta weather march
gozo malta	travel to malta from uk	hilton malta hotel	tripadvisor malta	easyjet malta	malta times	party food		malta flights	holidays in malta
malta country	malta map	st julians malta	mdina	flight to malta	malta airport map	malta sliema		hotels in malta	malta weather february

APPENDIX C

Appendix C.1: Results from Correlation Analysis

Table C.1.1a: Results from Correlation Analysis.

Lags	bugibba malta	cheap flights to malta	flight to malta	flights from malta	flights to malta	gozo
-7	-0.194	-0.260	0.077	0.066	-0.236	-0.096
-6	-0.365	-0.334	0.043	0.004	-0.284	-0.210
-5	-0.399	-0.383	0.025	-0.024	-0.306	-0.249
-4	-0.317	-0.406	0.081	-0.002	-0.291	-0.173
-3	-0.142	-0.366	0.157	0.072	-0.231	-0.021
-2	0.007	-0.349	0.234	0.102	-0.225	0.129
-1	0.176	-0.223	0.337	0.300	-0.055	0.271
Highest absolute correlation	0.399	0.406	0.337	0.300	0.306	0.271
Lag of highest absolute correlation	-5	-4	-1	-1	-5	-1

Lags	gozo malta	holidays in malta	holidays to malta	hotel in malta	hotels in malta	luqa
-7	-0.122	-0.175	-0.094	-0.301	-0.242	0.251
-6	-0.279	-0.259	-0.242	-0.385	-0.317	0.241
-5	-0.338	-0.276	-0.314	-0.423	-0.352	0.245
-4	-0.286	-0.249	-0.293	-0.393	-0.329	0.278
-3	-0.137	-0.187	-0.203	-0.312	-0.277	0.332
-2	0.035	-0.154	-0.150	-0.238	-0.208	0.375
-1	0.237	-0.009	0.051	-0.095	-0.107	0.416
Highest absolute correlation	0.338	0.276	0.314	0.423	0.352	0.416
Lag of highest absolute correlation	-5	-5	-5	-5	-5	-1

Table C.1.1b: Results from Correlation Analysis.

Lags	malta air	malta airlines	malta airport	malta beach	malta beaches	malta country
-7	-0.254	-0.218	-0.155	-0.061	-0.035	0.182
-6	-0.306	-0.271	-0.217	-0.203	-0.187	0.193
-5	-0.318	-0.298	-0.182	-0.272	-0.284	0.228
-4	-0.284	-0.269	-0.099	-0.210	-0.286	0.225
-3	-0.213	-0.210	0.016	-0.068	-0.170	0.233
-2	-0.138	-0.157	0.100	0.089	0.005	0.230
-1	-0.008	-0.063	0.233	0.296	0.207	0.234
Highest absolute correlation	0.318	0.298	0.233	0.296	0.286	0.234
Lag of highest absolute correlation	-5	-5	-1	-1	-4	-1

Lags	malta festival	malta flights	malta flights cheap	malta food	malta holiday	malta holidays
-7	-0.134	-0.233	-0.282	0.139	-0.246	-0.202
-6	-0.294	-0.274	-0.338	0.074	-0.367	-0.300
-5	-0.438	-0.291	-0.389	0.049	-0.416	-0.351
-4	-0.424	-0.276	-0.402	0.053	-0.382	-0.328
-3	-0.275	-0.224	-0.373	0.056	-0.292	-0.260
-2	-0.084	-0.206	-0.357	0.092	-0.250	-0.241
-1	0.068	-0.067	-0.236	0.159	-0.100	-0.092
Highest absolute correlation	0.438	0.291	0.402	0.159	0.416	0.351
Lag of highest absolute correlation	-5	-5	-4	-1	-5	-5

Table C.1.1c: Results from Correlation Analysis.

