Departamento de Matemáticas, Estadística
e Investigación Operativa

Elisa María Jorge González

Analysis and development of a tourism demand forecast system at the Canary Islands

Análisis y desarrollo de un sistema de predicción de demanda turística en Canarias

Under the guidance of:
Enrique González Dávila

2019

Universidad de La Laguna
Oficina de Sede Electrónica
Entrada
Nº registro: 2019/75736
Nº reg. oficina: OF002/2019/72835
Fecha: 20/09/2019 09:32:19

Elisa María Jorge González

Analysis and development of a tourism demand forecast system at the Canary Islands

Análisis y desarrollo de un sistema de predicción de demanda turística en Canarias

Director:
Enrique González Dávila

2019

D. Enrique Francisco González Dávila, Profesor Titular de la Universidad de La Laguna del Departamento de Matemáticas, Estadística e Investigación Operativa de la Universidad de La Laguna,

CERTIFICA:

Que la presente memoria, titulada "Análisis y desarrollo de un sistema de predicción de demanda turística en Canarias" (en inglés "Analysis and development of a tourism demand forecast system at the Canary Islands"), ha sido realizada bajo mi dirección por Dña. Elisa María Jorge González y constituye su Tesis para optar al grado de Doctor en Matemáticas y Estadística por la Universidad de La Laguna.

Y para que conste, en cumplimiento de la legislación vigente, y a efectos que hayan lugar, firmo la presente, en La Laguna, a 19 de septiembre de 2019.

Enrique Francisco González Dávila

Contents

List of Figures

List of Tables

Acknowledgements

I would like to take this opportunity to express my gratitude to many persons and institutions for their guidance and support.

This memory has been done under the supervision of professor Enrique González Dávila, to whom I want to express my most sincere gratitude. His guidance helped me in all the time of research and writing of this thesis. Without his support and guidance, this research would never have been possible. He taught me not only valuable knowledge but also the wisdom of life, which gave me interest and confidence in accomplishing this thesis. I am truly thankful and fortunate to be able to work with him and I hope to continue being it.

I appreciate the support by the CajaCanarias Foundation Project "Desarrollo de un sistema de predicción de demanda turística en Canarias", which has made this PhD-research possible, as well as the Joint Research Unit on Public Statistics (ISTAC-ULL), which have helped me in developing the PhD-research. Who stood behind me, and believed in me and my project.

I extend my heartfelt gratitude to Domingo Lorenzo Díaz. Thank you for providing me the opportunity to work with you, it is a great experience. Thanks for your generous advice on a wide range of matters, from the most technical issues to general guidance on life, and your excellent feedback and patience with me.

I also extend my gratitude to Raquel Rivero, Alberto Yanes and Yenis González for guiding me in developing a vision for the topic of this thesis. They were always full of ideas and gave me countless hours of feedback and constructive suggestions. Thanks for believed in this research and helped me to explain and show it to others.

During these three years of research, I was really fortunate to be able to travel and present my research at international and national academic conferences. Thanks to that I had the best opportunities to meet many researchers from around the world and make friends. I again would like to acknowledge the CajaCanarias Foundation Project for funding my conference traveling expenses.

Last but not least, I cannot forget to thank my family and my friends for their support during these years. Those who perhaps did not dare to ask "what exactly are you doing" but always believed in me and my potential more than I never did it. Without them I would not have been able to keep such high spirits, reminding me every day that I can with everything and more. Thank you for being so supportive.

Preface

Over the last thirty years, the phenomenal growth of both the world-wide tourism sector and academic interest in tourism has created great curiosity in tourism demand modelling and forecasting from both business and research sectors.

Tourism is one of the main sectors which has a direct impact on the economic and financial development of a vast variety of countries around the world. World Tourism Organization (2019) reported that Spain is the top second destination by international tourist arrivals and in turn, Canary Islands is one of the top three destinations choose in Spain.

Time series modelling and forecasting have a positive and effective impact on the elaboration and planification of all engagements of the tourism industry besides defining the relationship with other factors contributing to tourism demand. There are a vast variety of methods to understand and analyse the relationship between tourism and its determining factors, besides in recent times it has been applied several quantitative time series models and forecasting techniques to forecasting tourist arrivals.

This thesis studies time series analysis including Autoregressive Integrated Moving Average Models and Structural Time Series Models and its application. The techniques that are developed in this thesis aimed at practical application to real problems in applied time series analysis.

This thesis arose from the need expressed by Instituto Canario de Estadística, which based on international recommendations has been developing in recent years a research strategy to offer information on tourism with the best possible quality at a regional level, to respond to administrations as well a companies and business associations that ask them to know the future tourists demand. To this end, a research project arose in which I was lucky to participate, whose goal was to forecast the number of tourist entries by destination island and visitor nationality in the following twelve months from the last available data.

The research in this Ph.D. thesis has been done at the Department of Mathematics, Statistic and Operational Research of the University of La Laguna, being supervised by professor Enrique González Dávila.

The plan of the manuscript is as follows. The thesis contains eight chapters. Chapter 1 provides an introduction of a little background as well as a review of the relevant literature and the research objective. Chapter 2 describes the basic concepts and classical time series approach. In Chapter 3, a brief review of the Box-Jenkins method is provided. Chapter 4 describes the structural time series models. Chapter 5 shows a study of the trend in structural time series models as well as the followed simulation process. Chapter 6 contains the application of the methodology exposed in the previous chapter to real data, describing the selection of forecasting models, the data used and also presents and discusses the results obtained. In chapter 7, an explanation of the software used and developed in this thesis is exposed. Finally, chapter 8 summarizes the main results, provides the limitations of this study and it also provides a brief presentation of future research lines that stem from this thesis. Appendix A contains opperators of Box-Jenkins methodology and Appendix B contains the R code developed and the reference manual of package STSM.

During the course of this research the results obtained have been showed in the following national and international conferences:

- *Spring Symposium on Challenges in Sustainable Tourism Development.* Gran Canaria (Spain), May 2017.
- *XXXI Congreso Internacional de Economía Aplicada "Economía Real y Finanzas".* Lisboa (Portugal), July 2017.
- *10th International Conference of the ERCIM WG on Computational and Methodological Statistics.* London (United Kingdom), December 2017.
- *Séptimo Punto de Encuentro de Jóvenes Investigadores en Matemáticas.* Tenerife (Spain), December 2017.
- *Third International Conference on Tourism & Leisure Studies.* Lanzarote (Spain), May 2018.

- *Octavo Punto de Encuentro de Jóvenes Investigadores en Matemáticas.* Tenerife (Spain), December 2018.
- *Encuentro de Usuarios de R.* Gran Canaria (Spain), March 2019.

Finally, the main results of this thesis have been pusblished in:

- *Stochastic and deterministic trend in state space models.* Communications in Statistics - Simulation and Computation. (Jorge-González et al., 2019a)
- *Univariate and multivariate forecasting of tourism demand using state-space models.* Tourism Economics. (Jorge-González et al., 2019b)

Chapter 1
Introduction

The tourism sector plays an important role as one of the main driving forces in economic development since it contributes to three high-priority goals of economics develop: employment, generation of income and foreign-exchange earnings. Tourism clearly provides a meaningful variety of positive and beneficial economic impacts to any country which receives a stable flow of tourist.

During the last decades, tourism has become one of the most dynamic activities in the current economies (Travel and Tourism, 2018). Studies of the tourism sector focused on knowing the tourist demand, as well as what are the factors that influence it, are currently of crucial importance. The companies that work in the tourism industry and public institutions related to this sector will be more efficient if they are able to improve in the planning and diversification of products and services related to tourism. The improvement of business decision-making processes in this sector involves adapting the methodology applied to the circumstances of each tourist destination and thus reducing the strong uncertainty inherent in these studies (Hassani et al., 2017). Additional, it is fundamental to provide several models that allow describing and forecasting the tourist demand in the short and long term, even combining both temporal horizons (Andrawis et al., 2011). If the tourism demand is correctly modelling then tourism policy makers and business practitioners could make use of the model results to evaluate the scale of impacts of unexpected events and base their decisions of employment and investment on the expected values of future tourism demand and change in its determinants.

Three main reasons justify the increase in the number of published studies on tourism modeling and forecast (Song, 2011): first is the significant increase in tourism that has occurred worldwide. Since the 1960s, international tourist arrivals have been increasing with a positive growth rate per year, which has generated increasing interest among academics and professionals who want to understand the determinants of this growth as well as its trends. Second, there is the rising importance of tourism prediction in the business planning process and the direct effect that these forecasts have on the growth strategies of tourism companies. Lastly, the forecast of demand at the destination level also helps local governments to formulate appropriate strategies and policies with the objective of generating sustainable tourism development, keep in mind that tourism is an important source of income and an important generator of employment for many countries.

Tourism modelling and forecasting contribute to answering three vital questions in policy analysis and business strategy of tourism, which decisions are often formed based to elicit or adjust to changes in tourism demand:

1. How many tourists are likely to arrive at a destination in a given time period?
2. Which origin country represents the main marketing opportunities for a destination?
3. Which factors and events are most influential in establishing future tourists arrive at a destination?

This thesis focuses on answer these questions for the specific case of the Canary Islands. After Catalonia and the Balearic Islands, the Canary Islands are the third preferred autonomous community among international tourists visiting Spain. These volcanic islands located off the coast of Morocco and Western Sahara has become almost a required visit for tourists from all over Europe, especially the United Kingdom and Germany. Its subtropical climate and cultural and landscape richness make the popularly called "lucky islands" a perfect tourist destination that each year generates around more than 40% of employment.

In order to do that, it should be noted that there is a vast variety of ways to measure tourism demand. The number of arrivals and its variations, which is often recorded by frontier counts, is the predominant measured of tourism demand, in per capita terms. This is because the tourist volumes have a direct impact on the supply capacity of tourism products or services. For example, the decisions on the number of airplanes, flights or new hotels rely mostly on an accurate forecast of tourist arrivals.

The dynamic of tourism implies continuous efforts in searching for new approaches and tools in order to obtain new knowledge and a better understanding of it. Econometric models and time series models are mainly used in forecasting tourism demand according to Song and Li (2008). For many years the Box-Jenkins methodology was used for much of the modelling of time series data, which provided an important step in the development of time series. Nevertheless, the Box-Jenkins models require the stationary of the time series, which implies differentiating the time series and not always one is possible to choose the right integration order. The assumption of stationarity implies rigidity. The continued increase in the availability of data in recent years and the possibility to construct larger frequency time series have promoted the use of statistical techniques to treat them more accurately. Actually, given these type of series, it is becoming harder to keep the assumption of a fixed pattern or behaviour through time and, in this context, structural time series models represent themselves as the suitable technique, due to they allow that each unobserved components posses a stochastic nature. Namely, the unobserved components describing the evolution of a time series have been usually treated in a deterministic way, but when these series are sufficiently large, it may be reasonable to consider that they envolve randomly over time. Turner and Witt (2001), Kim and Moosa (2001), Greenidge (2001) and Greenidge and Jackman (2010) have shown that the structural time series models are able to open up satisfactorily accurate forecasts. In this context, this thesis focuses on structural models for modelling time series data.

Classical tourism demand models are successful in forecasting and measuring the casual effects of exogenous variables on tourism demand for a single origin-destination pair, however, this approach tends to miss the indirect effects onto other countries. It is not unreasonable to hope that destinations become interdependent in the current era of globalization.

Impacts on a specific country can be transmitted and influence on others. Different origin countries located in the same general area and a similar economic, political and social structure are likely to be affected in the same way by economic, environmental or safety factors. The engagements between tourist flows, as well as their determinants, at different destinations countries establish the behaviours of tourists when they decide where to travel. Consequently, modelling tourism demand requires taking into account interdependencies across multiple countries.

1.1 Literature review

Many studies have been focused on modeling and forecast of tourist demand. Wu et al. (2017) show a review of the studies published from 2007-2015 on tourism and hotel demand modeling and forecasting. In addition to indicating that non-causal time series models, econometric approaches and artificial intelligence-based methods continue to dominate this field, a detailed review of the new explanatory variables that are being considered is also included, highlighting climate variables and tourist online behaviour variables. Jiao and Chen (2019) carry out a detailed review of the methods developed in tourism demand forecasting during the period of 2008-2017, placing special emphasis on the artificial intelligence methods and combination forecasts with different methods, in addition to the use explanatory variables based on big data. That article includes a summary table of model performance ranking in numerous articles depending on the criteria used. An extensive bibliographical review from the 60s approximately until 2000 can be found in Crouch (1994), Witt and Witt (1995), Lim (1999) and Li et al. (2005). Song and Li (2008), Song (2011) and Song et al. (2012) present a bibliographic review of the first decade of the 21^{st} century. Zhou-Grundy and Turner (2014) document the existing information about regional tourism demand forecasting. Saayman and Botha (2017) include a review of non-linear models for tourism demand forecasting and Hassani et al. (2017) show a review on the importance of accurate tourism demand forecasting and time horizons. Additionally, Kim et al. (2018) provide a review of the type of studies included in leading tourism and hospitality journals.

Harrison and Stevens (1971, 1976) pioneered the state-space approach to the structural time series model. Harvey (1989) has provided an extensive account of the approach. A structural time series model aims to represent the most outstanding characteristics of a time series by decomposing them into components that, although unobservable, have an interest in themselves and a direct interpretation. In other words, in a structural time series model, the time series observed are considered a sum of unobserved components, such as the trend, the cycle, a seasonal component, and an irregular component also known as the basic structural model or BSM. Greenidge (2001) successfully applied BSM to forecasting tourist arrivals to Barbados, showing that it offered valuable insights into the understanding of tourist behaviour. However, the BSM does not account for the external effects determinants on the variable of interest. To overcome this limitation, exogenous variables can be included in BSM to form a structural time series with explanatory variables, known as the casual structural model or CSM. However, no evidence has yet emerged to suggest that the inclusion of explanatory variables improves the forecasting accuracy of the BSM model (Turner and Witt, 2001; Kulendran and Witt, 2003). For that reason, the time-varying parameter (TVP) approach was introduced into the tourism context in the late 1990s, which takes into consideration the possibility of parameters changes over time and hence overcomes the structural instability problem caused by external shocks.

In particular, Wu et al. (2011) were pioneers in applying the TVP approach on the STS models, obtaining that the empirical results of the TVP-STS model outperforms on different models for short-term forecasts, among which were, the BSM, CSM, TVP, seasonal autoregressive integrated moving average (SARIMA) and autoregressive distributed lag (ADL) models, among others. This model also has the advantage that depending on the deterministic components that are set, it includes the BSM, CSM and the TVP model. Harvey (1989) and Durbin and Koopman (2012) extend the STS model from a univariate formulation to a multivariate setting. Chen et al. (2019) propose a method analogous to multivariate one to predict seasonal tourism demand using a multiseries STS method on nonseasonal series extracted from the same original seasonal series.

The characteristics of each tourist destination and the specifications of each region of origin, makes several researchers have studied the tourism in specifics destinations, focusing on the cases of USA, United Kingdom, France, Australia, Hong Kong, China, Spain, among others (Chan et al., 2005; Cho, 2003; Gil-Alana, 2005; Kim and Moosa, 2001; Kulendran and Shan, 2002; Gunter and Önder, 2015). With each other, the airport traffic prediction studies stand out (Tsui et al., 2014; Tsui et al., 2017; and López et al., 2017), mainly in destinations with limited access by roads and/or rail. Regarding to the temporality of the data, there are studies that use monthly data (Burger et al., 2001; Chu, 2004; Du Preez and Witt, 2003), quarterly data (Wong et al., 2007; Kulendran and Wong, 2005) and annual data (Song et al., 2003a; Song et al., 2003b).

Spain is the third country more visited of the world and the second in international tourism revenues (Poinelli, 2015). Due to the importance of tourism in this country, there have been numerous studies focused on this destination. González and Moral (1995, 1996) use the structural time series approach and include economic variables to model irregular trends in demand patterns, also using their inclusion in the cycle component. García-Ferrer and Queralt (1997) employ structural time series including delays in the economic variables introduced in the models as exogenous variables. Rossello-Nadal (2001) studies the British and German tourist behaviour at the Balearic Islands using the leading indicator approach obtained by macro-economic variables. Medeiros et al. (2008) also consider tourist arrivals for the Balearic Islands, using a neural network forecasting model that incorporates the time-varying conditional volatility. Garín-Muñoz (2011) analyses the British tourism demand behaviour employed panel data and dynamic models including economic variables and the demand in the previous period. Lopez et al. (2017) compare global forecast passenger air traffic results using SARIMAX models with the addition of individual models results by the airport. Landaluce-Calvo (2017) shows a multivariate exploratory analysis to describe the temporal structures of the evolution of tourist demand.

Within Spain, Canary Islands are the third Spanish region that receives the greatest number of tourists, being exceeded only by Catalonia and Balearic Islands (INE, 2016). In the bibliography found about tourism demand forecast at Canary Islands highlight three previous works: Garín-Muñoz (2006) applies models based on Generalized Method of Moments (GMM-DIFF) to determine the significance of the variables that may be affecting in international tourism demand; Gil-Alana et al. (2008) compare the use of different statistical seasonal models to analyse the number of tourists traveling to the Canary Islands; and Hoti et al. (2004) evaluate the tourism demand using volatility models.

1.2 Research Aims and Objectives

The purpose of this study is to

- Identify a time series models for short-term forecasting of tourist demand forecast at the Canary Islands.
- Search and study the influence of external variables that can affect and help to explain the behaviour of the tourist demand. Develop a tourism demand model using structural time series models.
- Evaluate the use of multivariate or univariate state space models in tourism demand forecasting by destination countries, to quantify the interdependencies across major countries.
- Analyse the influence of the slope and level depending of its treatment as stochastic or deterministic to described the trend in structural time series.
- Promote the use of structural models for tourism demand forecasting against the most commonly used models such as those included in the Box-Jenkins methodology.
- Provide free software packages that facilitate the use of state space models to non-expert users.

Although the main objective of this thesis was the analysis and forecast of tourism demand at the Canary Islands, due to the great interest shown by the tourism industry on Brexit (British exit, refers to the United Kingdom leaving the European Union) and its influence on Canary tourism which main tourist markets are made up of Germans and British, in the Chapter 6, some results obtained for tourism demand forecasts at the Canary Islands are shown, emphasizing in the tourism analysis according to the main markets of the tourist's origin to analyse how a change in the flow of ones of the destination countries could affect the others.

Chapter 2
Classical time series approach

2.1 Introduction

A time series is an ordened sequence of values of a variable at equally spaced time intervals, or also, the evolution of a phenomenon or variable over time. Said variable can be economic (daily sales prices of a specific product, evolution of interest rates, ...), physics (evolution of the flow of a river, evolution of the temperature of a region, ...) or social (annual population of a specific region, number of students in certain studies, number of births and deaths that occur in Spain each year, ...).

In general, the different observations are done by regular time range and constant duration. In this way, time series can be annual, semestral, monthly or quarterly, etc., according to the period of time in which the data was collected.

Time series are analysed in order to understand the underlying structure and function that produce the observations. Understanding the mechanisms of a time series allows a mathematical model to be developed that explains the data in such a way that prediction or control can occur.

The objective of time series analysis, for which data are available in regular periods of time, is the knowledge of its behaviour pattern to anticipate the future evolution, assumption that the conditions will be the same according to current and past.

If upon knowing the evolution of time series in the past could forecast its future behaviour without any type of error, we would be facing a deterministic phenomenon whose study would not have any special interest. But generally time series of interest have stochastic phenomena in a way that the study of its behaviour in the past only allows to approach the structure or probabilistic model for the forecasts. In other words, time series will be integrated by a regular part, systematic or deterministic, modeled with relative ease and another part with stochastic character, lacking regularity thus difficult to know its pattern of behaviour.

2.2 Components of a time series

Classical approach considers that a time series in general is supposed to be affected by four main components: *Trend*, *Cyclical*, *Seasonal* and *Irregular* components (Peña, 2010; Montgomery et al., 2015). It is difficult to define rigorously each one of these components, although from an intuitive point of view in some cases they are identifiable in the practice. A brief description of these four components is given following.

- ***Trend***: the trend picks up the variation of the time series over time. It is said that a time series presents a tendency when there is a significant variation in the average value of the time series as we add data to it. Thus, it can be said that trend is a long term smooth movement in a time series. In this way can appreciate if a time series increase, decrease or stagnate over a long period of time although to observe this it is necessary a broad time horizon.

  For example, series relating to population growth show upward trend, whereas downward trend can be observed in series relating to mortality rates, etc.

  Forecast of this component is usually in many cases the objective of time series analysis through the trend adjustment models, where it is assumed that time series lacks seasonal and cyclical variations.