Lags	malta hotel	malta hotels	malta independent	malta island	malta map	malta maps
-7	-0.245	-0.263	-0.024	-0.005	-0.345	0.154
-6	-0.403	-0.337	-0.038	-0.053	-0.389	0.097
-5	-0.484	-0.368	-0.037	-0.107	-0.420	0.070
-4	-0.459	-0.363	-0.026	-0.162	-0.414	0.084
-3	-0.336	-0.329	0.006	-0.096	-0.395	0.122
-2	-0.232	-0.308	0.040	-0.080	-0.366	0.181
-1	-0.046	-0.203	0.072	0.037	-0.277	0.263
Highest absolute correlation	0.484	0.368	0.072	0.162	0.420	0.263
Lag of highest absolute correlation	-5	-5	-1	-4	-5	-1

Lags	malta news	malta park	malta time	malta times	malta today	malta tourism
-7	0.202	0.140	0.163	0.215	0.132	-0.067
-6	0.232	0.081	0.151	0.215	0.135	-0.083
-5	0.258	0.045	0.177	0.218	0.143	-0.133
-4	0.291	0.036	0.216	0.220	0.147	-0.146
-3	0.314	0.058	0.296	0.221	0.152	-0.160
-2	0.335	0.091	0.353	0.208	0.162	-0.186
-1	0.337	0.129	0.411	0.216	0.169	-0.166
Highest absolute correlation	0.337	0.140	0.411	0.221	0.169	0.186
Lag of highest absolute correlation	-1	-7	-1	-3	-1	-2

Table C.1.1d: Results from Correlation Analysis.

Lags	malta tourist	malta travel	malta visa	malta weather forecast	mdina	paceville
-7	-0.317	-0.222	0.002	0.158	0.066	-0.014
-6	-0.354	-0.324	-0.029	0.230	0.065	-0.179
-5	-0.397	-0.379	-0.039	0.302	0.087	-0.194
-4	-0.403	-0.388	-0.016	0.280	0.157	-0.114
-3	-0.406	-0.311	0.025	0.208	0.212	0.102
-2	-0.407	-0.271	0.090	0.199	0.306	0.225
-1	-0.372	-0.165	0.105	0.165	0.370	0.392
Highest absolute correlation	0.407	0.388	0.105	0.302	0.370	0.392
Lag of highest absolute correlation	-2	-4	-1	-5	-1	-1

Lags	party malta	qawra	qawra malta	restaurants malta	sliema	sliema hotel
-7	-0.006	-0.065	-0.029	0.134	-0.056	-0.204
-6	-0.086	-0.238	-0.176	0.124	-0.173	-0.337
-5	-0.099	-0.307	-0.227	0.152	-0.188	-0.376
-4	-0.118	-0.263	-0.211	0.197	-0.114	-0.304
-3	-0.075	-0.135	-0.102	0.234	0.031	-0.201
-2	-0.024	-0.022	-0.024	0.295	0.137	-0.161
-1	0.051	0.167	0.143	0.338	0.281	-0.040
Highest absolute correlation	0.118	0.307	0.227	0.338	0.281	0.376
Lag of highest absolute correlation	-4	-5	-5	-1	-1	-5

Table C.1.1e: Results from Correlation Analysis.

Lags	sliema malta	st julians	st julians malta	times of malta	valletta	valletta malta
-7	-0.150	0.153	0.107	0.207	0.145	0.196
-6	-0.316	0.035	-0.022	0.207	0.142	0.161
-5	-0.357	0.015	-0.087	0.213	0.147	0.148
-4	-0.303	0.017	-0.077	0.218	0.164	0.168
-3	-0.127	0.129	0.030	0.221	0.193	0.233
-2	0.001	0.225	0.141	0.210	0.226	0.312
-1	0.186	0.387	0.323	0.216	0.289	0.383
Highest absolute correlation	0.357	0.387	0.323	0.221	0.289	0.383
Lag of highest absolute correlation	-5	-1	-1	-3	-1	-1

Lags	visit malta	weather for malta	weather in malta	weather malta	where is malta
-7	0.155	-0.018	0.198	0.252	0.160
-6	0.095	0.085	0.324	0.402	0.148
-5	0.053	0.242	0.437	0.503	0.177
-4	0.064	0.217	0.471	0.511	0.192
-3	0.092	0.245	0.452	0.446	0.232
-2	0.157	0.238	0.402	0.381	0.233
-1	0.223	0.226	0.377	0.310	0.273
Highest absolute correlation	0.223	0.245	0.471	0.511	0.273
Lag of highest absolute correlation	-1	-3	-4	-4	-1

Appendix C.2: List of 33 Queries in the Final Google Trends Dataset

1. bugibba malta
2. cheap flights to malta
3. flight to malta
4. flights from malta
5. flights to malta
6. gozo malta
7. holidays to malta
8. hotel in malta
9. hotels in malta
10. luqa
11. malta air
12. malta festival
13. malta flights cheap
14. malta holiday
15. malta holidays
16. malta hotel
17. malta hotels
18. malta map
19. malta new
20. malta time
21. malta travel
22. malta weather forecast
23. mdina
24. paceville
25. qawra
26. restaurants malta
27. sliema hotel
28. sliema malta
29. st julians
30. st julians malta
31. valletta malta
32. weather in malta
33. weather malta

Appendix C.3: Stationarity Results from ADF Test on Tourist Arrivals and First Principal Components

Table C.3.1: ADF Test on Arrivals Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: ARRIVALS has a unit root			Null Hypothesis: D(ARRIVALS) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-1.31	0.88	-3.22	0.08
Test critical values:	1% level	-4.02		-4.02	
	5% level	-3.44		-3.44	
	10% level	-3.15		-3.15	

.

Table C.3.2: ADF Test on the Log of Arrivals Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: LOG(ARRIVALS) has a unit root			Null Hypothesis: D(LOG(ARRIVALS)) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-2.03	0.58	-2.16	0.03
Test critical values:	1% level	-4.02		-2.58	
	5% level	-3.44		-1.94	
	10% level	-3.15		-1.62	

Table C.3.3: ADF Test on the First Principal Component of the 5-Month Lagged Dataset in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: PC1_Lag_5 has a unit root			Null Hypothesis: D(PC1_LAG_5) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-2.23	0.47	-4.20	0.01
Test critical values:	1% level	-4.01		-4.01	
	5% level	-3.44		-3.44	
	10% level	-3.14		-3.14	

Table C.3.4: ADF Test on the First Principal Component of the 4-Month Lagged Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: PC1_LAG_4 has a unit root			Null Hypothesis: D(PC1_LAG_4) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-1.46	0.84	-16.62	0.00
Test critical values:	1% level	-4.01		-4.01	
	5% level	-3.44		-3.44	
	10% level	-3.14		-3.14	

.