- **Cyclical**: some time series show repetitive behaviour without a fixed period, maybe due to some other physics causes, the periodicity of the same is greater than one year, or by economics and social activities fluctuations in general. This effect is known as cyclical change or cyclical component.

  The duration of a cycle extends over longer period of time, usually two or more years. Most of the economic and financial time series show some kind of cyclical variation.

  It is usually harder to identify how much longer its period is, because usually the information collection time does not provide enough data, so sometimes can be confused with other components.

- **Seasonal**: seasonal variations in a time series are fluctuations within a year during the season, that are generally generated by periodic factors (quarterly, monthly, weekly, etc.) and that influence the behaviour of time series.

- **Irregular**: it is about the variations present in the data that are not explained by the seasonal and trend effects. Irregular variations in a time series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern. It is characterized, therefore, because it does not respond to a systematic or regular behaviour and hence its forecast would not be possible.

2.3 Objective

Time series analysis is useful when you want to extract information from a time series, to discover the characteristics of it, predict the changes of a time series or to improve control over the physical system studied. Though, the main objective of time series analysis is to be able to forecast the future values, on the basis of have a better knowledge of it in the time.

The objectives of time series analysis are as follows:

- **Description**. Summarize the information of a time series employing descriptive statistics and plots. The first step in the analysis is to plot the data and obtain simple descriptive measures of the main properties of the series, which will give us an idea about the main components of variation in it. In the same way, it could be observed if there are anomalous values, this is observations that do not seem to be consistent with the rest of the data. The treatment of these observations could be complex and it requires to be careful because could be valid observations or measurement errors.

- **Explanation**. Make a mathematical model from which the time series observed can be a sample element.

- **Prediction**. Given an observed time series, forecast the future values of the series.

- **Control**. When time series generated to measure the quality of a manufacturing process, then the aim objective of the analysis is to control the process.

2.4 Time series classification

A time series can be classified by different criteria. From a theoretical point of view, time series are classified in discrete or continuous according to the ability to observe of time. If, at least theoretically, it is possible observe the phenomenon that generate the time series at any time then it will be continuous. On the contrary, if the variable studied can only be observed in isolated moments of time then it will be discrete. Note that the discrete or continous nature of time series not be related with the discrete o continous nature of variable that generate it but with the observation capacity, discrete or continous of the time.

On the other hand, taking into account the number of variables observed at each instant of time. It can be collected observed values from a single source or simultaneously from two or more sources. Single-source observations generate **univariate** time series, and multi-source observantions form **multivariate** time series.

Furthermore, time series can be classified like:

- **Stationary**: a time series is stationary when contains no systematic change. This is reflected graphically in that the time series values oscillate around a constant average and the variability regarding that average stay constant over time. A time series is stationary when it is stable over time, without systematic increase or decrease of its values. For this type of time series, concepts such as mean and variance make sense. However, it is possible to apply the same methods to nonstationary time series if them are previously transformed into stationary.

- **Nonstationary**: a time series is nonstationary if its mean and/or variability change over time. Changes in the mean determine a increase or decrease trend long-term, so time series do not oscillate around a constant value.

Lastly, according to the probabilistic character it can be distinguish:

- **Deterministic models**: formal relationships that do not include random variables. Consequently, a model of this kind allows a relationship between values of the variable studied and fixed parameters throughout the time horizon considered: linear trend, fixed seasonal coefficients and constant frequency and amplitude cycles.

- **Stochastic models**: analytic relationships that connect values of the variable studied with cominations, linear or not, of parameters and variables among which at least one have probabilistic character, this is, it been generated by a random variable of a certain probability distribution although usually unknown. Therefore, the variable studied is a random variable subject to probability distribution laws with temporal reference. This statistics concept is known as stochastic process.

2.5 Model selection and estimates evaluation

Once a model has been implemented, it is necessary to check whether the assumptions underlying the model hold. Following, some model selection criteria and measures of accuracy are introduced.

2.5.1 Model selection criteria

There are many measures of the relative quality of a candidate model to be selected. In order to have a fair comparison between models with different numbers of parameters, the maximised likelihood functions may be used.

The Akaike information criterion, AIC, (Akaike, 1974) it is usually the most employed for this purpose, although it tends to overestimate the number of parameters in the model, therefore the corrected version is used, AICc (Burnham and Anderson, 2002). Added to that, another quality measure of the model is supplied employing the Bayesian information criterion, BIC (Schwarz, 1978). The formulations are given by (Harvey, 1989)

$$AIC = -2logL(\Psi) + 2n$$

$$AICc = -2logL(\Psi) + \frac{2nT}{T-(n+1)}$$

$$BIC = -2logL(\Psi) + nlogT$$

where $L(\Psi)$ represents the logarithm of the maximum likelihood function of the estimator Ψ of the model parameters, n is the number of parameters and T is the number of observations.

Whem comparing different models with this criterions, it has to be taken into consideration that smaller values denote better fitting models than larger ones. Also models with more parameters obtain a larger penalty. The lower the scores of AICc and BIC more convenient will be the model, both by adequacy as parsimony.

2.5.2 Measures of accuracy

Over the past few decades, diverse measures of accuracy to evaluate the performance of the estimation and prediction methods have been proposed, and which have been explicitly addressed by Armstrong and Collopy (1992) and further discussed by Fildes (1992) and Clements and Hendry (1993). In the 25 years' review of prevision carried out by Gooijer and Hyndman (2006) a table of the most measures employed can be found, between them, the *Mean Squared Error* (MSE), the *Root Mean Square Error* (RMSE), the *Mean Absolute Error* (MAE) or the *Mean Absolute Percentage Error* (MAPE). Due to the great diversity of existing tools, it is complicated to select the most suitable measure for a particular model. Depending on the importance given to the big errors, it should be selected one or another measure.

For a time series with T observations, let z_t denote the observation at time t and $\hat{z}_t$ denote the estimate of z_t ($\nu_t = z_t - \hat{z}_t$, the residual in periodo t), the most usually measures are

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\nu_t|$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\nu_t^2$$

$$RMSE = \sqrt{MSE}$$

However, these measures usually cause biased results when a time series presents extremely large outliers or just they are more vulnerable to outliers and to forecast it. For that reason, MAPE is one of the main prediction measures most used in practice, however, it presents some drawbacks. MAPE is defined by

$$MAPE(\%) = \frac{\sum_{t=1}^{T}|(z_t - \hat{z}_t)/z_t|}{T} \cdot 100$$

Being a percentage error, it can be excessively large when the time series to be estimated has values close or equal to zero. Besides, it has a bias that imposes a higher penalty on estimations above the real values. Nonetheless, it is one of the most used measures due to it is acceptable in terms of reliability and good in sensibility as was checked by Armstrong and Collopy (1992).

Additionally, to avoid potential problems of bias that MAPE can present, the Theil's U inequality coefficient could be used, defined as

$$U_t = \sqrt{\frac{\sum_{t=1}^{T-1}((z_{t+1}-\hat{z}_{t+1})/z_t)^2}{\sum_{t=1}^{T-1}((z_{t+1}-z_t)/z_t)^2}}$$

Theil's U can be interpreted as a quotient between the root mean square error (RMSE) of the proposed estimation method and the RMSE of the naïve model where estimate in a period is equal to the last observation of the variable. If the value is less than one, it means that the proposed model is more appropriate than the naïve model.

For both indexes, values close to zero would be indicative of models with more accurate estimates.

It should be stressed that the estimates evaluation can be carried out in two ways, the real or ex-ante and ex-post forecasts. Ex-ante forecast is a forecast based only on information available at the time of the forecast, while ex-post forecast uses the information beyond the period at which the forecast is carried out. Ex-post forecasts are most commonly in the context of modelling time series with exogenous variables (Song et al., 2011).

2.6 Outliers

During the observation period of a time series may occur external phenomena that affect the behaviour of it (a strike, a holiday period, a law change, etc.). For example, a terrorist attack on a tourist destination can have a drastic effect on the flow of tourists to that destination. These external phenomena are known as *interventions* and the technique that studies the effect or response of interventions are known as *intervention analysis*. Initially, in intervention analysis, it is assumed that the instant of time in which the intervention occurs is known. Nevertheless, a time series could be affected by a phenomenon whose existence is in principle unknown and be discovered in the intervention analysis of the residuals of an adjusted model. To cover these situations, the techniques of intervention analysis are generalized giving rise to what is known as *outliers analysis*.

An outlier is a value of a time series that has little chance of occurring given the usual evolution structure of the time series. The detection of outliers and their consideration in the adjustment of models is fundamental to make good forecasts.

The presence of outliers causes important problems in the univariate modeling of time series. Outlier detection influences modelling, inference, and data processing, because outliers can lead to biased parameter estimation and poor forecasts. Consequently, it becomes notorious that the outliers, or ignorance of them, will affect the predictions. This affectation will depend on different issues, such as the type of outliers not detected, the quantitative effect of the same, the period of occurrence (more or less close to the time horizon from which we began to calculate the predictions) and the process that generates the data.

2.6.1 Types of outliers

The most frequent atypical values are introduced below, four types of outliers in the context of atypical temporary or permanent changes of the time series. On the one hand, additive outliers and temporary changes are frequently related to the irregular component of the time series. On the other hand, level shifts and ramps are generally associated with the trend or cycle components.

Additive outlier

It will be said that an additive outlier (AO) has occurred on a time series at time t_0 if the value of the time series is generated at that time differently from the rest. This is, an AO represents a one-off peak or trough in the time series at a single point or observation, which can be defined as

$$AO_t = \begin{cases} 1 & \text{if } t = t_0 \\ 0 & \text{if } t \neq t_0 \end{cases}$$

Level shifts

The level shifts (LS) is an event that affects a time series in a given period, and whose effect is permanent. Therefore, it affects all observations from the moment it occurs. Then, an LS that permanently increase or decrease the time series by some constant factor prior to a specific observation can be expressed as

$$LS_t = \begin{cases} 0 & \text{if } t < t_0 \\ 1 & \text{if } t \geq t_0 \end{cases}$$

Temporary changes

A temporary change (TC) is an event that has an initial impact and whose effect decays exponentially by a damping factor (denoted by α). Therefore, it affects all observations from the moment it occurs, where the change in the level is not of permanent but temporary nature. A TC has the following representation

$$TC_t = \begin{cases} 0 & \text{if } t < t_0 \\ \alpha^{t-t_0} & \text{if } t \geq t_0 \end{cases}$$

Ramps

Ramps (RP) allows a linear transition between two time points, t_0 and t_1 the start and end points respectively. The effect of a ramp does not emerge immediately from one instant to the other but envolves linearly over time. It means no shock on the time series at t_0 and the influence increases continually until the effect achieves its final size at t_1. A RP has the following expression

$$RP_t = \begin{cases} 0 & \text{if } t < t_0 \\ (t-t_0)/(t_1-t_0) & \text{if } t_0 < t < t_1 \\ 1 & \text{if } t \geq t_1 \end{cases}$$

2.7 Explanatory Variables

Find the factors that impact on a specific time series is a crucial step in order to identify an appropriate model. These factors will depend and vary according to the time series nature. Taking into account the objective of this thesis, tourism demand forecast, a review of different factors that could affect the time series evolution is carried out bellow. Some of them may even be difficult to evaluate and therefore difficult to incorporate into the models.

Variation in tourism inflows is caused by many factors, economic and socio-psychological factors are considered in many studies. Over the past twenty years, several studies have been focused on this topic in particular, more concretely on economic factors such as relative prices, income, and exchange rates. However, other kinds of variables as the impact of wars, crisis or terrorism risk shouldn't be overlooked.

Some factors will be described below in more detail.

Income

In tourism demand models, income in the country of tourist's origin is generally included as a key explanatory variable, becuase usually it plays an essential role when it comes to travelling. Lim (1997) conducted a review and classification of the number of the explanatory variable used and conclude that income holds the title of the most frequently used variable.

In the case of attention focuses on business visits a broader income variable should be used. It is possible to apply income per capita, such as Gross Domestic Product (GDP). Song et al. (2010) chose the real GDP in their studies, focusing on identifying the key determinants of demand for Hong Kong tourism, due to the large proportion of people that visits Hong Kong for business trips, which represents the majority of its overall tourist arrivals.

Relative prices

When choosing possible determinants of tourism demand, price comes a close second to income. There are two elements of price to have in account: transportation costs and the cost of living for tourists at the destination (Martin and Witt, 1987).

The cost of international transportation not usually have included in the studies, due to the paucity of data and the difficulty to determine the actual price paid by tourists.

Regarding the cost of living at the destination, usually the Consumer Price Index (CPI) is a reasonable proxy for the cost of tourism in that country. Commonly, in order to reflect the differences of exchange rates between the destination countries and the currencies of origin, the following exchange rate adjusted CPI ratio formula, called Relative Consumer Price Index (RCPI), can be used

$$RCPI_{it} = \frac{CPI_{it}}{CPI_{jt}} ER_{it}$$

where $RCPI_{it}$ is relative price in destination i in period t; CPI_{it} is the CPI in destination country i in period t; CPI_{jt} is the CPI in origin country j in period t; and ER_{it} is an index of the price of origin country currency in terms of destination country i in period t.

Calendar effects

Calendar configurations can have a relevant impact on economic activity or tourism sector. The calendar effects component is that part of the seasonal component that represents calendar variations in a time series, such as moving holidays, changing month length, day-of-the-week (known as trading day) and other calendar effects (such as leap year).

Trading day variation occurs because the number of flights and arrivals varies with the day of the week. The number of tourists departures and arrivals to a destination in scheduled flights will depend on which days of the week occur four or five times in the month, for example.

A holiday day effect occurs because tourism behaviour patterns vary depending on whether a particular month contains a holiday or not, such as Easter week. For that reason, the displacements of festivities from one month to another according to the years is one of the most important sources to keep in mind within the calendar effects.

Interventions or Dummy variables

There is a lot of reasons why incorporate intervention variables into tourism demand models. Including them in the model allows capturing the impacts of various one-off events or the impacts of seasonality among others. Terrorist attacks, economic crises or wars are just some of many examples of one-off events, which can be included like a dummy variable. These one-off events usually can be reflected in the time series as outliers and should be checked and included in the model as was explained in 2.6.1.

2.8 Fundamental Concepts

In this section, some fundamental concepts that are necessary for a proper understanding of time series models are introduced.

2.8.1 Stochastic Processes

A *stochastic process* is a family of indexed random variables $Z(\omega, t)$, where t belongs to an index set and ω belongs to sample space. That is, a stochastic process is a set of time indexed random variables defined on a sample space.

From the probabilistic point of view, a stochastic process is perfectly characterized when its finite-dimensional distributions are known, that is, the joint distributions of any finite subset of process variables. Another way, generally more incomplete, to describe a stochastic process is to provide the moments of it, particularly, the first and second-order, what we will call functions of mean, of variance and autocovariance. (Brockwell and Davis, 2002; Peña, 2010)

Marginal distributions

The *mean function* of the process refers to a function of time representing the expected value of the marginal distributions z_t for each instant

$$\mu_t = E[z_t]$$

The *variance function* of the process give the variance at each point in time

$$\sigma_t^2 = Var[z_t]$$

The *autocovariance function* of the process refers to the covariance between two variables of the process at any two given times

$$\gamma_{t_1,t_2} = Cov(z_{t_1}, z_{t_2}) = E[(z_{t_1} - \mu_{t_1})(z_{t_2} - \mu_{t_2})]$$

In particular,

$$\gamma_{t,t} = Var[z_t] = \sigma_t^2$$

The *autocorrelation function* is the function of two plots that describe the autoccorelation coefficients for any two values of the variables

$$\rho_{(t_1,t_2)} = \frac{\gamma_{t_1,t_2}}{\sqrt{\sigma_{t_1}^2 \sigma_{t_2}^2}}$$

2.8.2 Stationary and nonstationary processes

One of the most important properties of any time series process is whether it is or not stationary. Intuitively, if a time series has the same probability distribution at every point in time, then that time series follows a stationary process. The idea of stationary is related to the stability of the time series. A stationary process is described as a sequence of data or values that do not present systematic change in the mean or the variance, it is said then that the series is stable.

One example of nonstationary time series are variables that have a trend component. Time series that present a trend component will be nonstationary, due to the mean of its distribution is not the same at all periods. However, many of the properties that the stationary time series meet could be applied to nonstationary time series once some components will be extracted to the original time series.

A time series process $\{z_t\}$ is *strictly stationary* if the joint probability distribution of any pair of observations from the process (z_t, z_{t-s}) depends only on the distance in time between the observations (s), and not on the time (t).

However, the concept of stationary in the strict sense implies compliance with several conditions that are excessive in practice. Therefore, in most of the cases *weak stationarity* or *covariance stationarity* is used.

Formally, a time series z_t is weak stationarity if

$$\begin{aligned} &E[z_t] = \mu, && \forall t \\ &Var[z_t] = \sigma_z^2, && \forall t \\ &Cov(z_t, z_{t-s}) = \sigma_s, && \forall t, s \end{aligned}$$

2.8.3 White noise

The simplest kind of time series process corresponds to the known as *white noise*. A time series a_t is called a white noise if $\{a_t\}$ is a sequence of independent and identically distributed random variables with finite mean and variance, in other words, each element has an identical, independent, mean-zero distribution.

Officially, it can be said that a_t is a white-noise process if

$$\begin{aligned} &E[a_t] = 0, && \forall t \\ &Var[a_t] = \sigma^2, && \forall t \\ &Cov(a_t, a_s) = 0, && \forall t \neq s \end{aligned}$$

In particular, if a_t is normally distributed with mean zero and variance σ^2, the series is known as *Gaussian white noise*.

2.8.4 Random Walk

One of the most important examples of nonstationary processes are *Random Walk* with or without a drift (a slow steady change).

Pure Random Walk

A time series said to follow a random walk if the first differences are random. Then, a random walk predicts that the value at time t will be equal to the last period value plus a stochastic component (a white noise).

$$z_t = z_{t-1} + \varepsilon_t$$

Random Walk with Drift

If the random walk model predicts that the value at time t will be equal the last period's value plus a constant or drift, and a white noise term, then it can be said that the process is a random walk with a drift.

$$z_t = \alpha + z_{t-1} + \varepsilon_t$$

Chapter 3
Box-Jenkins Methodology

3.1 Introduction

In the previous chapter, an introduction about the fundamentals of time series has been shown. The selection of an appropriate model is a key aspect to study and analyse the underlying structure of the time series, in addition, to be able to be used for prediction when that is its objective.

Altogether, time series models can have numerous forms and represent different stochastic processes. There are two extensively used linear time series models in the literature, *Autoregressive* (AR) and *Moving Average* (MA) models. Combining both it is possible to generalise a useful class of time series models, the *Autoregressive Moving Average* (ARMA) and the *Autoregressive Integrated Moving Average* (ARIMA) or the *Seasonal Autoregressive Integrated Moving Average* (SARIMA) for seasonal time series. ARIMA model and its different variations are based on the well-known Box-Jenkins Methodology, and so these are also extensively known as the Box-Jenkins models (Box and Jenkins, 1970).

In this chapter, these models are briefly introduced. Although they are not the focus of interest in this thesis, it is necessary to know about them and how they are applied since nowadays it is still a widely used methodology. These models have been broadly discussed in extensive literature (Box and Jenkins, 1970; Peña, 2010; Brockwell and Davis, 2002; Montgomery et al., 2015) and they will serve as a reference point in the methodologies described in the following chapters.

3.2 Univariate Linear time series model

The class of linear processes constitutes an important framework for the study of stationary stochastic processes. In fact, every stationary process is a linear process, or it can be transformed into a linear process, subtracting its deterministic component.

The formal representation of the random processes that generate real time series through the so-called linear models of time series is a classic result of theoretical statistics. However, its practical use has developed especially from the methodological proposal of Box and Jenkins.

The Box-Jenkins approach is one of the most widely used methodologies for stochastic time series modelling. It is popular because can handle many of the time series treated in practice, stationary and in some case nonstationary, and for having been implemented in numerous computer programs.

3.2.1 Autoregressive Models

An *Autoregressive Model* (AR) forecast future observations based on past observations, this is, a value from a time series is regressed on previous values from that same time series. An AR(p) or autoregressive model of order p is a discrete time linear equations with noise and can be formulated as

$$z_t = \phi_1 z_{t-1} + ... + \phi_p z_p + a_t$$

where p is the order, $\phi_1, ..., \phi_p$ are the unknown coefficients and a_t is a white noise term. The fact that this model forecast the time series using a linear combination of past values of the time series with decreasing coefficients causes them to be identified as long memory processes.

It may be desirable to introduce models with non-zero average, in order to model general situations,

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + ... + \phi_p \tilde{z}_p + a_t$$

where $\tilde{z}_t = z_t - \mu$; being μ the mean of the process.

A model AR(p) can be written in a more concise form using *time lag operator* as

$$\phi_p(B)\tilde{z}_t = a_t$$

being B the lag operator defined by $Bz_t = z_{t-1}$ for all $t \in \mathbb{Z}$ (see Appendix A.2), and

$$\phi_p(B) = 1 - \phi_1 B - ... - \phi_p B^p$$

which is so-called *AR-polynomial* in B of order p.

3.2.2 Moving Average Models

Due to an AR model has a relatively "long" memory, this model cannot represent short memory series, where the current observation of the time series is only correlated with a small number of previous observations. Unlike the AR models, the *Moving Average Model* (MA) have the property of very short memory. The MA model uses previous errors as predictors for future outcomes. The MA(q) or moving average model of order q is an explicit formula for $\tilde{z}_t$ in terms of noise as

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q}$$

Introducing the lag operator notation,

$$\tilde{z}_t = \theta_q(B)a_t$$

with $\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)$.

Thanks to a MA(q) is a finite linear combinations of a white noise sequence for which the first two moments are time invariant, then a MA model is always stationary.

3.2.3 Autoregressive Moving Average Models

AR(p) and MA(q) models can be effectualy blended together to form a general and useful class of time series models. This form combined is known as *Autoregressive Moving Average Models* (ARMA). An ARMA(p,q) expresses the value at time t as the sum of q terms that represents the average variation of random variation over q previous periods, the MA component, plus the sum of p AR terms that compute the present value of $\tilde{z}$ as the weighted sum of the p most recent values.

An ARMA(p,q) model is represented as

$$(1 - \phi_1 B - ... - \phi_p B^p)\tilde{z}_t = (1 - \theta_1 B - ... - \theta_q B^q)a_t$$

or using the lag operator notation as

$$\phi_p(B)\tilde{z}_t = \theta_q(B)a_t$$

This model is common in practice since the sum of both AR and MA processes results in ARMA processes (Peña, 2010).

3.2.4 Autoregressive Integrated Moving Average Models

An *Autoregressive Integrated Moving Average Models* (ARIMA) is the combination of AR(p) and MA(q) models as the ARMA, with the particularity of including a restoration process (called *integration*) of original instability present in a time series.

The general form of an ARIMA(p,d,q) is

$$(1 - \phi_1 B - ... - \phi_p B^p)(1 - B)^d z_t = c + (1 - \theta_1 B - ... - \theta_q B^q)a_t$$

where p is the order of the autoregressive part, d is the number of unit roots (order of the process integration) and q is the order of the moving average part. Using the *difference operator*, $\nabla = 1 - B$, the previous process is usually formulated as

$$\phi_p(B)\nabla^d z_t = c + \theta_q(B)a_t$$

3.2.5 Seasonal Autoregressive Integrated Moving Average Models

Sometimes the expected value of a time series does not remain constant but it varies predictably according to a cyclic pattern. In these situations can be said that the time series is seasonal.

The seasonal period s defines the number of observations that make up the seasonal cycle. The simplest way to introduce seasonality into a model is to incorporate it as a constant effect that adds to the values of the time series in the corresponding periods. Another possibility is to incorporate it directly into the model just as seasonal ARIMA models do.

When there is seasonal dependence, the ARMA model for stationary series can be generalized, incorporating in addition to the regular dependence, the one associated with the measuring intervals of the series, the seasonal dependence, associated with observations separated by s periods. The incorporation into the model of seasonal and regular dependence in a multiplicative way will allow the modeling of both separately. The multiplicative *Seasonal Autoregressive Integrated Moving Average Model* (SARIMA) has the form

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d z_t = \theta_q(B)\Theta_Q(B^s)a_t$$

where $\Phi_P(B^s) = (1 - \Phi_1 B^s - ... - \Phi_P B^{sP})$ is the seasonal AR operator of order P, $\phi_p(B) = (1 - \phi_1 B - ... - \phi_p B^p)$ is the regular AR operator of order p, $\nabla_s^D = (1 - B^s)^D$ represents the seasonal differences, and $\nabla^d = (1 - B)^d$ the regular differences, $\Theta_Q(B^s) = (1 - \Theta_1 B^s - ... - \Theta_Q B^{sQ})$ is the seasonal MA operator of order Q, $\theta_q(B) = (1 - \theta_1 B - ... - \theta_q B^q)$ is the regular MA operator of order q, and a_t a white noise process. These kinds of models were introduced by Box and Jenkins and it represents well numerous seasonal time series that can be found in practice. They can be identified as $ARIMA(p, d, q) \times (P, D, Q)_s$.

These models can be extend incorporating information provided by leading indicators or exogenous variables, resulting in the *Autoregressive Integrated Moving Average Models with exogenous variables* (SARIMAX).

3.2.6 Airline Passenger Model

The most used seasonal SARIMA model is the so-called *Airline Passenger Model*, which adapts well to many time series that can be found in practice and in particular, in studied of tourism demand. It is formulated as

$$\nabla\nabla_s z_t = (1-\theta B)(1-\Theta B^{12})a_t$$

that is representing as $ARIMA(0,1,1)\times(0,1,1)_{12}$ and whose equation to calculate the forecasts is, for $k>0$

$$\hat{z}_t(k) = \hat{z}_t(k-1)+\hat{z}_t(k-12)-\hat{z}_t(k-13)-\theta\hat{a}_t(k-1)-\Theta\hat{a}_t(k-12)+\theta\Theta\hat{a}_t(k-13)$$

This equation indicates that to forecast what happens in the period *t*, what happened in the period *t-1*, what happened in the previous year *t-12* and what happened in the month before the period *t-12* are used. For example, if you want to see what happens in the month of December 2016 then the information of November 2016, December 2015 and November 2015 will be used.

3.3 Box-Jenkins Methodology

Different methodologies help to choose the model that best represents the time series from among the various existing models. The most used and widespread methodology is that proposed by professors G.E.P Box and G.M Jenkins in the 1970s, in which they made great progress in identifying, adjusting and verifying the appropriate ARIMA models. It is commonly known as *Box-Jenkins Methodology*. Only the use of the ARIMA models is specified in this methodology since the AR, MA and ARMA models are considered special cases of the ARIMA models.

This methodology is based on trying to determine which is the probabilistic model that governs the behaviour of the process over time. It should be noted at this time the difference between process and model. It can be said that a *process* is the real thing, that is, the phenomenon itself, of which its generating mechanism is unknown. On the other hand, a model is only the imitation or representation of the process, which cause could be more than one model to represent a specific process. For the selection of the best model this methodology proposes five main stages: stationary, identification, estimation, evaluation and forecast (Figure 3.1).

The first step in Box-Jenkins methodology, **stationary**, is to determine if the time series is stationary. In case the time series is not stable over time, it is necessary to apply a transformation to the series to induce it. This step is essential to be able to apply the other stages of this methodology. The main two ways in practice to know if a series is stationary are through the plot of the time series or the exploration of the *autocorrelation function* (ACF).

In the next step, **model identification**, the identification of ARIMA models more probable to govern the time series process is carried out through comparing the autocorrelation function (ACF) and the *partial autocorrelation function* (PACF).

In the third step, **estimation**, the coefficients of the tentatively chosen model in the previous step are estimated and should be check that covers different criteria such as parsimony, stationary and invertible, statistically significance and appropriate adjustment.

Once the coefficients of the proposed model have been estimated, the next stage in the Box-Jenkins is its **evaluation** or verification. In this step, the efficiency of the model is checked and it is decided if it is statistically adequate.

Finally, the last phase of this approach is to **forecast** future values of the time series, usually supplying confidence interval to known if the forecasts are satisfactory or not.

After a suitable model has been found, forecasts for a period, or several, can be carried out in the future.

If the pattern of the series seems to change over time, the new data could be used to re-estimate the model parameters or, if necessary, develop a completely new model.

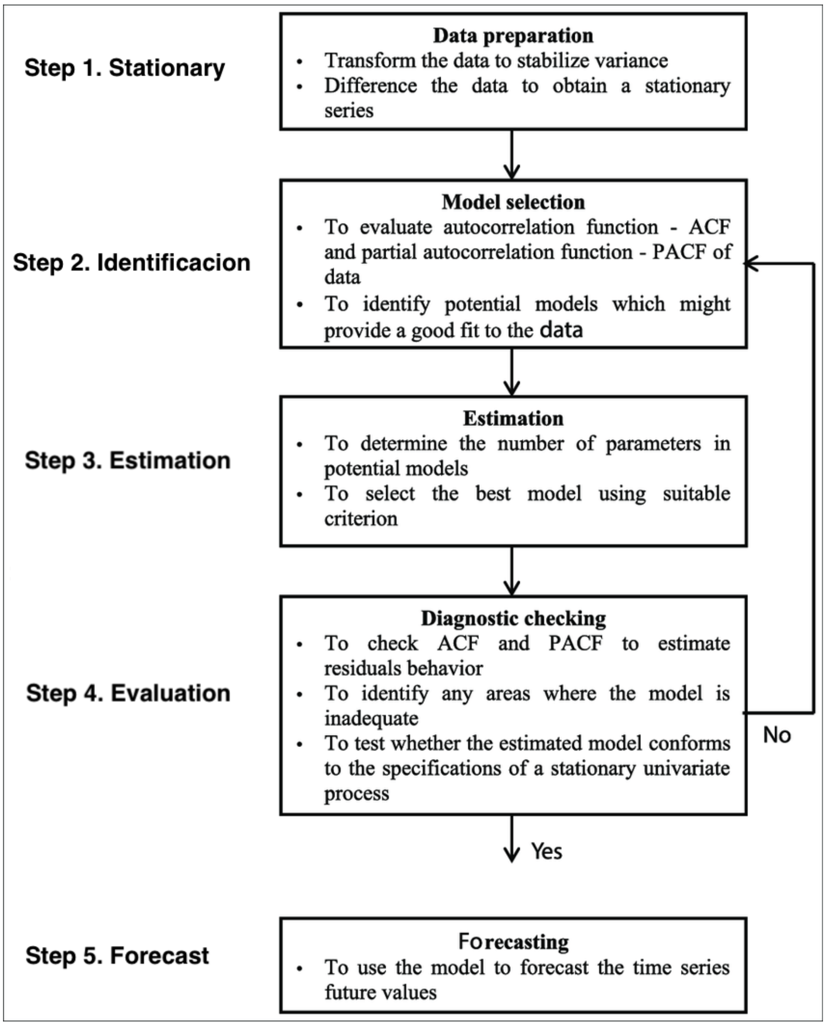

Figure 3.1: *Steps of the Box-Jenkins Methodology (Box and Jenkins, 1970).*

3.4 Multivariate time series model

Unlike the Univariate Model, a *Multivariate time series model* (MARIMA) has more than one time-dependent variable, where each variable depends not only on its past values but also has some dependency on other variables. The MARIMA model use information from other related time series to attempt to obtain better forecasts.

If considerer a k-variable random vector z_t of observations and, correspondingly, a k-variable random vector, a_t, of unknown innovations with zero mean

$$z_t = \begin{pmatrix} z_{1,t} \\ z_{2,t} \\ \cdots \\ z_{k,t} \end{pmatrix}, \quad and \quad a_t = \begin{pmatrix} a_{1,t} \\ a_{2,t} \\ \cdots \\ a_{k,t} \end{pmatrix}$$

the random vector z_t can be generated through the multivariate ARMA(p,q) model as

$$z_t + \phi_1 z_{t-1} + ... + \phi_p z_{t-p} = a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q}$$

where the coefficient matrices $\phi_1, ..., \phi_p$, and $\theta_1, ..., \theta_q$ are all of dimensions $k \times k$.

Similarly to the Univariate Models, it can be defined as the *k-variate backward shift operator* B. The result of multiply B on a time indexed k-variate random variable should be interpreted as the variable lagged a single time step. In general, lagging r time steps are accomplished using

$$B^r z_t = z_{t-r}$$

Then, if the operator B is introduced into ARMA(p,q) representation, the model can be written as

$$(I + \phi_1 B + ... + \phi_p B^p) z_t = (I + \theta_1 B + ... + \theta_q B^q) a_t$$

where I is the $k \times k$ unity matrix. Defining the *matrix polynomials* $\phi(B) = I + \phi_1 B + ... + \phi_p B^p$ and $\theta(B) = I + \theta_1 B + ... + \theta_q B^q$, the general multivariate ARMA(p,q) model can be formulated in operator form as

$$\phi(B) z_t = \theta(B) a_t$$

This model, as well as the Univariate Model, can be extended to include external regression variables or incorporate a seasonal cycle or include an integration.

It should be noted that the Multivariate Models have more parameters than Univariate Models since every additional parameter is an unknown value and has to be estimated (Brockwell and Davis, 1991). Multivariate Models usually have a simpler structure than Univariate Models to overcome the additional complexity of being multivariate, whereby Univariate Models generally are preferred in practice.

Chapter 4
Structural Time Series

4.1 Introduction

Structural time series models (Harvey, 1989) are models that are formulated directly in terms of components such as trend, seasonality and cycle. These models, therefore, offer clear interpretations through the decomposition into components, which ability is a major attraction for time series forecasting. Considering a time series that can be decomposed as the sum of a trend, seasonal and irregular component, then if these components are not stable it may be reasonable to consider that they envolve randomly over time and it would be necessary for them to change over time. This flexibility is possible with structural models and it is the starting assumption of it. These models are included in the so-called state space models (Harvey and Shepard, 1993). The components of the series are generally unobserved, and although there are different techniques employed for its modelling, the representation in state space form is increasingly often. In addition to all these unobserved components, the inclusion of exogenous variables that could be explaining the relationship between the series and other phenomena that have occurred, even if a sporadic character, is frequent. Autoregressive terms of the own series are also often included. Other components not observed for a periodic time series analysis also can be introduced (Koopman and Ooms, 2006).

The state space models applied to structural time series are explained with rigor in Harvey (1989), as well as in many later references (Harvey et al., 2004; Durbin and Koopman, 2012). Kitagawa (1998) proposes calculation methods of filtering and smoothing functions when we have non-Gaussian state space models. Among others, Harvey and Koopman

(1996) extended this methodology to multivariate time series. There are also numerous references that adapt the classic Box-Jenkins methodology to study time series using autoregressive integrated moving average (ARIMA) models to their equivalents in state space (Casals et al., 2016).

The structural time series models introduced before are special cases of state space models (SSM). State space models have their origin in the field of control engineering, beginning with the pioneering report of Kalman (1960). Famous applications were initially and still are, developed to accurately track the position and velocity of moving objects such as airplanes, rockets or missiles, in astronautics. Later, they were adapted to treating econometrics time series data in Harvey (1989), West and Harrison (1997) and Hannan and Deistler (1988).

Altogether, state space methodology provides a successful approach for areas that can generate longitudinal data. Nevertheless, applications of state space models in the social sciences, except econometrics, are still uncommon.

State Space Model formulation is based on the *measurement equation* and the *transition or state equation.* The measurement equation relates the observed variables to the elements of the state vector, and the transition equations describe the evolution of the unobserved state vector, like the trend or seasonal component.

The general state space model can be written in a variety of ways, but ones of the most employed is the form

$$z_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \tag{4.1}$$

$$\alpha_{t+1} = T_t\alpha_t + \eta_t, \quad \eta_t \sim N(0, Q_t), t = 1, ..., T \tag{4.2}$$

where z_t is the dependent variable and α_t is an unobserved vector called *state vector*. The matrices Z_t, T_t, H_t and Q_t are initially assumed to be known, and ε_t and η_t are the disturbances, which are serially independent and independent of each other at all time points.

The rationale behind this model is that the development of the concerning system over time is determined by α_t according to equation 4.2, known as the *state equation*. Meanwhile, α_t is unobservable, the analysis must be based on the observations z_t. Then, equation 4.1 is called the *observation equation*.

The main reason for setting a structural time series model in a state-space form is susceptible to the application of the Kalman filter, which is the main algorithm to estimate dynamic systems in state-space form. This filter can be used both for obtaining the unobserved components and for estimating the parameters of the model. The Kalman filter is a recursive procedure that given all the information available at time t to calculate the optimal estimator of the state vector. Full illustration of the Kalman filter technique is available from Harvey (1989). This allows that different methods applied in time series analysis, such as VARMA models in general, time-varying parameters, etc., being susceptibles to be written in state space form, can be analysed using this approach.

Despite state space model covers a wide field, the objective in the present manuscript will be the application of structural time series models both in univariate and multivariate versions. Following, the basic formulation of structural time series models will be introduced.

4.2 Univariate time series models

A structural model for the time series $\{z_t\}_{t=1,\dots,n}$ can be defined as

$$z_t = \mu_t + \psi_t + \gamma_t + \varepsilon_t,$$

where, μ_t is a slowly varying component called the *trend*, ψ_t is a cyclical component called the *cycle*, γ_t is a periodic component of fixed period called the *seasonal* and ε_t is an irregular component called the *disturbance*.

As well as in the classical time series approach (Section 2.2), in structural time series models, the dependent variable is decomposed into a trend, seasonal, cycle and irregular component. A trend is the long-term growth of the model, and it usually be divided into two factors: level, it is an actual value of the trend; and slope, it is a tendency to grow of the trend. Due to some time series move around their level randomly without any tendency to grow, a slope is not always necessary, it is only used when the time series have continuous growth. A seasonal component reflects regular variations that recur every year to the same extent.

Structural models stand out for their flexibility to catch the behaviour changes of the series due to the consideration of the different components as stochastic processes directed by random disturbances. A structural model is not always defined by all unobserved components, factors as the periodicity or the sample length will influence in the components to consider when formulate the model. Depending of which are the main characteristics of the series that it wants to catch in the model, these unobserved components are combined in diverse ways giving rise to the different structural models (Koopman and Durbin, 2012). The following will describe some of the possible different formulations for each component or combinations among them.

The simplest case resulting of considerer a time series whose observations oscillate around of a medium level that can remain constant or vary over time. One of the simplest structural models is obtained by adding noise to a random walk, known as the *Random Walk plus Noise Model.* It is suggested by the nonseasonal classical decomposition model

$$z_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$$

where $\mu_t = \mu_0$, the deterministic level for all $t = 1, ..., n$.

If the level component is allowed to vary in time, the state space model as called the *Local Level Model*, in which the level at each moment in time is the sum of the previous period value and a random element, and it can be formulated as

$$\begin{aligned} z_t &= \mu_t + \varepsilon_t, \quad && \varepsilon_t \sim NID(0, \sigma^2_\varepsilon) \\ \mu_{t+1} &= \mu_t + \xi_t, && \xi_t \sim NID(0, \sigma^2_\xi) \end{aligned}$$

being σ^2_ε and σ^2_ξ the variances of the irregular component of the series and the level, respectively. Representing μ_t a random walk noise. If the variance of the level is zero, the deterministic level model as scpecified before is obtained. The first equation is called the *observation* or *measurement* equation, while the second equation is called the *state equation.*

If we add a slope β_t to the level component as described above, the local level model is extended to the *Local Linear Trend Model*, as follows

$$\begin{array}{ll} z_t = \mu_t + \varepsilon_t, & \varepsilon_t \sim NID(0, \sigma^2_\varepsilon) \\ \mu_{t+1} = \mu_t + \beta_t + \xi_t, & \xi_t \sim NID(0, \sigma^2_\xi) \\ \beta_{t+1} = \beta_t + \zeta_t, & \zeta_t \sim NID(0, \sigma^2_\zeta) \end{array}$$

The local linear trend model contains two state equations: one for modelling the level, and one for modelling the slope. The disturbance term ζ_t gives to the slope a random character. If σ^2_ζ is zero, such that the slope is constant, then the model takes a stochastic level together with a fixed slope (known as local level with drift), so that

$$\mu_{t+1} = \mu_t + \beta_0 + \xi_t$$

In many time series it is important to distinguish between a long run trend and the cyclical or short-run movements. A cycle describes behaviour patterns that are repeated over the years, which can be seen as a long-term seasonal component with a perfectly periodic behaviour and with a given frequency. The cycle component can be expressed as a function of sines and cosines of the following type

$$\psi_t = \varpi_1 cos(\lambda t) + \varpi_2 sin(\lambda t)$$

where λ represents the angle frequency of the cycle, in radians; therefore, the ratio $\lambda/2\pi$ expresses the number of times the cycle repeats per unit of time; the amplitude of the cycle is $(\varpi_1^2 + \varpi_2^2)$; and $tan^{-1}(\varpi_2/\varpi_1)$ its phase. Just as with the unobserved components described above, the cycle components admits as well a deterministic or stochastic specification. If we assume that the parameters ϖ_1 and ϖ_2 envolve randomly through time, then the cycle has a stochastic nature and in this case it is convenient to express the cycle component in a recursive way such as

$$\begin{cases} \psi_t = \rho cos\lambda \psi_{t-1} + \rho sin\lambda \psi^*_{t-1} + \kappa_t \\ \psi^*_t = -\rho sin\lambda \psi_{t-1} + \rho cos\lambda \psi^*_{t-1} + \kappa^*_t \end{cases}$$

where ψ_{t-1} is the value of the cycle at $t-1$; ψ^*_{t-1} appears by construction; ρ is a damping factor between 0 and 1; and the error κ_t and κ^*_t are white noise Gaussian errors, that are mutually uncorrelated and possess the same variance σ^2_κ.