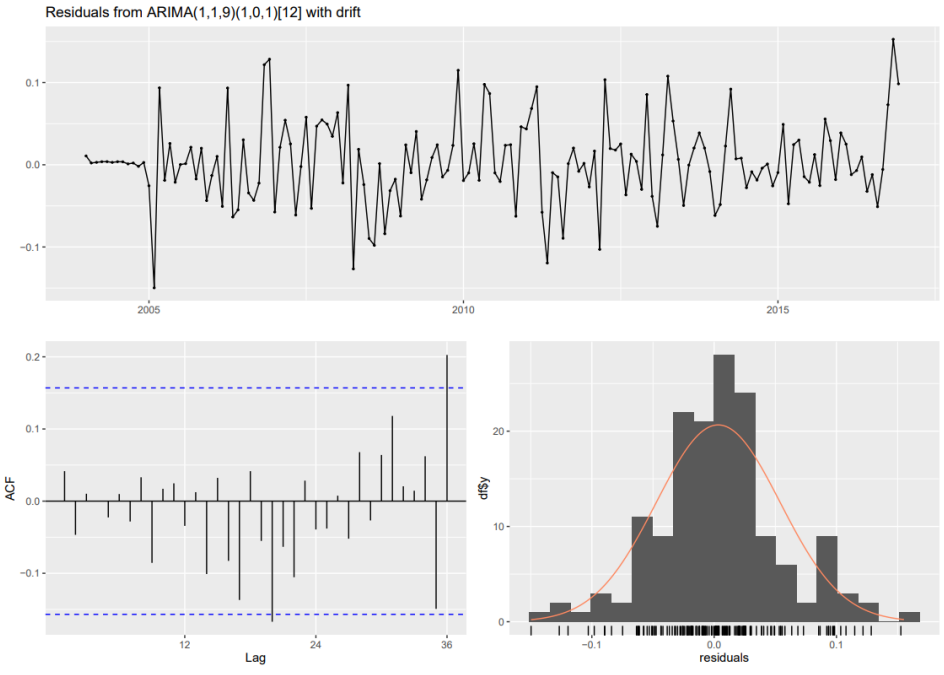
Table C.3.5: ADF Test on the First Principal Component of the 1-Month Lagged Dataset in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: PC1_LAG_1 has a unit root			Null Hypothesis: D(PC1_LAG_1) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-0.92	0.95	-4.87	0.00
Test critical values:	1% level	-4.01		-4.01	
	5% level	-3.44		-3.44	
	10% level	-3.14		-3.14	

.

Appendix C.4: Diagnostic Test Results on the $SARIMA(1,1,9)(1,0,1)$

Figure C.4.1: Residual Plot, ACF and PACF for $SARIMA(1,1,9)(1,0,1)$ Estimating on Training Sample.



Source: Author's calculations.

Table C.4.1: Ljung-Box Test Results on $SARIMA(1,1,9)(1,0,1)$ Estimated on Training Sample.

Q* value	Degrees of freedom	p-value	Model degrees of freedom	Total lags used
37.536	23	0.029	13	36

Appendix C.5: Adjustment to Training Sample in the Estimation of $SARIMA(1,1,0)(0,1,1)$ Model

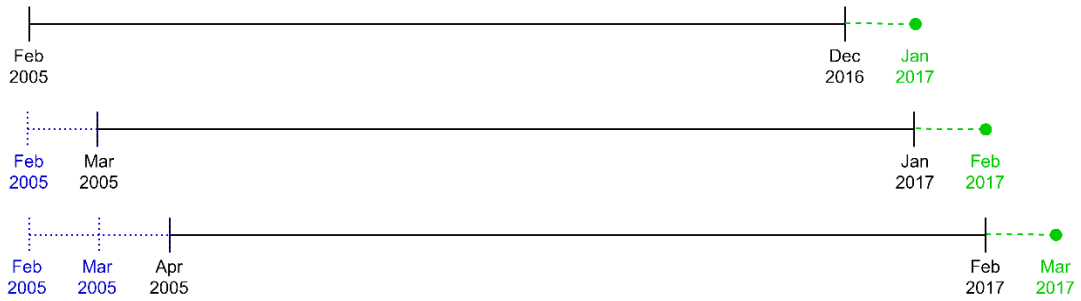
Due to seasonal differencing and differencing of the tourist arrival series as dictated by the $SARIMA(1,1,0)(0,1,1)$ model, the training sample starts from February 2005 and ends in December 2016 as shown in Diagram C.5.1. This is because twelve observations are lost due to seasonal differencing and another observation is lost because of non-seasonal differencing. Therefore, the observations from January 2004 till January 2005 (red dotted line) are lost. The testing sample is the same as described in chapter three, that is, beginning from January 2017 and ending December 2019 (green dashed line).

Diagram C.5.1: Adjustment to Training Sample.

Source: Author's illustration.

In view of this, the training window in each model estimated is made up of 143 observations as shown by the black solid line in Diagram C.5.2. With each forecasted month (green dashed line), the training window is rolled forward by one month such that an observation from the beginning of the training set is dropped (blue dotted line) and an actual observation is added to the end of the training set.