The seasonal component represents the patterns that are repeated every certain period of time less than one year. Assuming that seasonal fluctuations are stable, namely they manifest regularly every certain period of time, then the seasonal component at time t with a s fixed number of seasons, it expressed as

$$\gamma_t = \sum_{j=1}^{s-1} \lambda_{j,t}\gamma_j$$

where γ_t represents a fictitious variable that is only valid (non-zero) in a certain season j; and $\lambda_{j,t}$ is defined as

$$\lambda_{j,t} = \begin{cases} 1 & \text{if t=j} \\ -1 & \text{if j=s} \\ 0 & \text{other case} \end{cases}$$

such that,

$$\sum_{j=1}^{s} \gamma_j = 0$$

Seasonality expressed in this way guarantees that the sum of seasonal effects over the period s is null, which it involves that fluctuations are deterministics.

In some cases is more convenient to express the seasonal in a trigonometric formulation as the quasi-random walk model

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{j,t}$$

with $[s/2]$ the integer part of $s/2$ and $\gamma_{j,t}$ is generated by

$$\begin{pmatrix} \gamma_{j,t} \\ \gamma^*_{j,t} \end{pmatrix} = \begin{pmatrix} cos\lambda_j & sin\lambda_j \\ -sin\lambda_j & cos\lambda_j \end{pmatrix} \begin{pmatrix} \gamma_{j,t-1} \\ \gamma^*_{j,t-1} \end{pmatrix} + \begin{pmatrix} \omega_{j,t} \\ \omega^*_{j,t} \end{pmatrix}$$

for $j = 1, ..., [s/2]$ and $t = 1, ..., T$. Where $\lambda_j = 2\pi j/s$ is the frequency, in radians, and the $\omega_{j,t}$ and $\omega^*_{j,t}$ terms are independent $N(0, \sigma^2_\omega)$ variables.

Combining the trend to be formulated as a local linear trend model and the seasonal component as ones of the ways explained before, it results in the model known as *Basic Structural Time Series Model* (BSM), which plays a prominent part in the approach of Harvey (1989) to structural time series analysis and can be express as

$$
\begin{array}{ll}
z_t = \mu_t + \gamma_t + \varepsilon_t, & \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
\mu_{t+1} = \mu_t + \beta_t + \xi_t, & \xi_t \sim NID(0, \sigma_\xi^2) \\
\beta_{t+1} = \beta_t + \varsigma_t, & \varsigma_t \sim NID(0, \sigma_\varsigma^2) \\
\gamma_t = \sum_{j=1}^{s-1} \lambda_{j,t}\gamma_t + \omega_t, & \omega_t \sim NID(0, \sigma_\omega^2)
\end{array}
$$

where $\lambda_{j,t}$, $j = 1, ..., s-1$, is equal to 1 if t in the station j, -1 if t in station s and 0 in other case; β_t the stochastic slope of the linear trend; and σ_ε^2, σ_ξ^2, σ_ς^2 and σ_ω^2 the variances of the irregular component, the level, the slope and the seasonal of the series, respectively.

The BSM, as well as, univariate time series models in general, does not account for the effects of external determinants on the variable of interest. To resolve this limitation, it is possible to include casual variables in the BSM model, generating the model known as the *Casual Structural Model* (CSM), which is express as follow,

$$
\begin{array}{ll}
z_t = \mu_t + \gamma_t + \sum_{i=1}^{k} \delta_i x_{i,t} + \varepsilon_t, & \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
\mu_{t+1} = \mu_t + \beta_t + \xi_t, & \xi_t \sim NID(0, \sigma_\xi^2) \\
\beta_{t+1} = \beta_t + \varsigma_t, & \varsigma_t \sim NID(0, \sigma_\varsigma^2) \\
\gamma_t = \sum_{j=1}^{s-1} \lambda_{j,t}\gamma_t + \omega_t, & \omega_t \sim NID(0, \sigma_\omega^2)
\end{array}
$$

with $x_{i,t}$ the exogenous variables or interventions and δ_i the unknown parameteres to estimate.

However, in some times, the CSM can not generate more accurate forecasts than the BSM. One possible reason for that maybe is related to its treatment of the explanatory variables. In the CSM the parameters of the explanatory variables are treated as being constant over time, unlike the seasonality or trend, which could be treated as stochastic (González and Moral, 1995; Kulendran and Witt, 2001; Turner and Witt, 2001).

To overcome the limitations of the traditional fixed-parameter estimation, the Time-Varying Parameter (TVP) modelling approach was introduced (Song et al., 2003b; Song and Wong, 2003). This approach allows for stochastic parameters, relaxing the restriction on the constancy of the parameters of the exogenous variables. Therefore, the coefficients of the exogenous variables will be make time-varying by incorporating them directly into the state vector, such a way as the variance of the corresponding state errors has to be nonzero.

4.3 Multivariate time series models

In the previous section, all state space models dicussed are concerned with the analysis of only one time series. However, the methodology of structural time series model lends itself easily to generalisation to multivariate time series (Harvey, 1989; Commandeur and Koopman, 2007), which can also be represented in the state space form by

$$z_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, H_t) \tag{4.3}$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim NID(0, Q_t) \tag{4.4}$$

for $t = 1, ..., n$, where

z_t contains the values of the p observed time series at time point t.

ε_t is a $p \times 1$ irregular vector contains the p observation disturbances.

H_t is a $p \times p$ matrix with an unknown variance-covariance structure of the p observation disturbances.

α_t is a $m \times 1$ vector contains unobserved variables and unknown fixed effects.

Z_t is a $p \times m$ matrix links the unobservable factors and regression effects of the state vector with the observation vector.

T_t is a $m \times m$ matrix called the transition matrix.

η_t is a $r \times 1$ vector contains the state disturbances.

Q_t is a $r \times r$ matrix contains the covariances of the state disturbances.

Equation 4.3 is the observation or measurement equation, and equation 4.4 is the state or transition equation.

To clarify the use of state space model for multivariate time series analyses, we considerer a case with $p = 2$ of the multivariate trend model with regression effects.

If we denote the first time series considered by A_t, and the second by B_t $(t = 1, ..., n)$, the measurement equations of this bivariate state space model can be written as

$$A_t = \mu_t^{(1)} + \delta_t^{(1)} x_t^{(1)} + \varepsilon_t^{(1)},$$
$$B_t = \mu_t^{(2)} + \delta_t^{(2)} x_t^{(2)} + \varepsilon_t^{(2)},$$

with $\varepsilon_t^{(i)}$ $(i = 1, 2)$ denote the observation disturbances, $x_t^{(i)}$ $(i = 1, 2)$ are explanatory variables, which can be either continous or dummy intervention variables, and $\delta_t^{(i)}$ $(i = 1, 2)$ are unknown regression coefficients. The superscripts (1) and (2) denote whether they belong to the first or to the second series, respectively. Modelling the unobserved $\mu_t^{(i)}$ as local lineal trends, those for matrix equation 4.3 result in the following four state equations

$$\mu_{t+1}^{(i)} = \mu_t^{(i)} + \beta_t^{(i)} + \xi_t^{(i)},$$
$$\beta_{t+1}^{(i)} = \beta_t^{(i)} + \varsigma_t^{(i)},$$

for $i = 1, 2$, where $\mu_t^{(i)}$ is a time-varying level component, $\beta_t^{(i)}$ is a time-varying slope component, and $\xi_t^{(i)}$ and $\varsigma_t^{(i)}$ are level and slope disturbances, respectively. The vectors and matrices in equation 4.3 and 4.4 can be set up as

$$z_t = \begin{pmatrix} A_t \\ B_t \end{pmatrix}, \quad \alpha_t = \begin{pmatrix} \mu_t^{(1)} \\ \beta_t^{(1)} \\ \delta_t^{(1)} \\ \mu_t^{(2)} \\ \beta_t^{(2)} \\ \delta_t^{(2)} \end{pmatrix}, \quad \eta_t = \begin{pmatrix} \xi_t^{(1)} \\ \varsigma_t^{(1)} \\ \xi_t^{(2)} \\ \varsigma_t^{(2)} \end{pmatrix}, \quad T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$R_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, Z_t = \begin{bmatrix} 1 & 0 & x_t^{(1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & x_t^{(2)} \end{bmatrix},$$

$$H_t = \begin{bmatrix} \sigma^2_{\varepsilon^{(1)}} & cov(\varepsilon^{(1)}, \varepsilon^{(2)}) \\ cov(\varepsilon^{(1)}, \varepsilon^{(2)}) & \sigma^2_{\varepsilon^{(2)}} \end{bmatrix},$$

$$Q_t = \begin{bmatrix} \sigma^2_{\xi^{(1)}} & 0 & cov(\xi^{(1)}, \xi^{(2)}) & 0 \\ 0 & \sigma^2_{\varsigma^{(1)}} & 0 & cov(\varsigma^{(1)}, \varsigma^{(2)}) \\ cov(\xi^{(1)}, \xi^{(2)}) & 0 & \sigma^2_{\xi^{(2)}} & 0 \\ 0 & cov(\varsigma^{(1)}, \varsigma^{(2)}) & 0 & \sigma^2_{\varsigma^{(2)}} \end{bmatrix}$$

The values of the variances of the disturbances on the diagonal of matrix Q_t determine whether the corresponding level and slope components are stochastic or deterministic. If a disturbance variance is zero, then the corresponding component is deterministic and fixed. If the value of the variance is larger than zero, on the other hand, then the corresponding component is stochastic, and its value therefore changes over time.

The observation equations of this bivariate local linear trend model only contain two explanatory variables, nevertheless the model can easily be extended to incorporate any number of explanatory variables, as well as include the seasonal component.

4.4 Goodness of fit

The goodness of fit of a model describes how it fits a time series of interest. The basic measure of goodness of fit is the *prediction error variance* (p.e.v), which is the variance of the one-step-ahead prediction errors in the steady state (Harvey, 1989).

However, the conventional measure of goodness of fit statistics for a structural model is the R^2 coefficient. Harvey (1989) recommends using the variance of the first difference rather than variances of the series to calculate the coefficients of determination R_s^2. The R_s^2 value measures the goodness of fit of the equations, with a random walk plus drift model in first differences around the seasonal means as benchmarks. This goodness-of-fit criterion is defined as

$$R_s^2 = 1 - SSE/SSDSM$$

where *SSE* is the residual sum of squares for a univariate time series model, and *SSDSM* is the sum of squares of first differences around the seasonal means. Notes that any model which has R_s^2 negative should be rejected.

On the other hand, it is important to run post-estimation diagnostics on way to assess whether or not the model adequately describes the data. Significance tests are based on three assumptions concerning the residuals of the analysis, which should satisfy the following three properties listed in decreasing order of importance: independence, homoscedasticity and normality (Durbin and Koopman, 2012). These three tests are based on what are known as the *standardised prediction errors*,

$$e_t = \frac{v_t}{\sqrt{F_t}}$$

where v_t are the prediction errors and F_t the variance of the one-step prediction errors v_t.

The assumption of independence of the residuals will be checked by the sample autocorrelation values r_k at lags 1 and 24, and the Box-Ljung portmanteau test based on 24 lags of the sample autocorrelation function.

The Box-Ljung statistic for lags $l = 1, ..., k$ is

$$Q(k) = n(n+2)\sum_{l=1}^{k} \frac{r_l^2}{n-l}$$

where, the residual autocorrelation for lag k its defined as

$$r_k = \frac{\sum_{t=1}^{n-k}(e_t - \bar{e})(e_{t+k} - \bar{e})}{\sum_{t=1}^{n-k}(e_t - \bar{e})^2}$$

with $\bar{e}$ the mean of the n standardised residuals.

That should be tested against a χ^2-distribution with $(k - w + 1)$ degrees of freedom, w the number of estimated hyperparameters, and α the critial value (usually at the 5% level).

$$Q(k) < \chi^2_{(k-w+1;\alpha)}$$

The second assumption, homoscedasticity, will be checked with the H(h) statistics of Goldfeld-Quandt test,

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=d+1}^{d+h} e_t^2}$$

with d the number of diffuse inital elements, and h is the nearest integer to $(n-d)/3$.

This statistic should be tested against an F-distribution with (h, h) degrees of freedom and $\alpha/2$ as the critical value. Notes that if H(h) is larger than 1, it have to check $H(h) < F(h, h; \alpha/2)$, otherwise have to use the reciprocal of H(h), and check $1/H(h) < F(h, h; \alpha/2)$.

Finally, the Bowman-Shenton test (N statistics) will be used to test the normally distribution of the residuals.

$$N = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right)$$

where S is the skewness of the residuals, and K the kurtosis.

$$S = \frac{\frac{1}{n}\sum_{t=1}^{n}(e_t - \bar{e}_t)^3}{\sqrt{(\frac{1}{n}\sum_{t=1}^{n}(e_t - \bar{e}_t)^2)^3}}, \quad K = \frac{\frac{1}{n}\sum_{t=1}^{n}(e_t - \bar{e}_t)^4}{(\frac{1}{n}\sum_{t=1}^{n}(e_t - \bar{e}_t)^2)^2} \ .$$

This statistic asymptotically has a χ^2 distribution with two degrees of freedom. Then it has to check if $N < \chi^2_{(2;\alpha)}$ to no reject the null hypothesis of normality.

4.5 Structural Models versus ARIMA

As was introduced in Section 4.1, ARIMA models can be formulated in state space form, as well as many structural models admit an ARIMA representation. Actually, in the case of time series have a simple underlying structure both state space formulation and ARIMA formulation are basically equivalent to each other, by contrast, the differences between both approaches become more apparent when the structure of the time series is more complex.

In the ARIMA methodology, the trend and seasonal components can be eliminated by applying differences to the original time series, which may have strong inconveniences cause in many applications, like in tourism or econometric context, those components could have an interest in themselves. Furthermore, in some cases, this methodology implies differentiating the time series to satisfy the requirement of stationary. In sum, this approach is commonly known as a "black-box" in which the finally selected model does not take account of a prior analysis to study the structure underlying the generating system since the adopted model depends entirely on the data. In this sense, structural models are more flexible and transparent.

On the other hand, the incorporation of explanatory variables, structural breaks or missing observations is more difficult and not always are immediate in the ARIMA approach in comparation with structural models (Brockwell and Davis, 1991; Harvey and Koopman, 1992; Durbin and Koopman, 2001). Also, the structural models have the advantage of the associated regression coefficients of the explanatory variables can be treated stochastically.

Finally, it should be noted that despite state space models and in particular structural models have many advantages comparing with ARIMA models cause the first ones are very general and cover very wide range of models including all ARIMA models, the structural models have de disadvantage of the relative lack in the knowledge and software regarding these models. In this sense, ARIMA models count with the advantages of having numerous textbooks and software available about it (Durbin and Koopman, 2001).

To summarise, structural models are based on modelling the structure of the time series and its components, being more general, flexible and transparent than ARIMA models. However, nowadays the use of ARIMA models are more frequents due to its wide knowledge and the vast variety of software available (Durbin, 2000).

Chapter 5
Study of stochastic and deterministic trend in State Space Models

5.1 Introduction

As explained in Chapter 4, the descriptions of the seasonal and cycle components are understood as repetitive or stationary behaviour in short and long periods, respectively. Although they can be introduced with different formulations in state space models, there is usually a broad agreement. The functions demanding that the sum of the stationary effects on average be equal to zero, or trigonometric formulations, are commonly used (Durbin and Koopman, 2012). Another kind of function can also be used to express these components, such as polynomial functions by sections or splines (Koopman, 1992; Martin and Murillo, 2005).

However, there is usually a bit more diversity and controversy on the trend component. Highlight the section entitled "*What is a trend?*" in Harvey (1989, p. 284), where among other possibilities, the trend is defined as "*a general direction and tendency*", "*a trend is a trend...*" and, from the prediction approach, "*part of the series which when extrapolated gives the clearest indication of the future long-term movements in the series*".

The trend description of the series points out the particular state space model that will be used. The simplest model considers that trend is explained by a level component, μ_t, that can be constant or variable over time. In the local level model, the trend variation over time is linearly related to the level information in the previous period. In order to explain the trend behaviour, aside from the level of the series, the slope of the trend is usually included, β_t, which can again be constant or variable over time. If the variation over time is related to the slope in the previous period, then the local linear trend model will be developed. On the local linear trend model, Durbin and Koopman (2012) expose that

"*Applied workers sometimes complain that the series of values of μ_t obtained by fitting this model does not look smooth enough to represent their idea of what a trend should look like*".

Other possibilities to express the trend are developed by Young and collaborators and, in particular, Young et al. (1991) propose models where the level and slope state system is replaced by an autoregressive model of order r given by $\bigtriangledown^r \mu_t = \zeta_t$, with ζ_t normally and independently distributed random error with mean zero and variance σ^2_ζ. It is also possible to introduce a damping factor in the slope component (Harvey, 1989, p. 46) or, to use alternative formulations with growth curves (Harvey, 1989, p. 295).

The incorporation of the slope in state space models of the deterministic way together with stochastic level produces a random walk plus drift (Harvey, 1989, p. 170). Other similar versions for this situation were already introduced in Harrison and Stevens (1976) and West and Harrison (1997) in a Bayesian context and dynamic linear model, respectively.

In applications on real data, the use of models including stochastic level and slope to describe the trend of the series, both in the univariate and multivariate versions, are frequent in the bibliography (Hghiem et al., 2016; Antoniou and Yannis, 2013). In some of these articles, the interpretation of the results is made based on the estimated slope values obtained. Highlight that, in particular cases of state space models in continuous time, the model can be interpreted as a local linear model with stochastic slope, the level and slope being correlated (Koopman et al., 2017; Harvey, 1989, p. 487).

This chapter aims to analyse the influence of the inclusion of the stochastic slope and level to describe the trend in structural time series using state space models. The purpose is to check if the estimates obtained by the Kalman filter in these situations are able to discriminate between the slope and the level or, if these models can be simplified by models with deterministic slope, including stochastic level with a biggest variance. A secondary objective is to show how some goodness-of-fit criteria can promote the use of these models even when reality presents a steep slope. In addition, it should be noted that the inclusion of the stochastic slope, in general, helps to better describe the trend of the series, but that its estimates should not be used for purpose of discrimination.

The software for all the analyses presented in this chapter was written in R, with the *KFAS* package as it contains the main functions for filtering and smoothing structural time series and apply the Kalman Filter. Furthermore, it was used the accuracy function of the *forecast* package for calculating the goodness of fit of the estimates. Additionally, it was validated using STAMP 8.30 software (Koopman et al., 2010). For a list of candidate software to analyse this kind of data see Commandeur et al. (2011).

5.2 Methodology

For the purposes of this study, it considered the simplest case where observations $\{z_t\}_{t=1,..,T}$ oscillate around of a medium level that can remain constant or vary over time.

There are many variants of the local level model, which allows the level of the series varies in time depending on the level of the previous time plus a random element. The most usually used incorporate the slope associated with the level of the series. This slope can remain constant or change over time and it will explain, together with the level, the trend of the series (Section 4.2). This model, the local linear trend model, can be formulated as follows

$$\begin{aligned} z_t &= \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_t + \eta_t \\ \beta_t &= \beta_{t-1} + \xi_t \end{aligned}$$

with $\varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$, $\eta_t \sim NID(0, \sigma^2_\eta)$ and $\xi_t \sim NID(0, \sigma^2_\xi)$ beign σ^2_ε, σ^2_η and σ^2_ξ the variances of the irregular component of the series, the level and slope, respectively.

The considerations over σ^2_η and σ^2_ξ, that both, one or none are equal to zero, will lead to the different studied models:

1) Deterministic ($L_D S_D$), where both are fixed and equal to zero;

2) Stochastic level and deterministic slope ($L_S S_D$), σ^2_η is fixed equal to zero;

3) Deterministic level and stochastic slope ($L_D S_S$), σ^2_ξ is fixed equal to zero; and

4) Stochastic level and slope ($L_S S_S$).

5.2.1 Simulation process

In the generation of the simulations used, following a similar process described by Durbin and Koopman (2012, p. 26), it has fixed the value of σ^2_ε and the q-ratio$_\eta$ and q-ratio$_\xi$. The latter are defined as the quotient between the level and slope variances, respectively, against the irregular component variance, σ^2_ε, this is, q-ratio$_\eta = \sigma^2_\eta / \sigma^2_\varepsilon$ and q-ratio$_\xi = \sigma^2_\xi / \sigma^2_\varepsilon$.

The value of the σ^2_ε was fixed at 25 and the values of both q-ratios varied between 0 and 20 with a step of 1. In this way, models with q-ratio$_\eta$ close to 20 will involve a strong presence of the level component in the series and q-ratio$_\xi$ close to 20 of the slope component. For the beginning of the simulations it has been considered that $\mu_0 = 3000$, $\beta_0 = 0$, and $T = 143$. The fixed value of σ^2_ε to 25 does not influence the results. Its effect is simply picked up in relation to the rest of the variances through the definition of the q-ratios. The value of μ_0 has been set to 3000 with the intention of avoiding that the series present values close to zero and could interfere in the calculation of the goodness of fit measures. The fixed size of the series is a usual value in real situations, corresponding approximately to 12 years with monthly data.

Although a sensitivity analysis is included for the different models and grid of q- ratios considered, seven scenarios have been selected on which data are shown in more detail:

1. q-ratio$_\eta$ = 0 and q-ratio$_\xi$ = 20. Scenario with strong presence of the slope component with deterministic level.
2. q-ratio$_\eta$ = 1 and q-ratio$_\xi$ = 1. Scenario with similar presence of the level and slope components to the irregular component.
3. q-ratio$_\eta$ = 1 and q-ratio$_\xi$ = 20. Scenario with strong presence of the slope component.
4. q-ratio$_\eta$ = 1 and q-ratio$_\xi$ = 10. Scenario with moderate presence of the level component.
5. q-ratio$_\eta$ = 10 and q-ratio$_\xi$ = 10. Scenario with moderate presence of both the level component and the slope.
6. q-ratio$_\eta$ = 20 and q-ratio$_\xi$ = 1. Scenario with strong presence of the level component.
7. q-ratio$_\eta$ = 20 and q-ratio$_\xi$ = 20. Scenario with strong presence of both the level component and the slope.

Additionally, the scenario with both q-ratios equal to zero, where both the level and the slope are deterministic, is also included. All the scenarios are assuming that the data generation model is stochastic both in level and slope, except those in which one or both q-ratios are equal to zero.

The evaluation calculations of the different models were carried out over 100 randomly generated series. All of them maintained similar behaviour patterns when using the models described. In particular, Figure 5.1 displays examples of the series in the eight selected scenarios.

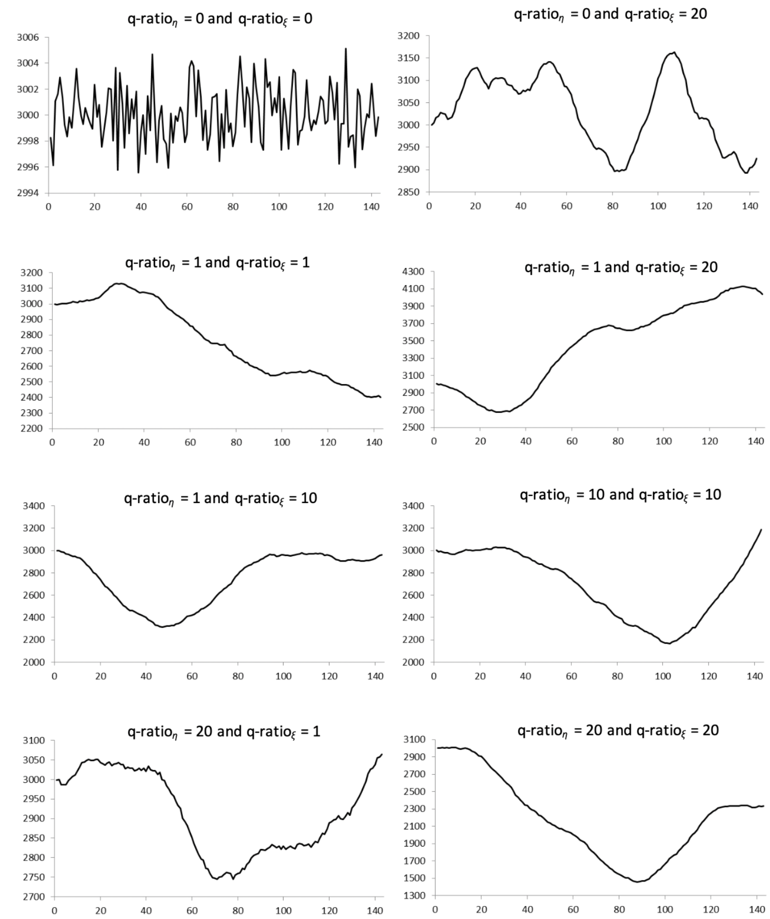

Figure 5.1: *Behaviour of time series studied based on the q-ratio$_\eta$ and q-ratio$_\xi$ used in the simulation process.*

5.2.2 Model selection and estimates evaluation

As was explained in Section 2.5 there are numerous measures of the relative quality of a candidate model and of the measures of accuracy. This study has taken into account the scores of AICc and BIC and the values of MAPE and Theil's U.

Aditionally, the Mean Magnitude of Relative Error, MMRE, will be employed for the evaluation of the estimated values of the level and slope in the different periods (Conte et al. 1986). It is defined as

$$MMRE_{\mu}(\%) = \frac{\sum_{t=1}^{T} |(\mu_t - \hat{\mu}_t)/\mu_t|}{T}100$$
$$MMRE_{\beta}(\%) = \frac{\sum_{t=1}^{T} |(\beta_t - \hat{\beta}_t)/\beta_t|}{T}100$$

Low values close to zero will be an indicative of good estimate.

5.3 Results

The estimated values of the fixed parameters of the models considered in the eight scenarios selected are shown in Table 5.1 and Table 5.2. Furthermore, Table 5.1 includes the scores reached of the selection criteria of the models used, as well as goodness of fit of the estimates. The boxes with the minimum values have been marked in gray for simple reading.

The variance of the irregular component, σ_{ε}^2, was always underestimated in the $L_S S_D$ model, compensated with a high overestimation of the level variance. The exception occurs for those instances where the real slope is deterministic ($\sigma_{\varepsilon} = 0$), as can be seen in Figure 5.2b), in which $\hat{\sigma}_{\varepsilon}$ remains close to the real value 5 and $\hat{\sigma}_{\eta}$ close to its real values too. In particular, in the scenario with $\sigma_{\varepsilon} = 0$ and $\sigma_{\eta} = 0$, that is, when both the slope and the level are deterministic, this model provides a estimation value of 4.82 for σ_{ε} and 0.0145 for σ_{η} (see Table 5.1).

Nevertheless, both values of MAPE and Theil's U for $L_S S_D$ model, submitted lowest values compared with the rest of the models at all points in the studied grid. In the scenarios shown in Table 5.1, it can be seen how these values can be even up to 10000 times lower compared to the models that follow it. The exception occurs in the scenario where both level and slope are deterministic, in which the four considered models presents similar values.

The L_SS_S and L_DS_S models, i.e. models with stochastic slope, have the lowest values of both AICc and BIC. The second of the models, L_DS_S, reaching the minimum values in those scenarios in which the level variance is relatively low in comparison to slope variance. However, the values of MAPE and Theil's U are higher than obtained with the L_SS_D model. The values of estimated variances in these two models, L_SS_S and L_DS_S, it remains more stable and close to the real values used in the simulation than in the L_SS_D model (Figure 5.2c and Figure 5.2d). In the L_SS_S model case, the value of $\hat{\sigma}_\varepsilon$ remained close to the real value 5, showing some overestimation when σ_η is high with respect to σ_ξ and underestimating in the opposite. In this case, the values of $\hat{\sigma}_\eta$ compensate for loss or excess shown in $\hat{\sigma}_\varepsilon$. The values of $\hat{\sigma}_\xi$ are close to their real values, although, in general, with certain underestimation.

When the model with deterministic level and stochastic slope is applied, L_DS_S, the results are like those of the L_SS_S model. In the scenarios where σ_η is high, the estimated values of σ_ε are overestimated, about twice, also compensating the non-inclusion of the level in the model with an overestimation of σ_ξ. While in scenarios with σ_η low, the results are fairly close to the real values. For example, in the scenarios with $\sigma_\eta = 0$, both values of $\hat{\sigma}_\varepsilon$ and $\hat{\sigma}_\xi$ quasi agree with their real values (Figure 5.2c). Thus, in the scenario with q-ratio$_\eta = 0$ and q-ratio$_\xi = 20$, $\hat{\sigma}_\varepsilon = 5.05$ and $\hat{\sigma}_\xi = 22.53$ (see Table 5.1), quite similar to the real values 5 and 22.4, respectively.

Figure 5.3 shows the residuals of the series using the four considered models in the scenario corresponding to a q-ratio$_\eta = 20$ and q-ratio$_\xi = 1$. As shown by MAPE and Theil's U values, the L_SS_D model is the best fitting the original series, while L_DS_D model presents greater disturbances, as expected. The L_DS_S and L_SS_S models graphically are very similar, both with residuals greater than those obtained with the L_SS_D model, even though the values of AICc and BIC for these models are the lowest.

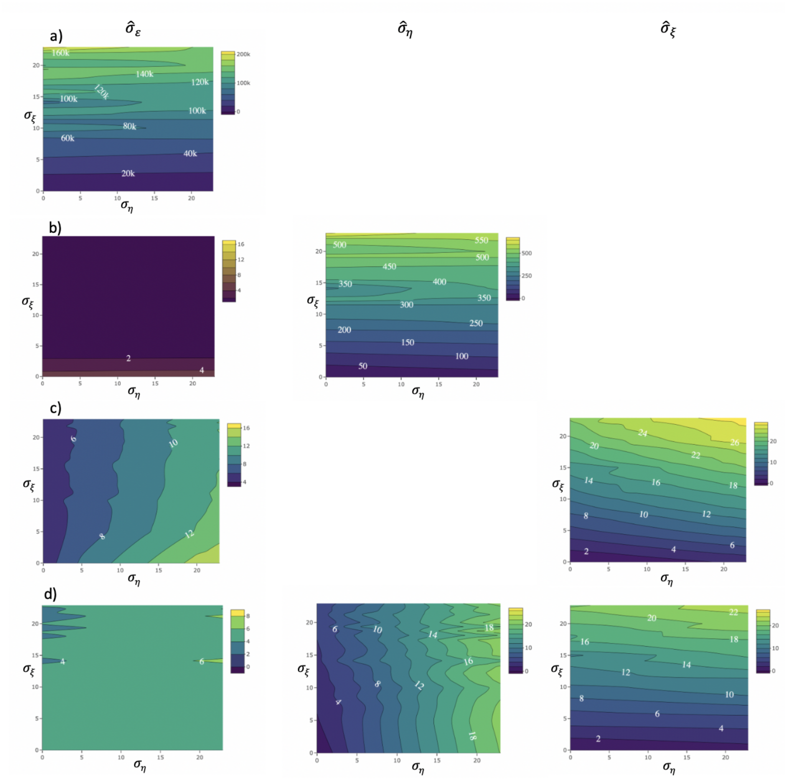

Figure 5.2: *Behaviour of standard deviation of the irregular, level and slope component ($\hat{\sigma}_\varepsilon$; $\hat{\sigma}_\eta$ and $\hat{\sigma}_\xi$; respectively, in each column) for the different fixed models: a) deterministic; b) stochastic level and deterministic slope; c) deterministic level and stochastic slope; and d) stochastic, depending on the real values of standard deviation of the level and slope (x and y axis, respectively).*

Table 5.2 shows the real and estimated values of the slope in the first period, as well as the average of the difference relativized between them in all of times, MMRE. Of all the scenarios and models tested the best estimates of the slope, as well as the level, it was obtained in the $L_D S_S$ and $L_S S_S$ models applied to the generated data in scenarios with small q-ratio$_\eta$ and high q-ratio$_\xi$. In the scenarios with q-ratio$_\eta$ and q-ratio$_\xi$ both high, $L_S S_S$ model presents the lowest values in the $MMRE_\beta$, although

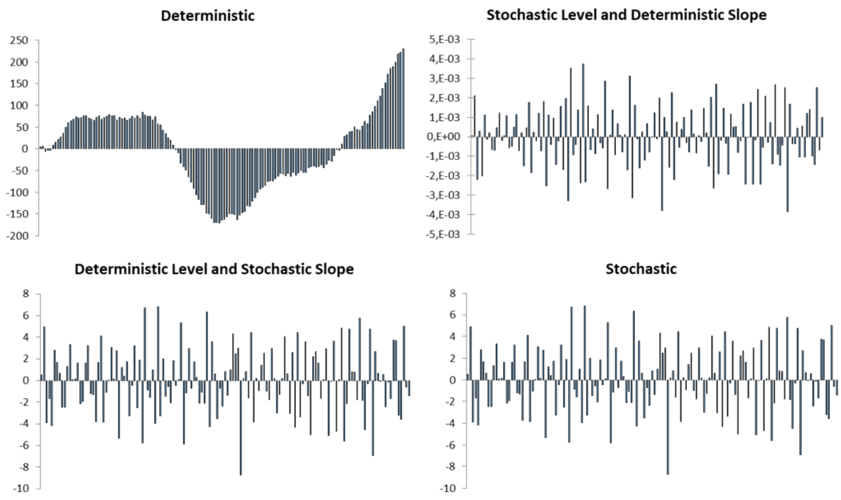

Figure 5.3: *Residual, difference between real value and estimations, of the simulation with q-ratio$_\eta = 20$ and q-ratio$_\xi = 1$.*

close to those obtained with the $L_D S_S$ model. In the case of $MMRE_\eta$, the behaviour is similar but with the roles reversed. For the scenarios with high q-ratio$_\eta$ versus q-ratio$_\xi$, $L_S S_D$ model is the best, both in slope and level estimation. Note that in the case of deterministic model, the goodness of fit values were always greater than in the rest of models, although with worse results in the estimation of the level than in the slope.

5.4 Discussion

The present study shows that the inclusion of stochastic slope in state space models, even when this component does not have a predominant role, it provides models with low values of AICc and BIC. Nonetheless, this kind of models can provide values of MAPE and Theil's U higher than the models with deterministic slope, even in situations where reality shows a strong component of the slope. Harvey (1989, p. 169) indicates that the formulation of unobserved components does not have to be unique. In Lawson et al. (2011), the comparison of different models applied to real data is done, providing tables where analogous situations to those obtained in this paper can be seen. That is, models with the lowest values of MAPE and RMSE that do not correspond with the minimum values of AIC, although the interpretation and selection of the models in said paper was done using the MAPE and RMSE.

Many references after analyzing different models on the available data determine that the models without stochastic slope are the best. Wang and Getz (2007) compare different models of state space and other types of models to explain seasonal fluctuations of vole and shrew populations, choosing them with non-stochastic slope. The data typically used to introduce state space models, United Kindow drivers KSI and Norwegian fatalities (Commandeur et al., 2011), are other examples of preferences of models with deterministic slope. In the analysis of these data, the authors explicitly indicate (p. 28) that: "*the addition of a slope component to the local level model is not effective in improving the description of the observed time series*". In all these examples, the criterion employed for its selection was to choose the model with the lowest AICc. Commandeur et al. (2013) compare regression, ARIMA and state space models, among others, on data of the evolution of road safety, concluding that saving the differences in their formulations, these have a lot in common. State space models employed for this kind of problems do not include the slope as it was introduced in this paper; instead, it is included the time as an exogenous variable fixing the corresponding parameter in a deterministic way.

Despite the fact that in the vast majority of the selected models suggest stochastic slope using the AICc or BIC criterion, this is not equally reflected in the use of MAPE and Theil's U estimation criteria. In addition, as Table 5.2 shows, the estimates of the level and slope can even not being estimate coherently. Thus, for example, in situations where the variances of the slope and the level are high, the best values of the level estimates are obtained in models with deterministic level. When reality presents a trend with a low level against the slope then models that present stochastic slope produce best precisions on the slope as well as on the level. In a multivariate context, Nghiem et al. (2016) make part of the conclusions of their study based on the estimated values of stochastic slopes. Harvey (1985) in the context of macroeconomic time series had already noted that: "*The structural time series model is not intended to represent the underlying data generation process. Rather, it aims to present the facts about the series in terms of a decomposition into trend, cycle, seasonal, and irregular components. These quantities are of interest to economists in themselves*".

5.5 Concluding remarks

The trend of the series discomposed in its level and slope has been simulated in different scenarios, checking how classic state space models work on them. The selected models estimate the trend accurately but not necessarily the contribution of the level or slope. The tests performed shows how the AICc or BIC criteria tend to suggest the use of models with stochastic slope even in situations where it does not have a relevant presence. In addition, models with a deterministic level, L_DS_S, could be preferred to stochastics in many of the situations where reality introduced both components in a random way. The non-introduction of the stochastic level would produce an increase in the variance of the slope as well as the variance of the irregular component. Unlike model selection criteria, the goodness of fit measures of the estimates, such as MAPE and Theil's U, prioritize mainly models with deterministic slope. In the latter case, the estimated values of the level variance tend to be much larger than the real values, and even producing an underestimation of the irregular component variance. Based on the estimated parameters of level and slope, results tend to show success in situations where just one of these components is large but not where both are high, being able in these last situations to have an information transfer among them and so making difficult their interpretation.

Table 5.1: *Estimated parameters, model selection criteria and goodness of fit.*

Real	Models	Estimations						
$\sigma_\varepsilon = 5$		σ_ε	σ_η	σ_ξ	AICc	BIC	MAPE	Theil's U
q-ratio$_\eta = 0$	L_DS_D	4.91	-	-	642.47	645.41	0.059	0.707
q-ratio$_\xi = 0$	L_SS_D	4.82	0.0145	-	644.08	649.92	0.058	0.696
($\sigma_\eta = 0$)	L_DS_S	4.86	-	0.0002	644.52	650.36	0.059	0.701
($\sigma_\xi = 0$)	L_SS_S	4.82	0.0124	0.0014	646.14	654.86	0.058	0.696
q-ratio$_\eta = 0$	L_DS_D	275354.52	-	-	2056.65	2059.58	23.562	10.670
q-ratio$_\xi = 20$	L_SS_D	0.006	811.95	-	1278.70	1284.54	7.39E-06	55.10E-06
($\sigma_\eta = 0$)	L_DS_S	5.05	-	22.53	939.05	944.89	0.122	0.053
($\sigma_\xi = 22.4$)	L_SS_S	4.05	4.93	19.74	940.83	949.55	0.072	0.041
q-ratio$_\eta = 1$	L_DS_D	29321.21	-	-	1778.13	1781.09	10.993	14.837
q-ratio$_\xi = 1$	L_SS_D	0.003	113.19	-	1046.76	1052.60	6.83E-06	2.30E-05
($\sigma_\eta = 5$)	L_DS_S	6.26	-	6.77	872.11	877.95	0.092	0.183
($\sigma_\xi = 5$)	L_SS_S	4.29	6.11	4.98	872.56	881.28	0.067	0.122
q-ratio$_\eta = 1$	L_DS_D	140308.82	-	-	2020.64	2023.58	13.620	11.523
q-ratio$_\xi = 20$	L_SS_D	0.008	473.44	-	1244.32	1250.16	5.85E-06	1.07E-05
($\sigma_\eta = 5$)	L_DS_S	6.55	-	22.96	956.41	962.25	0.063	0.068
($\sigma_\xi = 22.4$)	L_SS_S	4.49	8.20	19.32	957.86	966.58	0.040	0.051
q-ratio$_\eta = 1$	L_DS_D	130281.78	-	-	2027.08	2030.02	12.763	10.866
q-ratio$_\xi = 10$	L_SS_D	0.012	445.59	-	1244.52	1250.36	1.38E-04	1.35E-05
($\sigma_\eta = 5$)	L_DS_S	6.79	-	16.71	935.14	940.98	0.804	0.067
($\sigma_\xi = 15.8$)	L_SS_S	4.39	6.58	14.14	936.62	945.34	0.486	0.052
q-ratio$_\eta = 10$	L_DS_D	83390.47	-	-	1969.22	1972.16	12.453	10.088
q-ratio$_\xi = 10$	L_SS_D	0.008	384.45	-	1223.04	1228.88	2.25E-05	9.21E-06
($\sigma_\eta = 15.8$)	L_DS_S	10.42	-	21.20	983.31	989.15	0.191	0.096
($\sigma_\xi = 15.8$)	L_SS_S	4.83	15.14	15.67	983.70	992.41	0.123	0.049
q-ratio$_\eta = 20$	L_DS_D	21006.03	-	-	1767.52	1770.46	4.848	9.149
q-ratio$_\xi = 1$	L_SS_D	0.0004	113.01	-	1063.92	1069.76	1.38E-05	4.10-05
($\sigma_\eta = 22.4$)	L_DS_S	12.68	-	9.48	951.73	957.57	0.117	0.218
($\sigma_\xi = 5$)	L_SS_S	5.91	19.51	4.77	948.79	957.51	0.057	0.098
q-ratio$_\eta = 20$	L_DS_D	128770.94	-	-	1989.23	1992.17	12.770	10.621
q-ratio$_\xi = 20$	L_SS_D	0.008	516.60	-	1249.71	1255.55	9.28E-06	6.77E-06
($\sigma_\eta = 22.4$)	L_DS_S	12.10	-	28.54	1013.54	1019.38	0.108	0.101
($\sigma_\xi = 22.4$)	L_SS_S	4.81	20.84	25.70	1015.50	1024.22	0.039	0.039