Diagram C.5.2: One-step Ahead Rolling Window Forecasting Methodology.



Source: Author's illustration.

Appendix C.6: Stationarity Results from ADF Test on GDP and the REER

Table C.6.1: ADF Test on GDP Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: REALGDP has a unit root			Null Hypothesis: D(REALGDP) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-2.97	0.14	-3.06	0.12
Test critical values:	1% level	-4.02		-4.02	
	5% level	-3.44		-3.44	
	10% level	-3.15		-3.15	

Table C.6.2: ADF Test on the log of GDP Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A		Panel B	
	Null Hypothesis: LOGREALGDP has a unit root		Null Hypothesis: D(LOGREALGDP) has a unit root	
	t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic	-2.90	0.17	-1.95	0.049
Test critical values: 1% level	-4.02		-2.58	
5% level	-3.44		-1.94	
10% level	-3.15		-1.62	

Table C.6.3: ADF Test on REER Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A		Panel B	
	Null Hypothesis: REER has a unit root		Null Hypothesis: D(REER) has a unit root	
	t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic	-2.38	0.15	-3.00	0.04
Test critical values: 1% level	-3.48		-3.48	
5% level	-2.88		-2.88	
10% level	-2.58		-2.58	

Table C.6.4: ADF Test on the log of REER Series in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: LOGREER has a unit root			Null Hypothesis: D(LOGREER) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-2.38	0.15	-3.00	0.04
Test critical values:	1% level	-3.48		-3.48	
	5% level	-2.88		-2.88	
	10% level	-2.58		-2.58	

Appendix C.7: Results from Correlation Analysis, PCA, Stationarity Tests and Model Estimation for the Nowcasting Experiment

Table C.7.1a: Results from Correlation Analysis.

Lags	flights to malta	gozo	gozo malta	holidays in malta	hotel in malta	malta air
-7	-0.106	0.048	0.033	-0.144	-0.270	-0.305
-6	-0.105	-0.027	-0.071	-0.135	-0.274	-0.335
-5	-0.101	-0.039	-0.097	-0.105	-0.276	-0.342
-4	-0.177	0.010	-0.079	-0.156	-0.283	-0.359
-3	-0.145	0.151	0.053	-0.117	-0.205	-0.299
-2	-0.137	0.324	0.228	-0.128	-0.167	-0.223
-1	-0.007	0.538	0.465	-0.017	-0.026	-0.096
Highest absolute correlation	0.177	0.538	0.465	0.156	0.283	0.359
Lag of highest absolute correlation	-4	-1	-1	-4	-4	-4

Table C.7.1b: Results from Correlation Analysis.

Lags	malta airport	malta Flights	malta flights cheap	malta holiday	malta holidays	malta hotel
-7	-0.189	-0.116	-0.235	-0.171	-0.194	-0.236
-6	-0.187	-0.120	-0.255	-0.209	-0.211	-0.308
-5	-0.142	-0.119	-0.285	-0.218	-0.207	-0.342
-4	-0.107	-0.184	-0.324	-0.264	-0.273	-0.349
-3	-0.006	-0.156	-0.300	-0.208	-0.232	-0.256
-2	0.101	-0.145	-0.292	-0.193	-0.224	-0.163
-1	0.251	-0.033	-0.207	-0.060	-0.100	0.016
Highest absolute correlation	0.251	0.184	0.324	0.264	0.273	0.349
Lag of highest absolute correlation	-1	-4	-4	-4	-4	-4