L_DS_D Deterministic; L_SS_D Stochastic level and deterministic slope; L_DS_S Deterministic level and stochastic slope; L_SS_S Stochastic

Table 5.2: *Estimated and real values of the level and slope.*

$\sigma^2_\varepsilon = 25$	Real μ_1	Real β_1	Models	Estimations μ_1	Estimations β_1	$MMRE_\mu$	$MMRE_\beta$
q-ratio$_\eta = 0$ q-ratio$_\xi = 0$	3000	0	L_DS_D	3000.04	0.0009	0.0037	-
			L_SS_D	3000.04	0.0009	0.0037	-
			L_DS_S	3000.04	0.0009	0.0037	-
			L_SS_S	3000.04	0.0009	0.0037	-
q-ratio$_\eta = 0$ q-ratio$_\xi = 20$	3004.406	4.406	L_DS_D	3005.29	-0.125	12.4506	209.2250
			L_SS_D	2999.76	-0.109	2.5402	198.0151
			L_DS_S	2999.70	0.844	2.4941	97.1942
			L_SS_S	2999.76	-0.108	2.4939	98.0260
q-ratio$_\eta = 1$ q-ratio$_\xi = 1$	2999.567	-0.946	L_DS_D	3165.20	-5.576	2.1128	257.1898
			L_SS_D	2998.32	-4.204	0.2549	134.1081
			L_DS_S	2997.65	-0.002	0.2422	170.3788
			L_SS_S	2997.65	-0.0004	0.2422	170.2909
q-ratio$_\eta = 1$ q-ratio$_\xi = 20$	3002.587	-3.608	L_DS_D	2646.90	11.257	13.7845	186.1222
			L_SS_D	3003.54	7.301	0.4288	178.1078
			L_DS_S	3002.45	-4.006	0.4245	85.9449
			L_SS_S	3002.47	-4.027	0.4245	84.5617
q-ratio$_\eta = 1$ q-ratio$_\xi = 10$	2999.441	-1.917	L_DS_D	2559.49	2.587	6.3551	325.959
			L_SS_D	2999.65	-0.281	0.4087	161.230
			L_DS_S	3000.22	-4.657	0.3987	53.323
			L_SS_S	2999.78	-5.343	0.3961	53.651
q-ratio$_\eta = 10$ q-ratio$_\xi = 10$	3002.870	-4.156	L_DS_D	3371.85	-12.571	10.0855	145.087
			L_SS_D	2999.54	-10.797	0.8664	114.359
			L_DS_S	3000.48	4.015	0.8659	94.601
			L_SS_S	2999.54	2.440	0.8664	94.293
q-ratio$_\eta = 20$ q-ratio$_\xi = 1$	3000.384	-0.665	L_DS_D	2994.61	-1.122	2.7136	161.0802
			L_SS_D	2999.24	0.461	0.1924	117.2115
			L_DS_S	2998.72	-4.194	0.2048	179.4998
			L_SS_S	2998.77	-4.187	0.2048	179.4302
q-ratio$_\eta = 20$ q-ratio$_\xi = 20$	2998.802	-0.536	L_DS_D	2640.80	-6.361	17.6268	181.5151
			L_SS_D	3000.63	-4.692	0.8872	153.9145
			L_DS_S	3000.81	2.732	0.8777	144.3008
			L_SS_S	3000.63	2.109	0.8851	143.5288

L_DS_D Deterministic; L_SS_D Stochastic level and deterministic slope; L_DS_S Deterministic level and stochastic slope; L_SS_S Stochastic

Chapter 6
Tourism demand forecast at the Canary Islands

6.1 Introduction

International tourism contributes significantly to the economic growth of many countries and regions, becoming increasingly important to worldwide economic development. Taking into account the fast increase in international tourism demand over the last few decades, accurate forecasts of future tourism demand trends are necessary to make good decisions.

Traditional quantitative forecasting methods include time series analysis and econometric models as well as a combination of both. However, in terms of forecasting, it is difficult to find a model that is better in all the criteria, and each model has its pros and cons (Song and Li, 2008).

Excessively large forecast horizons in tourism studies are often unreliable unless we are in very stable scenarios. There are different types of methods that could be employed to carry out the prediction models. Particularly, integrated autoregressive methods of seasonal moving averages (SARIMA and SARIMAX) in situations of stability usually produce satisfactory results short-term predictions (Peña, 2010; López et al., 2017; and Box et al., 2016). These type of methods has often been criticised for operating as a "black box", as was explained in Section 4.5, where the time series components, seasonality, trend, cycle, etc., are combined such as is not possible a disaggregation of the effect of each of them (Commandeur and Koopman, 2007). For these reasons, an approach using both the state space model and the SARIMAX model is carried out and the results are compared.

International tourists who arrive at a certain tourist destination may come from countries or regions with similar or different customs and behaviours, and therefore be able to present correlated arrival patterns (Chen et al., 2019). The existence of related patterns will boost the use of multivariate models to estimate and forecast, which take into account and benefit from the contribution of extra information that these relationships entail. On the contrary, trying to use additional information provided by origins with unrelated patterns in multivariate models could prejudice the goodness of fit of the models.

One of the main advantages of structural time series models is the capacity of modelling multivariate models. Many countries are connected touristically, having the same behaviour, and as a result, one event in a country can affect other countries' behaviour. In these cases, the multivariate model will give appropriate information about how the behaviour of the countries move together under each other's effects and structural time series models are a useful tool for it. These models are a powerful method for forecasting and analyzing multivariate time series because it detects common behaviours explicitly to show the short-term dynamics of the time series and it forecasts each component separately. With the purpose of measure of the influence of the tourist behaviour of one country concerning another, in this thesis individual univariate structural time series models are built for each country and then extend the model to multivariate case.

This chapter aims to analyse the reliability of the ARIMA models and structural models in the tourism demand problems in which there is a strong presence of a seasonal effect, as well as the influence of different exogenous variables that can model their behaviour. With this aim in mind, the arrival of national and international tourists to the Canary Islands is used. It also aims to evaluate the use of multivariate or univariate state space models in tourism demand forecasting and how their choice may be affected by the ratio of passengers in one country to the others. For this purpose, the structural time series model applies to the entrance of German and British tourists at Canary Islands, including exogenous variables and the results are compared with the airline passenger model with the inclusion of exogenous variables. Lastly, this chapter aims to promote structural models that not only allows estimating and forecasting the global tourist demand but also allows the components of the trend and seasonality.

In order to assess the effectiveness of the forecasting in the case of the entrance of German and British tourists at Canary Islands, for both the Time-Varying Parameter Structural Time Series model (TVP-STS) and the SARIMAX, the MAPE and Theil's U will be evaluated in different time horizons (one-, three-, six- and twelve step-ahead). In addition to these descriptive measures, we have proceeded to the comparison of the forecasting results of the different models using the modified Diebold-Mariano (M-DM) test given in Harvey et al. (1997).

All the analyses and results presented in this chapter were obtained with the computer package STAMP 8.30 (Koopman et al., 2010) and the R software. The passenger entry time series, as well as the values of the capacity variables, were logarithmically transformed to stabilize the variances. The calculation of MAPE, Theil's U and M-DM test were carried out on the untransformed values. To obtain the M-DM test and to evaluate the multivariate SARIMAX model, the *dm.test* function and *marima* library, respectively, of the R software was used.

6.2 Data and variables

Before detailing the information related to the data and variables, it should be clarified that reference will be made to the data until 2016 since it was a year of change in the tourist flow at the Canary Islands and it was considered a good time to testing the selected models and checking if they could be able to collect changes in both the trend and the pattern of the flow that had been followed so far.

Tourism plays a major role in Canary Islands because of its economic and employment potential, as well as its social and environmental impacts. It was responsible for 39.7% of the employment generated in that Autonomous Community in 2016, due to the 14.9 million tourists who visited the islands during that year. The number of tourists entering the Canary Islands has increased by 27.3% since 2012 to 2016. From the supply side, these islands have a tourist infrastructure that offered a total of 246,476 hotel beds and 169,634 extra-hotel accommodations in 2016 (ISTAC, 2017a), and therefore hotel beds representing the most important offer to stay in the Canary Islands with almost 110 million overnight stays (ISTAC, 2017a). Furthermore, the presence of tourism in the islands' economy has been increasing in recent years. In 2016, its contribution to the productive system reached 34.3%, with an average annual growth of 6% in the Tourist Gross Domestic Product (IMPACTUR, 2016). However, it is important to note that as with the economy as a whole, tourism activity has also suffered the consequences of the terrible economic crisis in which the world and Spain, in particular, were involved, with ups and downs periods in its activity since 2008.

The determination of tourism demand is basically an empirical issue. The passenger arrivals to a tourist destination isolated from the continent, such as the Canary Islands, is a good thermometer of tourism demand. The number of tourist arrivals at Canary Islands have been increasing during recent years, both in the period of world crisis and in the recovery period, surpassing 13 million of tourists in 2015 and reaching figures not seen until that moment in the summer of 2016. The number of tourists visiting the Canary Islands has increased by 13% from 2010 to 2015 (Figure 6.1). That record figures were favored by several aspects, among which the events that occurred in competing destinations stand out, economic recovery or the ever-increasing connectivity with archipelago.

The data used to analysis and forecast the Total Canary Islands of tourist are the passenger arrivals to the Autonomous Community of the Canary Islands provided by AENA (ISTAC, 2016c) for the period from January 2005 to May 2016.

To measure the degree of consumer response to the different factors that may condition their decision to travel, a series of aspects must be taken into account prior to the analysis. Stand out: study and detection of possible sporadic events or level changes that require the incorporation of intervention variables; selection of exogenous variables that could be affecting the demand trend; and lastly determination of the model and components that should be considered.

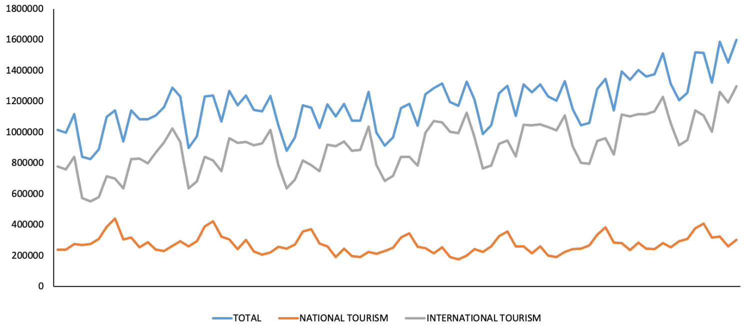

Figure 6.1: *Monthly number of passenger arrivals, for January 2010 to December 2016.*

6.2.1 Canary Islands inbound tourism

Canary Islands inbound tourism is diverse, receiving tourists from various destinations around the world, the contribution of countries not belonging to the European continent is simply 1% of the total since 2010. Ninety-nine percent comes from Europe, with 77% being international tourism and being national tourism of the rest of Spain the other 23% (Figure 6.2). There are two main international markets, United Kingdom and Germany, which contribute more than 63% on European international tourism, contributing with 27% and 19% of the total' tourism respectively. The rest of the origin countries have a contribution each less than 2%, except Netherlands and the Nordic countries, with 3% and 11% respectively as can be see in Figure 6.3 (INE, 2016).

Due to the good economy reflected in Germany, most of its inhabitants make at least one outward journey each year. This leads to Germany to become one of the countries that most travel abroad (GNTB, 2016). In 2016 there were around 94 million Germans who decided to cross the border to visit another country, being Spain, behind Australia, the second most visited by Germans. Regarding British tourists, Spain is still the most visited country (ONS, 2016), with 14,7 million visits and an increase of 13% between 2015 and 2016.

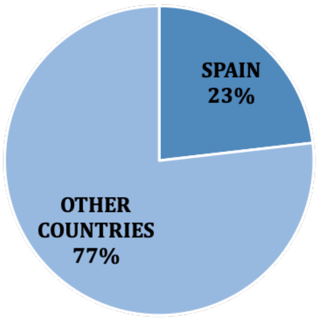

Figure 6.2: *Contribution of international and national tourism since 2010.*

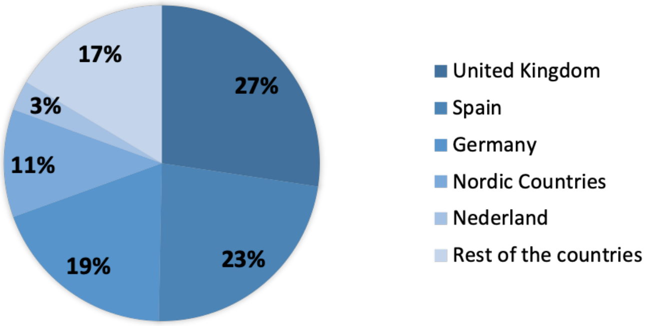

Figure 6.3: *Main international markets since 2010.*

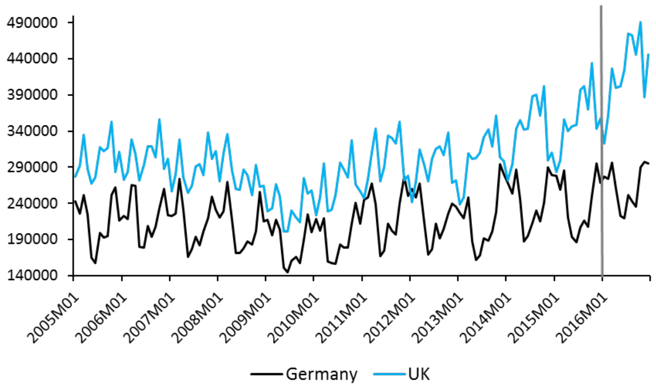

Figure 6.4: *Monthly number of Germans and British passenger arrivals, for January 2005 to December 2016.*

The data that will be modelled and forecasted, in this case, belong to the monthly arrival of German and British passengers to the Autonomous Community of Canary, provided by AENA (ISTAC, 2017b), using monthly data for the time 2005:01 to 2016:12, which generates 144 observations, leaving the last 12 (2016:01-2016:12) for forecasting purposes (Figure 6.4). For the calculation of the values of MAPE and Theil's U in the different time horizons, the model will be re-estimated adding each time one month. Obtaining thus, 12 one-step-ahead forecasts, 10 three-step-ahead forecasts, 7 six-step-ahead forecasts, and 1 twelve-step-ahead forecast.

6.2.2 Interventions

As explained in Section 2.6, a change in the trend of the time series by particular events can be modelled using variables that collect these effects. In order to better account for these external shocks to the process, it was decided to seek possible events that could be identified and explicitly entered into the model.

Interventions that are assumed to have a transitory effect on the level are coded as a pulse variable being 1 for the time point of the effect and 0 for the rest of the time points. On the other hand, interventions that are assumed to have a permanent effect, level change, are coded as a step variable being 0 for all time points prior to the event and 1 for the time point of the event and those following.

Five main events were detected that could be entered as interventions in the model for the period and data that are being analysed: terroristic attacks open competing destinations, Spanish economic recession, world economic recession, Japan earthquake and volcanic eruption in Iceland. As a confirmation process in the intervention study, the time series were deseasonalized and verified that these events effectively produced a change in trend (Figures 6.5, 6.6 and 6.7). An overview of all these interventions is given in Table 6.1, indicating the period of the intervention as well as the external event with which was associated.

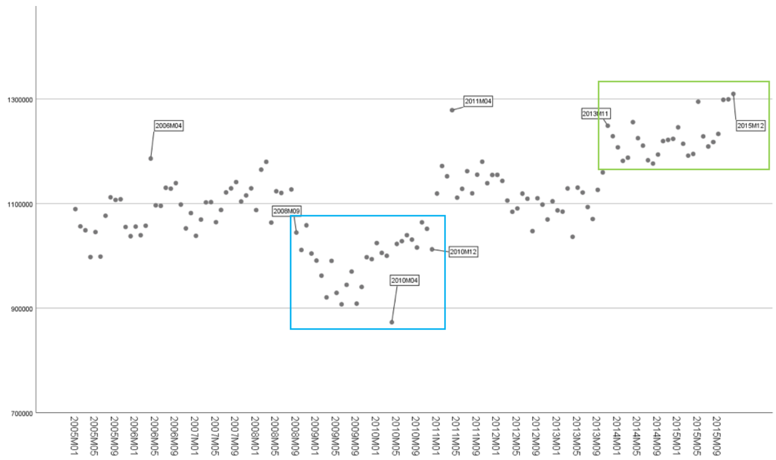

Figure 6.5: *Deseasonalized total of tourist at Canary Islands time series.*

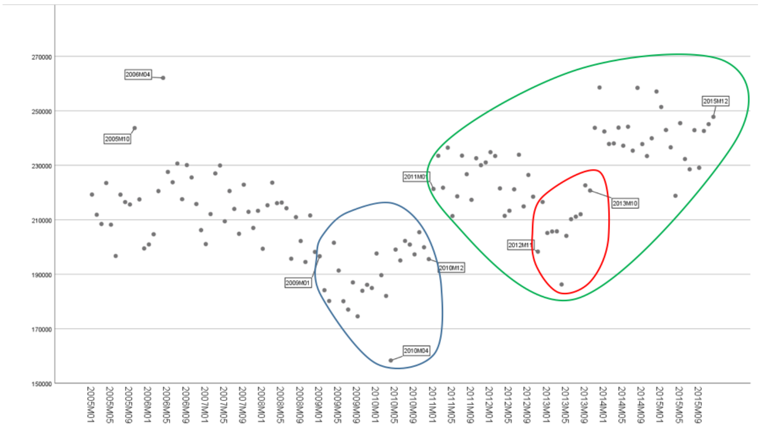

Figure 6.6: *Deseasonalized total Germans tourists time series.*

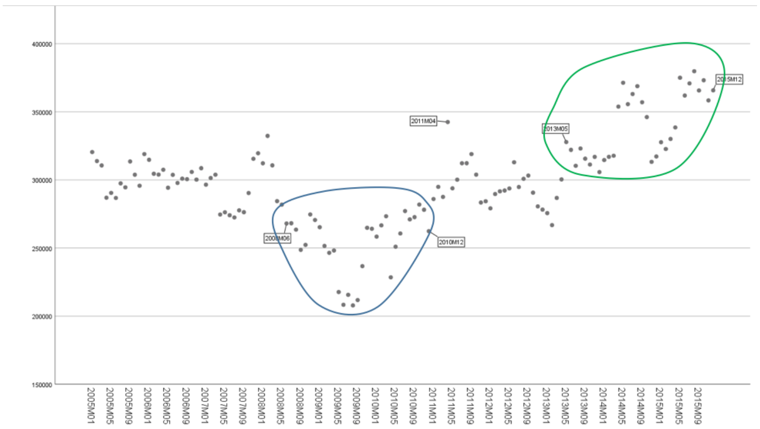

Figure 6.7: *Deseasonalized total British tourists time series.*

Despite detecting the same particular events in different time series, this event does not necessarily affect them in the same way. The Spanish economic recession is a clear example of this, which affects the time series of total tourists to a greater extent as it includes Spanish tourism. This event affected the total tourism for two years, had an effect of one year on Germans' arrived and did not affect British tourism (Table 6.1).

Table 6.1: *Overview of the intervention variables for the countries studied, including their type and cause.*

Reason	Period	Type	Total	Germany	United Kingdom
Terroristic attacks on	10:2005	Transitory effects		✓	
competing destinations	04:2006		✓	✓	
	05:2013 to 12:2016	Level change			✓
	11:2013 to 12:2016			✓	
Spanish economic recession	11:2012 to 10:2013	Level change		✓	
	11:2013 to 12:2015		✓		
World economic recession	09:2008 to 12:2010	Level change	✓		
	01:2009 to 12:2010	Level change		✓	
	06:2008 to 12:2010				✓
	05:2009 to 09:2009				✓
Japan earthquake	04:2011	Tansitory effect	✓		✓
Volcanic eruption in Iceland	04:2010	Tansitory effect	✓	✓	✓

6.2.3 Exogenous variables

With the purpose of identifying structure and pattern in the data, so to construct a model to predict future patterns, the estimation of the future values is accomplished using past observations from the series of interest. However, in some cases time series are influenced by other factors and past observations are not enough to explain its behaviour, being necessary the use of exogenous variables (Section 2.7).

Exogenous variables considered in this study have been subdivided into three sets.

Economic variables

Economic variables considered have been obtained from EUROSTAT (2017a). In particular, three variables have been included as proxy variables of the economic situation

- *Relative Consumer Price Index* (RCPI), defined as the ratio of Canary CPI against CPI of the country of origin corrected by the exchange rate, it measures the ratio of the price of good in the source market relative to the good or similar good in the destination market.

- *Economic Sentiment Indicator* (ESI), which is a composite indicator made up of five sectoral confidence indicators with different weights (Industrial, Services, Consumer, Construction and Retail trade confidence indicators).
- Standardized *Consumer Confidence Indicator* (CCI), that it is defined as the degree of optimism on the state of the economy that consumers are expressing through their activities of savings and spending.

Calendar variables

Time series like tourist arrivals are often influenced by seasonal effects or calendar effects (Koc and Altinay, 2007) that prevent a clear understanding of the behaviour of the series. Calendar effects are defined as the impact that occurs in the time series due to the different structure presented by the months in different years. Calendar variables considered are

- *Easter week*, defined as the number of days of the seven that are in the month of March and April in each of the years included.
- *Carnivals*, defined in a similar way account three day of festive.
- *Trading day*, representing the number of days of the week in each of the corresponding months, with respect to the number of Sundays of that month.
- Leap Year, indicating every four years in the month of February such situation.

Capacity variables

Capacity variables represent the number of airline seat reservations for regular flights, known as “slots”. The companies reserve in the different airports the number of flights and the size of the aircraft, which is translated into a number of possible seats. These variables are provided by Instituto Canario de Estadística (ISTAC), enabling with one-year advance to know the estimated reserves of the number of arrivals by origin country. In this study, we have considered three set variables corresponding to

1. Slots of Germany.
2. Slots of United Kingdom.
3. Slots of the total of tourist, that was introduced as five variables of the main markets (Figure 6.3).

6.3 Results of the Total Canary Islands of tourist analysis

In this section, the main results of the Box-Jenkins approach using the airline passenger models (3.2.6) and structural models to forecast the Total passengers' arrivals at the Canary Islands are presented. For that purpose, different models with the inclusion of the intervention variables detected in the previous section were tested by introducing the set of exogenous variables considered and its combinations. The possible combinations of the exogenous variables are the followings

- Without inclusion of exogenous variables (*Without*)
- Calendar Effects (*CE*)
- Capacity Variables (*SLOTS*)
- Economic Variables (EV)
- Calendar Effects and Capacity Variables (*CE and SLOTS*)
- Calendar Effects and Economic Variables (*CE and EV*)

- Capacity Variables and Economic Variables (*SLOTS and EV*)
- Calendar Effects, Capacity Variables and Economic Variables (*CE, SLOTS and EV*)

Table 6.2 shows the estimates and forecasts values of MAPE and Theil's U for the different airline passenger models according to the inclusion of the possible combinations of the exogenous variables. For the estimates, that is, in the period 2005:01 to 2015:12, it can be seen that the use of the economic variables only and the combination of the calendar effects with economic and capacity variables obtain the lower MAPE values, with values lower than 2%. This means that these models explain better than the others the past behaviour of the time series analysed, but this does not imply that they perform the best forecasts. The values of MAPE and Theil's U in the validation period (2016:01 to 2016:12) show that the airline passenger model with the intervention and the capacity variables is the best in the forecasts, with 2.39% and 0.26 values of MAPE and Theil's U respectively.

The results obtained using structural models are shown in Table 6.3. In this case, the use of the calendar effects was the best option to explain the behaviour of the time series in the estimates, followed by the model without the inclusion of any exogenous variables, having both values of MAPE lower than 1% and values of Theils' U close to 0. However, in the forecasting part, these models obtain the worse values of both measures. The model that includes only the capacity variables as exogenous variables is the best in the forecast.

The results with both approaches in the validation period are similar, achieving a little lower values with the structural models. However, in the estimates period, structural models present lower values of both accuracy measures than the airline passenger model. While the lowest MAPE value in the case of the airline passenger model is around 2%, the structural model taking into account the calendar effects obtain 0.32%.

Table 6.2: *Measures of fit for airline passenger models.*

Exogenous Variables	Estimates		Forecasts	
	MAPE	THEIL'S U	MAPE	THEIL'S U
Without	3.12%	0.30	5.96%	0.58
CE	2.38%	0.23	5.28%	0.53
SLOTS	2.56%	0.25	2.39%	0.26
EV	1.94%	0.19	4.53%	0.47
CE and SLOTS	2.01%	0.19	2.55%	0.28
CE and EV	2.18%	0.214	4.68%	0.48
SLOTS and EV	2.41%	0.24	2.84%	0.27
CE, SLOTS and EV	1.97%	0.19	3.08%	0.32

CE Calendar Effects; SLOTS Capacity Variables; EV Economic Variables

Table 6.3: *Measures of fit for casual structural model.*

Exogenous Variables	Estimates		Forecasts	
	MAPE	THEIL'S U	MAPE	THEIL'S U
Without	0.88%	0.00	5.75%	0.63
CE	0.32%	0.00	5.11%	0.51
SLOTS	1.23%	0.08	2.11%	0.21
EV	2.10%	0.19	4.58%	0.49
CE and SLOTS	2.05%	0.20	2.45%	0.27
CE and EV	2.36%	0.22	4.15%	0.36
SLOTS and EV	1.41%	0.11	2.56%	0.28
CE, SLOTS and EV	1.93%	0.18	3.05%	0.31

CE Calendar Effects; SLOTS Capacity Variables; EV Economic Variables

6.4 Results of the main markets analysis

Depending on the treatment of the unobserved components: the level, seasonal and slope components, in a stochastically or deterministic way, several types of models can be obtained (Section 4.2). In order to select the best model, taking into account the selection criteria of models and goodness of fit measures described, it was concluded that a stochastically treated of the components was the better model in the case of UK, whereas for the Germany data the best model was the one that fixed on zero the slope variance and treated the level and seasonal components stochastically. Additionally, the economic and capacity variables were initially introduced stochastically, obtaining in all cases variances of δ_i coefficients of the exogenous variables less than 1.175e-7. This prompted that they were fixed in a deterministic way.

This section presents the main estimation results of the univariate and multivariate state space models applied, displaying the explicative power of the models and conducting diagnostic tests. Table 6.4 shows the estimates of the hyperparameters obtained for both Germany and the United Kingdom in the univariate and multivariate versions. In addition, evaluation parameters of the Goodness of fit are included, such as Prediction Error Variance (PEV), Akaike Information Criteria (AIC) and Coefficient of global and seasonal determination (R^2 and R_s^2). In the case of Germany, the values obtained for all the parameters shown in the univariate and multivariate version are very similar, while in the case of the United Kingdom the parameters in their multivariate version have a better behaviour, with lower values of AIC, variances of the hyperparameters and PEV, and higher coefficients of determination. The coefficient of determination in Germany is 0.97 and 0.96 in the United Kingdom, while the coefficient of differences around a seasonal mean is 0.79 and 0.69 respectively. The German and United Kingdom models present very similar AIC values, around -6.4 in both.

The estimated coefficients for the different variables included: interventions, economics, calendar and capacity are collected in Table 6.5. Capacity variable (slots) was significant, presenting a direct relationship with the entrance of passengers, both in Germany and the United Kingdom in its univariate and multivariate versions, being of the set of exogenous

Table 6.4: *Overview of the intervention variables for the countries studied, including their type and cause.*

PARAMETERS	Germany		United Kingdom	
	UNIVARIATE	MULTIVARIATE	UNIVARIATE	MULTIVARIATE
Hyperparameters estimated				
σ^2_ε	2.07E-04	1.93E-04	6.74E-05	4.63E-06
σ^2_η	3.95E-04	3.89E-04	7.64E-04	2.33E-04
σ^2_ξ	-	-	6.38E-07	6,30E-07
σ^2_ω	2.75E-06	3.51E-06	1.77E-06	1.44E-06
Goodness of fit				
PEV	1.07E-3	1.02E-3	1.17E-3	1.08E-3
R^2	0.971	0.973	0.958	0.961
R^2_s	0.787	0.804	0.663	0.689
AIC	-6.377	-6.452	-6.307	-6.388

σ^2_ε variance of irregular component; σ^2_η variance of level; σ^2_ξ variance of slope; σ^2_ω variance of seasonal;

PEV Prediction Error Variance; R^2 Coefficient of determination; R^2_s Coefficient of differences around a seasonal mean;

AIC Akaike Information Criteria

variables the one that most strongly explained the behaviour of the demand. All the variables that were significant at 1% and 5% in the univariate version were also significant in the multivariate and only the intervention of October 2005 reduced its p-value in the multivariate version.

The intervention causing by the volcanic eruption in Iceland and the closure of European airspace (04:2010) was included in any models, producing a significant decrease in the entrance of passengers. The Japan earthquake was also included meaningfully in the United Kingdom models but producing an increase in the demand. The effect of terrorist attacks on competing destinations reflected in the 10:2005 intervention brings a significant increase in Germany passengers' entry, unlike the intervention of 04:2006.

The intervention periods corresponding to the stage after the 11:2012 and identified as terrorist attacks on competing destinations and the Spanish economic recession were not significant in any of the models of Germany or the United Kingdom. Unlike what happened in the associated period with the world economic recession from 06:2008 to 12:2010 which caused a significant decrease, aggravated in the United Kingdom, in the corresponding period from 05:2009 to 09:2009.

Table 6.5: *Estimated coefficients for the set of exogenous variables and interventions considered.*

Set of exogenous variables	Variable		Germany		United Kingdom	
			Univariate	Multivariate	Univariate	Multivariate
Calendar variables	Trading day	**Monday**	-0.00023	-0.00039	-0.00373	-0.00399
		Tuesday	-0.00057	-0.00182	0.00131	0.00032
		Wednesday	-0.00619	-0.00538	-0.00942	-0.00754
		Thursday	-0.01293**	-0.01407**	0.00420	0.00336
		Friday	-0.00006	0.00136	0.00634	0.00567
		Saturday	0.01678***	0.01650***	0.00043	0.00143
	Leap Year		0.0313	0.02501	0.05969***	0.06068***
	Carnivals		0.00031	0.00083	-0.00826**	-0.01006**
	Easter Week		0.00414**	0.00416**	0.00192	0.00214
Intervention variables	**10:2005**		0.06005*	0.07749**		
	04:2006		0.03129	0.03902		
	01:2009 to 12:2010		-0.04572**	-0.03933**		
	11:2012 to 10:2013		-0.04402	-0.04388		
	11:2013 to 12:2016		0.02036	0.01004		
	04:2010		-0.18367***	-0.18351***	-0.15561***	-0.15049***
	06:2008 to 12:2010				-0.06582**	-0.05246**
	05:2009 to 09:2009				-0.14581***	-0.13636***
	04:2011				0.11835***	0.11265***
	05:2013 to 12:2016				0.05688	0.03857
Capacity variables	**Slots**		0.32581***	0.30252***	0.11082***	0.12139***
Economic variables	**ESI**		0.00102	-0.00057	0.00091	0.00032
	CCI		-0.00067	0.00066	-0.00073	-0.00006
	RCPI		0.00673	0.01030	0.00157	0.00081

ESI Economic Sentiment Indicator; CCI Consumer Confidence Indicator; $RCPI$ Relative Consumer Price Index

*** Statistically significant at 1%; ** Statistically significant at 5%; * Statistically significant at 10%

Calendar variables had different behaviour depending on the country of origin. In Germany, Easter Week and the effect of have a larger number of Thursdays in a certain month (trading day Thursday is defined as the difference of Sundays versus Thursdays in a certain month and the variable would be negative if there is a higher number of Thursdays) were positively significant, while the effect of having a larger number of Saturdays was negative. In the United Kingdom, none of the trading days was significant, but Leap Year was positively significant, and Carnivals were negatively significant.

None of the economic variables considered ESI, CCI, and RCPI presented a significant association with tourism demand. All of them presented p-values higher than 0.4.

Table 6.6 shows the diagnostic tests for model estimation, including the critical values of the different tests considered. Normality, homoscedasticity, and independence are fulfilled in the univariate and multivariate models proposed for Germany and the United Kingdom, except independence in the univariate model for the United Kingdom. Autocorrelation function and histogram plus estimated density and theoretical normality curve for both countries of the univariate models are shown in Figure 6.8, while Figure 6.9 shows it for the multivariate models. Although Box-Ljung statistic (Q) and autocorrelation with 1 and 24 lags, r(1) and r(24) respectively, show forecast errors are serially uncorrelated, autocorrelation with 3 and 11 lags in Germany and with 6 and 11 in the United Kingdom show slight significant autocorrelations. The estimated density and theoretical normality curve are practically coincident in all the models, as the N statistic already ratified.

Table 6.6: *Diagnostic tests for model estimation.*

				UNIVARIATE		MULTIVARIATE	
		Statistic	Critical value	Value	Assumption satisfied	Value	Assumption satisfied
Germany	Independence	$Q(24)$	32.7 (33.9*)	31.405	+	24.304	+
		$r(1)$	±0.174	-0.031	+	-0.007	+
		$r(24)$	±0.174	-0.028	+	-0.016	+
	Homoscedasticity	$H(33)$	2.074	1.016	+	1.203	+
	Normality	N	5.99	0.335	+	0.856	+
United Kingdom	Independence	$Q(24)$	32.7	33.381	-	31.867	+
		$r(1)$	±0.174	0.072	+	0.062	+
		$r(24)$	±0.174	-0.0003	+	-0.028	+
	Homoscedasticity	$H(33)$	2.074	0.913	+	0.811	+
	Normality	N	5.99	0.879	+	0.813	+

* Critical value for multivariate model of Germany

The values of MAPE and Theil's U for the models considered in the different forecasting horizons are shown in Table 6.7. This information is also included for the SARIMA model. Graphically the considered models estimate successfully the real data, so that if they were represented on

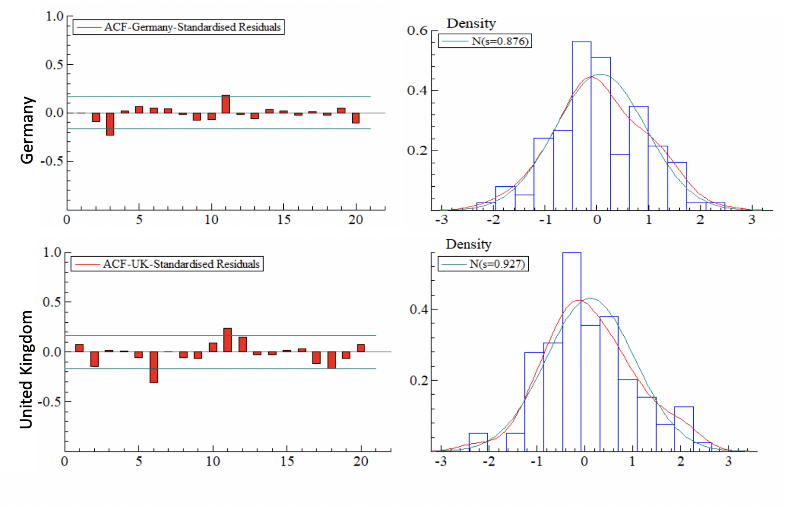

Figure 6.8: *Autocorrelation functions and histograms of standardized residuals of the univariate model of Germany and the United Kingdom.*

Figure 6.4, the real data would be hidden by the estimated series except for small peaks practically imperceptible.

Table 6.7: *Measures of fit for univariate and multivariate model in different forecasting horizons between TVP-STS and SARIMA models.*

		TVP-STS				SARIMA			
		Univariate		Multivariate		Univariate		Multivariate	
	Horizon	MAPE	THEIL'S U	MAPE	THEIL'S U	MAPE	THEIL'S U	MAPE	THEIL'S U
Germany	**1-month**	3.10	0.40	3.44	0.41	3.73	0.45	4.05	0.49
	3-month	3.47	0.42	3.76	0.47	4.04	0.54	4.44	0.52
	6-month	4.68	0.53	4.89	0.53	5.03	0.57	4.97	0.54
	12-month	5.02	0.58	6.01	0.69	5.52	0.60	5.73	0.66
United Kingdom	**1-month**	3.17	0.40	2.02	0.37	3.64	0.49	3.04	0.39
	3-month	3.95	0.45	2.91	0.41	3.83	0.51	3.39	0.47
	6-month	5.04	0.50	3.47	0.46	4.20	0.57	4.85	0.53
	12-month	5.10	0.51	3.79	0.48	4.89	0.61	5.32	0.63

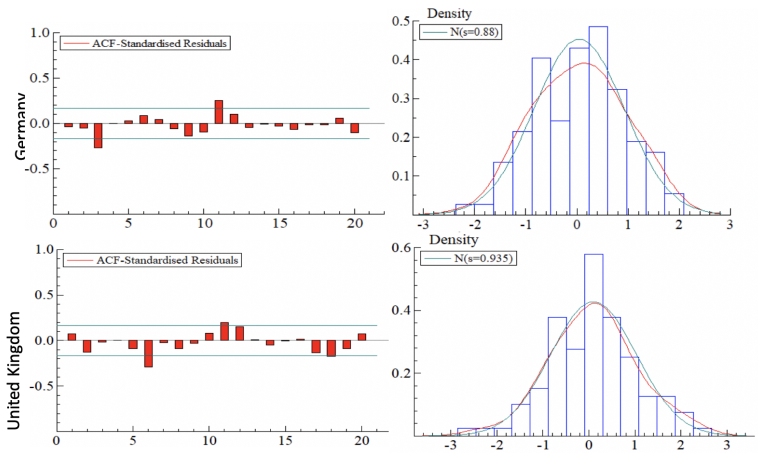

Figure 6.9: *Autocorrelation functions and histograms of standardized residuals of the multivariate model of Germany and the United Kingdom.*

Regarding the forecast period, the 2016 year, it becomes notorious that the multivariate model obtained a better prediction of the behaviour of the United Kingdom series than the univariate model, showing both MAPE as U-Theil values lower in all forecasting horizons. With regard to the SARIMA model, these values are somewhat worse, although for the last prediction horizons used, the MAPE values obtained with the univariate SARIMA improve those of the univariate TVP-STS model, although worse than the multivariate TVP-STS model. In view of the values obtained from MAPE and Theil's U for Germany in the univariate and multivariate model, with increase in both indices over those of the United Kingdom, it was selected as the best model to explain the behaviour of the prediction the univariate model, unlike the United Kingdom where multivariate model obtained better results. The univariate and multivariate SARIMA models for Germany were also somewhat worse than the univariate TVP-STS model. The results on TVP-STS models also coincided with the values of the M-DM test, obtaining for the United Kingdom a M-DM statistic of -3.5694 (p-value=0.0002, the multivariate model is more accurate than univariate one) and for Germany a M-DM statistic of 2.4767 (p-value=0.0064, the univariate model is more accurate than multivariate one).

For the selected models, the univariate model in the case of Germany and the multivariate model in the case of the United Kingdom, the unobserved components are shown in Figure 6.10. This figure includes the representation of the following components: level plus exogenous variables (Level + Reg), level, seasonal and irregular component. The level of both series suffered a decrease between 2008 and 2010, which coincides with the period of the world economic recession, being the case of the United Kingdom more noticeable. However, tourist arrivals have been increasing since then especially in the last years. Furthermore, clearly seasonal patterns can be observed in both series, being -0.6636 the seasonal correlation value. As of 2010, the seasonal pattern remains fairly stable in both countries, fulfilling that winter season (October to April) preferred choice by a German tourist, unlike British tourist that have a more irregular behaviour, whose seasonal peaks are recorded in the summer months, mainly.

Figure 6.11 shows the real and ex-post predicted (Song et al., 2011) values of the series during 2016 using the selected models built with the data of 2005:01 to 2015:12. For both countries, the twelve-month ahead forecasts underestimate the observed series mainly in the summer period and adjust more accurately in autumn and winter. These same characteristics are maintained when one ahead-month forecast is used, although with more accuracy in the summer.