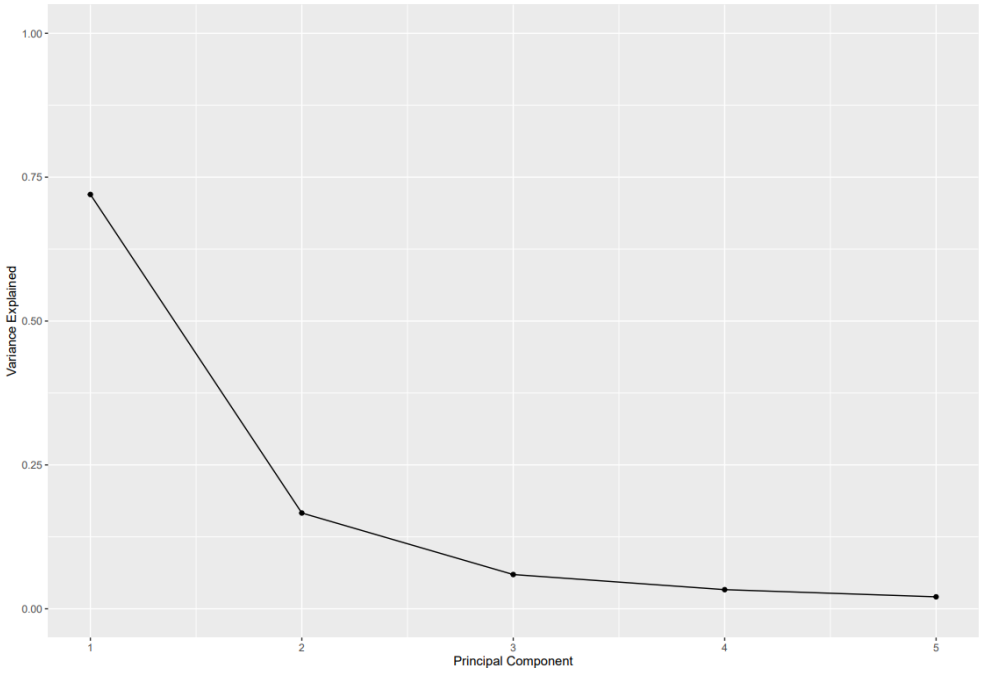
Table C.7.1c: Results from Correlation Analysis.

Lags	malta hotels	malta independent	malta map	malta times	malta tourism	malta travel
-7	-0.279	0.397	-0.202	0.151	-0.231	-0.342
-6	-0.299	0.349	-0.210	0.114	-0.241	-0.388
-5	-0.306	0.304	-0.220	0.079	-0.281	-0.422
-4	-0.343	0.278	-0.239	0.036	-0.315	-0.483
-3	-0.312	0.267	-0.209	0.027	-0.325	-0.449
-2	-0.300	0.282	-0.175	0.032	-0.355	-0.397
-1	-0.205	0.303	-0.095	0.047	-0.314	-0.273
Highest absolute correlation	0.343	0.397	0.239	0.151	0.355	0.483
Lag of highest absolute correlation	-4	-7	-4	-7	-2	-4

Table C.7.1d: Results from Correlation Analysis.

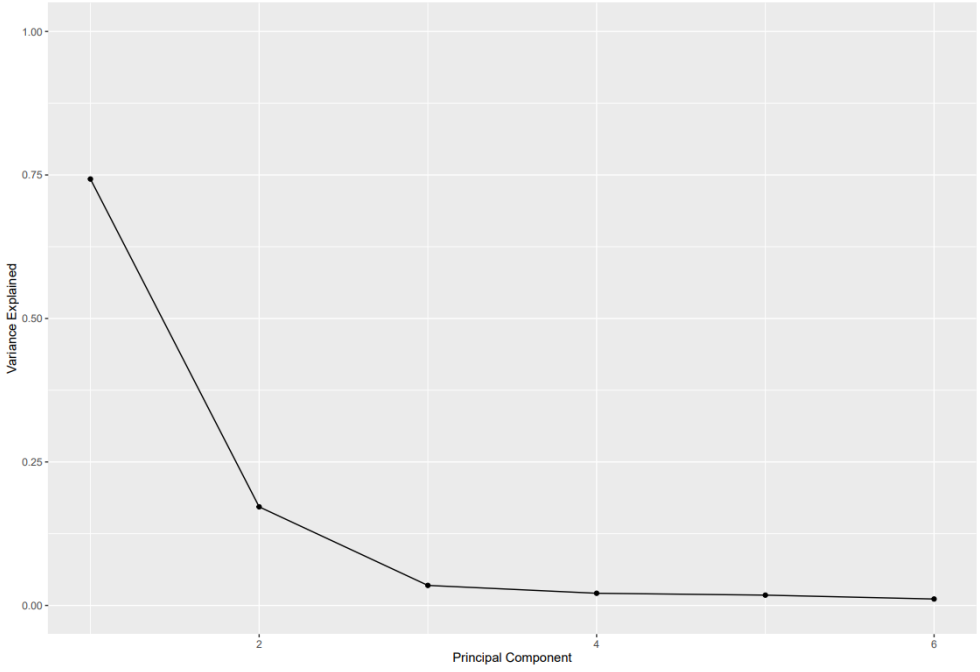
Lags	mdina	sliema	times of malta	valletta	valletta malta	weather in malta	weather malta
-7	0.284	0.170	0.155	0.312	0.358	0.300	0.270
-6	0.296	0.117	0.130	0.349	0.372	0.473	0.439
-5	0.316	0.112	0.103	0.337	0.373	0.574	0.517
-4	0.366	0.162	0.069	0.348	0.391	0.565	0.511
-3	0.463	0.272	0.065	0.386	0.460	0.570	0.501
-2	0.566	0.409	0.074	0.437	0.552	0.532	0.489
-1	0.700	0.596	0.094	0.545	0.673	0.530	0.466
Highest absolute correlation	0.700	0.596	0.155	0.545	0.673	0.574	0.517
Lag of highest absolute correlation	-1	-1	-7	-1	-1	-5	-5

Figure C.7.1: Scree Plot of 5 Principal Components from the PCA on Dataset Lagged by 4 Months.



Source: Author's calculations.

Figure C.7.2: Scree Plot of 6 Principal Components from the PCA on Dataset Lagged by 1 Month.



Source: Author's calculations.

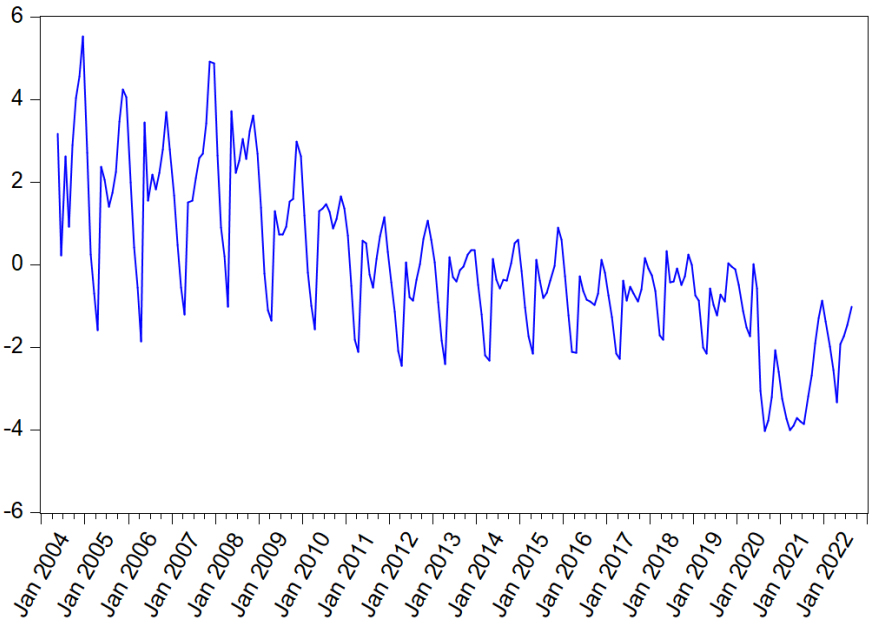
Table C.7.2: Loadings of the First Principal Component for the Dataset Lagged by 1 Month.

	Loading
flights to malta	-0.43
gozo	-0.07
malta holiday	-0.46
malta holidays	-0.46
malta hotel	-0.45
malta hotels	-0.43

Table C.7.3: Loadings of the First Principal Component for the Dataset Lagged by 4 Months.

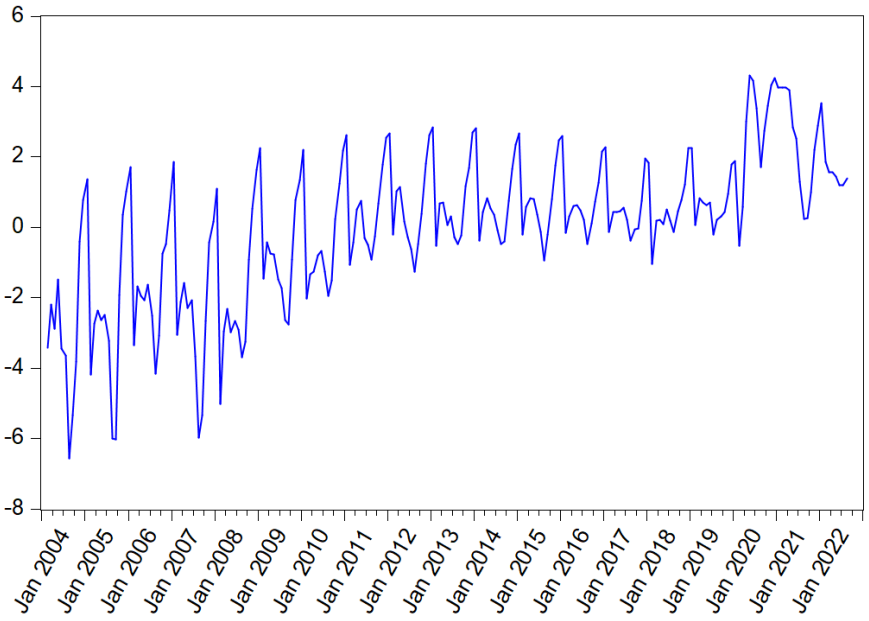
	Loading
gozo malta	0.29
holidays in malta	0.46
hotel in malta	0.49
malta air	0.49
malta flights cheap	0.47

Figure C.7.3: Time Series Plot of the First Principal Component of the Lagged by 4 Months Dataset.



Source: Author's calculations.

Figure C.7.4: Time Series Plot of the First Principal Component of the Lagged by 1 Month Dataset.



Source: Author's calculations.

Table C.7.4: ADF Test on the First Principal Component of the 1 Month Lagged Dataset in Levels (Panel A) and in First Difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: PC1Lag1 has a unit root			Null Hypothesis: D(PC1Lag1) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-3.27	0.07	-4.67	0.00
Test critical values:	1% level	-4.00		-4.00	
	5% level	-3.43		-3.43	
	10% level	-3.14		-3.14	

Table C.7.5: ADF test on the first principal component of the 4 month lagged dataset in levels (Panel A) and in first difference (Panel B).

	Panel A			Panel B	
	Null Hypothesis: PC1Lag4 has a unit root			Null Hypothesis: D(PC1Lag4) has a unit root	
		t-Statistic	Prob.	t-Statistic	Prob.
Augmented Dickey-Fuller test statistic		-3.68	0.03	-4.86	0.00
Test critical values:	1% level	-4.00		-4.00	
	5% level	-3.43		-3.43	
	10% level	-3.14		-3.14	

Table C.7.6: Google Trends Model with Dummy Variable for Nowcasting Experiment Estimation Results.

	AR1	SMA1	D(PC1 Lag 4)	D(PC1 Lag 1)	Covid Dummy
Coefficient	0.1886***	-0.8593***	-0.0020	-0.0130	-17.8185***
Standard error	0.0776	0.0674	0.0109	0.0123	0.1325
p-value	0.015	0.000	0.854	0.292	0.000

*** denotes statistical significance at the 0.01 level.