6.5 Discussion

This chapter starts analysing the total tourism at the Canary Islands, using the most seasonal ARIMA model, the airline passenger model, and the structural model. Concerning the results obtained, the best model to explain the behaviour of the passenger arrivals to the Canary Islands was the structural model with the inclusion of the interventions detected and the calendar variables. Although this model fits almost perfectly to the data in the part of the estimates, in the forecasting is one of the models that worse fits. Which it appears that despite it correctly explaining past behaviour, to forecast is necessary to introduce other types of exogenous variables such as

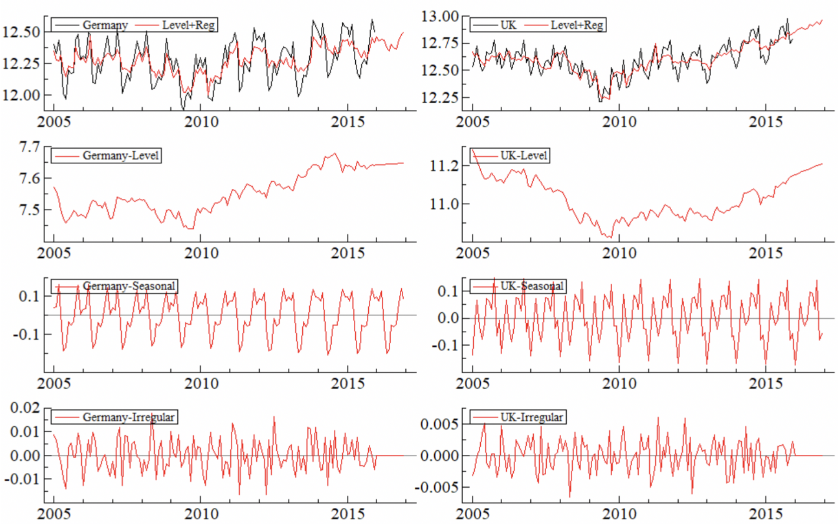

Figure 6.10: *Unobserved components of Germany and UK series in the univariate and multivariate model respectively.*

the capacity variables.

Furthermore, this study investigates the predictive power of structural time series model in both its univariate and multivariate form to predict tourism demand. In view of the empirical results, there is evidence that structural time series models perform consistently well in forecasting as showed by Li et al. (2005). There are several tourism forecasting studies that compare the accuracy of short-term forecast generated by multivariate models with the accuracy by univariate models (Song and Witt, 2000; Kulendran and Witt, 2001). Nevertheless, in tourism demand studies, the applications of multivariate time-series models are rarer. Pfefferman and Allon (1989) are the first to apply a multivariate time series model to tourism data, comparing the forecasting performance of the multivariate exponential smoothing model in state space form and the vector ARIMA models with their univariate counterparts and concluding that these multivariate procedures in practice are adequate. Kulendran and Witt

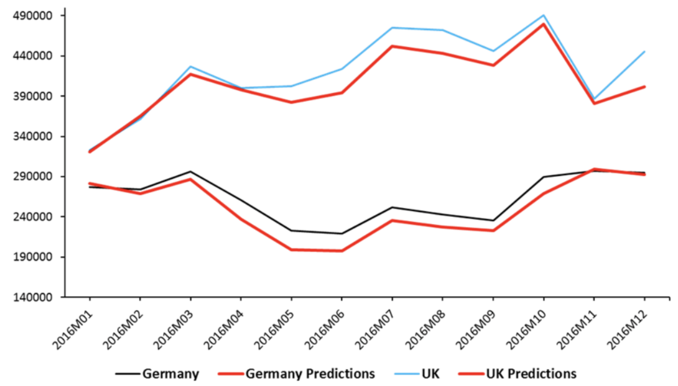

Figure 6.11: *Twelve-month ahead ex-post forecasts of the selected models.*

(2003) and Song et al. (2011) present empirical evidence of the Structural Time Series Model's superiority against the SARIMA model, due they capture the dynamic interrelationship between time series and may results in more accurate forecast (Athanasopoulos and de Silva, 2012). This result was also observed in this work.

It is found that based on the Akaike Information Criterion (AIC) there are not significant differences between the multivariate model and the univariate model. However, evaluating the forecast accuracy, the multivariate approach compared to the univariate one obtained lower forecast errors in the case of the United Kingdom and higher for Germany. Given that the degree of relationship between both countries is reciprocal, this result could be due to the strong influence that the number of British tourists would be exerting on German in the multivariate model. Silva et al. (2017) introduce an algorithm to identify leading indicators and they obtain that the German series is causing the British series significantly, but not the other way around. Despite this, it can be concluded that both models are useful in forecasting tourism demand. Especially the multivariate model is useful when the observed time series are strongly correlated and the series to be predicted is not extremely conditioned by the others.

Clearly, the model cannot be expected to foresee drastic exogenous forces, affecting the modeled level as it happened in 2016. For that reason, the model also allows the incorporation of interventions that can affect the modeled time series. Prideaux et al. (2003) argued that current forecasting methods have little ability to cope with unexpected crises and disasters. In the application presented in this paper, interventions have been constructed based on real events that are expected to have affected the tourism demand.

In this study, three were the external effects that most influenced in German and British tourism demand causing an increase or decrease in the number of passengers' entry. During the period of the global crisis, tourism was in the background not being an option for many families who experienced difficulties, causing a drastic decrease in passengers' entry. Furthermore, in March 2010 the eruption of the Iceland volcano occurred, which entailed the closure of the aerial space and therefore a deviation from tourism that originally intended to visit the islands and had to travel to other destinations, or simply not do it. This effect was reflected in April of that year. The third factor was the Japan earthquake, which affected British tourism, since it is one of the main emitters of tourists to Japan and redirecting the tourist flows to the Canary Islands, another of the main destination chosen by British. Additionally, it should be noted that in the period of the possibility of an economic rescue by the European Union in Spain (11:2012 to 10:2013) there was a reduction of German passengers' entry, but not significant.

In addition, the effect of terrorist attacks on competing destinations must be taken into account. This produces a feeling of insecurity in tourists, causing an emotional response of fear and anxiety at the idea of visiting those countries, choosing a destination in which they feel safer and risk-free (Baker, 2014). Attacks like the one that happened in September 2005 or those that have occurred in recent years had a significant effect on tourism, increasing the entry of passengers choosing the Canary Islands against competing destinations that have suffered attacks. The need for safety has become in one of the main factors when choosing a travel destination.

The analysis of the contribution of the explanatory variables to the growth of tourist arrivals has shown that economic variables have no effect on the demand for tourism, not having a statistically significant impact. In order to obtain a greater explanation of the variation and reduce the prediction errors, economic variables were included to improve the behaviour estimate of the time series. However, as it happens in our case, the inclusion of economic variables that a priori will help a better understanding of the behaviour does not always have to improve forecasting performance, as Kulendran and Witt (2003), Turner and Witt (2001) and Athanasopoulos et al. (2011) concluded. Nevertheless, in some case the inclusion of economic variables can help to explain the behaviour of the cyclical component (González and Moral, 1996; Guizzardi and Mazzocchi, 2010). This approach can be used and obtain good results provided that a sufficiently long study period is available.

Calendar variations can affect the behaviour of tourism demand, the fact that some vacations extend two consecutive months from year to year or that one year has one day more than another may have a positive or negative effect on the entry of passengers (EUROSTAT, 2017b). For example, the fact that the year is a leap year has a positive effect on the entry of British tourists due to the great volume of tourists on average that arrive at the islands per day from that country. Moreover, the entry of passengers may also be influenced by other variations associated with calendar composition such as trading day variations. There are days of the week that are more important than others. This effect can be seen reflected in the case of German tourists, for whom the trading day of Thursday and Saturday are significant. The largest number of flights from Germany recorded on Thursday, while the largest number of departures concentrated on Saturdays (AENA, 2017). However, the effect of the trading day does not affect British tourism since the frequency of flights is more regular throughout the week.

Among all the exogenous and intervention variables considered, the number of airline seat reservations for regular flights (slots) was the one that best explained the behaviour of the series studied. Prior knowledge of the slots is a strong explanatory variable highly correlated with the passenger input that contributes significantly to the modeling and predicting of this. Although this variable is available at different destination airports, it has not been used previously in tourism demand studies, and we believe should be taken into account. In spite of being given one year in advance, it is subject to variations throughout the year depending on the events that may occur. However, due to the increase in tourism worldwide, the reservations of runways at airports by airlines must increasingly be made with greater advance and precision, therefore, the reliability of this variable each day will be higher.

Some remarks should be made about the evolution of the distribution of the seasonal factors throughout the years since it is becoming more homogenous. Modeling seasonal variations has become an important issue in tourism forecasting in recent years (Kulendran and Wong, 2005). The stochastic specification of the seasonal component allowed us to perform a detailed analysis of the evolution of seasonal behaviour, being the assumption of stochastic seasonality more popular in recent studies (Lim and McAleer, 2001; Song et al., 2011). Since 2010, stability seasonal may be observed in both time series (Figure 6.10). Despite any preconception that can be had about the similar behaviour of German and British tourist arrivals, these present different seasonal patterns, being the winter season (October to April) preferred choice by a German tourist. Unlike British tourist that have a more irregular behaviour, whose seasonal peaks are recorded in March, July, August, and October.

The recent years of observations are characterised by a more pronounced increase in the number of tourist arrivals. The high entrance of tourists in 2016 was unexpected, causing a sudden change in the level of the time series. During the last years, the tendency had stabilized in both countries, in the German tourists quite flat and in the British with a constant increase since 2013. This fact is reflected in the results of the intervention variable associated with the terrorist attacks of the last period analysed, starting in 2013, which no longer presents statistical significance. These changes in 2016 seem to be related to the conduct of tour operators, which maintains the inertia of recent years. For this reason, in the present study, it was decided to use the 2016 data to test the forecast capacity of the models, so that the responsiveness capacity of the state space models could be analysed in the face of sudden and unexpected changes in the trend of tourism demand. However, stability in the future tourist arrivals number is expected for both countries. If on the contrary, the aim would have been to forecast the tourist arrivals in 2017, the MAPE values lower than 1% obtained in estimates, would also remain in forecasts since in this way the anomalous boom produced in 2016 would be explained in estimates.

6.6 Concluding remarks

This study has developed structural models, both univariate and multivariate versions, in which all the unknown parameters of the level, slope and seasonal are treated stochastically, except in case of Germany's univariate model where the slope was treated deterministically. The explanatory variables were initially introduced stochastically, although the results led to their treatment deterministically.

The empirical results show that TVP-STS and CSM models are capable of producing adequate forecasts. The structural equations forecasting component of the integrated model is useful for quantifying the separate effects of changes in key drivers of tourism demand. The advantages of structural time series model to formulate directly in terms of unobserved components treated stochastically, such as trend and seasonal, has turned out to be a powerful tool to proxy the influence of unobserved variables and evaluate changes in tourist preferences.

The forecast of the tourism demand at the Canary Islands obtained similar results using both approaches the Box-Jenkins methodology and structural models. However, structural models presented better results explaining the behaviour of the studied time series, fitting better to the original values of the time series in the estimates period.

Tourism demand from different origin countries with similar behaviour can improve the prediction if analysed jointly. The covariation of time series that follow similar time-based patterns is a source of information that can improve forecast accuracy in certain situations. The results indicated that the univariate model forecasts better than multivariate model when the observed time series are not strongly correlated and the number of passengers is lower compared to the other origins.

Although this study shows the advantages that the use of these models could bring, more deep research still needs to be carried out to study what variables can influence the behaviour of tourism demand. Despite the great contribution of the slots variable, it is data that can be modified according to external circumstances, such as crisis, terrorist attacks, etc.

Chapter 7
R scripts and packages

7.1 Introduction

This thesis has involved not only the acquisition of theoretical and practical knowledge about the analysis of time series and its application but also has led to the acquisition of knowledge about programming, in particular, the free software environment for statistical computing R.

During the thesis and as the study of structural models deepened the need arose for and use specific software that would allow carrying out the necessary analyses to achieve the stated objectives. However, as commented in Section 4.5 unlike the case of ARIMA models, nowadays there is not so much software focused on structural models.

There are some payment programs that apply structural models such as STAMP 8.30 software. We advocated for use free software, not only because much free software is heavily used all over the world and in this way they have more online support available and there are many access to a programs' source code to obtain and study, but also because the Instituto Canario de Estadística has an internal policy to bet on free software as much as possible.

Among the possible options offered by free software, finally R software was chosen, where two packages or libraries focused on the application of structural models and the Kalman filter can be found, the dlm package and KFAS package. Initially, we started using the dlm package which focused on Bayesian analysis of dynamic linear models, allowing develop a variety of constant or time-varying univariate and multivariate models. This package was employed to carry out the analysis exposed in the study of the stochastic and deterministic trend in Chapter 5.

Later, we saw that the use of the KFAS package was more intuitive and offered more information than that provided by the dlm package. The KFAS package has been created for state space modelling with the observations from an exponential family, not only for linear Gaussian state space models. The scripts developed for obtaining forecast tourist movements in frontiers of the Canary Islands for the Instituto Canario de Estadística are based on this package.

Finally, once put into practice and analysed both packages, they still did not fully comply with the needs that have been raised to us during these years of research. Although they are very complete packages, the user needs to have a deep knowledge of the package, of all its functions and how to combine them to be able to apply the desired model such as Local Linear Trend Model. So if we want to apply a change in the model, for example, including exogenous variables, the code had to be modified. Testing the goodness of fit of the different models became a cumbersome task. Therefore, a package called STSM was developed based on the KFAS package. It executes the structural models most used in practice, such as the Local Linear Trend Model or the Basic Structural Model, simply calling a function. It is posibble to indicate the desired peculiarities such as the stochastic or deterministic treatment of the components of the time series.

Since the development of the codes used to carry out the tests and obtaining results described in the previous chapters has been an important part of the work done in this thesis, then for each case they are presented below in Appendix B.

7.2 R Scripts to generate structural time series and study time-varying treatment of the trend in state space models

To achieve the objective of Chapter 5 to analyse the influence of the inclusion of the stochastic slope and level to describe the trend in structural time series using state space models, it was necessary to create simulations that allow checking if the estimated obtained by the Kalman filter could be able to discriminate between the slope and the level. To do that, R scripts to generate simulations, following a similar process described by Durbin and Koopman (2012, p. 26), were generated (Appendix B.1).

Three main scripts were developed with the purpose of

- Create functions for the Kalman Filter depending on the stochastic or deterministic treatment of the level and slope
- Generate time series simulations based on the specified q-ratios
- Generate contour graphics for sensitivity analysis.

7.3 R Scripts to forecast tourist movements in the border of the Canary Islands

As mentioned before, this thesis arose as a result of a project proposed by the Instituto Canario de Estadística, whose main objective was to forecast the number of tourist movements in frontiers of the Canary Islands (FRONTUR) according to nationality and destination island, using econometric models. The second objective was the creation of a system that allowed to generate forecasts automatically, through the development of a computer application.

With these objectives in mind and having studied the adequacy of structural models for tourism demand, R scripts were developed to make the 6-month forecasts of FRONTUR. Currently the Instituto Canario de Estadística makes and publishes the FRONTUR forecasts using the scripts set out in B.2, which can be found published in the GitHub public repository

($FRONTUR - forecast$, 2019).

7.4 Package STSM

As previously introduced, the STSM package was developed with the purpose of applying the different structural models without the needed of many changes in the codes, only calling a function specifying some characteristics of the studied time series, like the time-varying treatment of the components or the use or not of exogenous variables.

The models that can be found in the STSM package (B.3) are

- The Local Level Model (LLM)
- The Local Level Model with the inclusion of exogenous variables
- The Local Linear Trend Model (LLTM)
- The Local Linear Trend Model with the inclusion of exogenous variables
- The Basic Structural Model (BSM)
- The Basic Structural Model with the inclusion of exogenous variables
- The Cycle Plus Noise Model (CNM)
- The Cycle Plus Noise Model with the inclusion of exogenous variables
- The Trend Plus Cycle Model (TCM)

Chapter 8

Conclusions, limitations and future research

8.1 Conclusions

In summary, the following outcomes are concluded from this research

- Care must be taken with the use of some goodness-of-fit criteria ore accuracy measures when selecting the best structural model to forecast. Criteria such as AICc or BIC usually tend to suggest the use of models that include stochastic slope even in cases where it does not have a relevant presence. Unlike measures such as MAPE or Theil's U, which prioritize mainly models with deterministic slope.
- In the case of selecting a structural model with stochastic level and slope, if the values of their variances are high then an information transfer among them is produced and makes difficult their interpretation.
- Structural models present better results explaining the behaviour of the studied time series, thanks to its capacity to formulate the time series studied directly in terms of unobserved components.
- Tourism demand from different origin countries with similar behaviour can improve the prediction if analysed jointly.
- STSM package facilitates the use of state space models in the study of time series.

8.2 Limitations

During the research in this Ph.D. thesis, some limitations have been found. Following some of the limitations of the work presented in this manuscript are highlighted.

- As it is common in time series studies, if the exogenous variables are not able to detect changes in the time series, then the models probably will tend to keep the past patterns that might not fit that alteration, in case of changes in the trend or seasonal patterns.
- As seen in Chapter 6, the use of an exogenous variable highly correlated with the data such as the number of airline seat reservations for regular flights (slots) can be highly beneficial when making forecasts. Its variable character, due to the reservations of runways at airports by airlines must increasingly be made with greater advance, makes the reliability of this type of variable should be carefully questioned and analysed.
- Another limitation related to finding variables that explain the time series studied is the problem of temporality. Some of the variables that were initially thought of as possible candidates to explain tourism demand were finally discarded due to temporary problems, either because the data was not found on the same frequency or because the period of available data did not coincide.
- In this thesis, the cyclical component has not been included because the studied periods could not be extended caused by the non-availability of all the exogenous variables used. In the cases that it will be possible, it would be advisable to include the cycle component to analyse its repercussions in the model and detect new patterns in the time series.
- One of the limitations of time series analysis, in general, is the difficulty of correctly outliers identified. A time series can be affected by uncontrolled or unexpected events, producing observations which clearly differ from the behaviour patterns of the time series. Identifying these specific events and analysing their possible causes can be a

challenge that involves not only an analysis of the time series history and its behaviour and also an exhaustive search for events that occurred in the periods detected and that may be related to the object of study.

8.3 Future research

This thesis address different issues related to tourism and met the objective of giving the Instituto Canario de Estadística a tool that could use to forecast the entry of tourists at Canary Islands. However, as this thesis progressed, other issues have arisen that may be of interest to both the academic and tourism industry. Today we are still working on them and we would like to delve deeper into them and look for answers to certain questions and ideas that have come up and which are detailed below.

- Study on the feasibility of including in the model currently used in the Instituto Canario de Estadística for forecast a variable that contains information on hotel reservations, number of reserved accommodations, hotel expense forecast, etc., and if it can be done, analyse its influence.

- Analyse the influences of closure of the main countries' airlines of tourists' origins. Since in the last year several German airlines have been closed and it would be of interest to study whether or not this fact really influences the entry of tourists.

- Analyse the influence of the trend in competing destinations. Through a study of potential future scenarios, it would try to analyse how the increase or decrease in the tourist flow to direct competitor destinations, such as Turkey or Egypt, would influence in the Canary Islands.

- Perform a multivariate analysis of the tourist's entrance individually by islands with the restriction that the sum of the tourist entrance by islands is equal to the total tourists' entrance in the Canary Islands. Comparing the results of a post-adjustment approach and a parallel modelling adjustment (Benchmarking or Multivariate Cholette).

- Analyse the possible influences of the tourist flows from the point of view of the country of origin, which could explain in a multivariate way the tourist demand in one of the destination countries.
- Carry out a study to verify if the effect of passenger volume by countries in the multivariate models observed in this thesis is maintained.

Appendix

Appendix A
ARIMA Operators

The treatment of ARIMA models is greatly facilitated by the use of both difference and time lag operators. The characteristics of these operators are showed below.

A.1 Difference operator. Integration

If the difference operator ∇ is applied to a time series in a concrete time, z_t, it has the following effect

$$\nabla z_t = z_t - z_{t-1}$$

That is, the difference between the value referred to the current moment and the value taken by the time series in the immediately preceding period is obtained, and it is known as the *first difference*.

Then, applying ∇ to ∇z_t

$$\nabla[\nabla z_t] = \nabla[z_t - z_{t-1}] = \nabla z_t - \nabla z_{t-1} = [z_t - z_{t-1}] - [z_{t-1} - z_{t-2}]$$

This result is called the *second differences* and it is expressed as ∇^2. This can be generalized to ∇^d, for any positive integer d (the lag order).

Note that in each differentiation one period of time is lost.

In the application of operator ∇ differences are taken between each pair of consecutive values, either from the original time series or from those obtained in previous differences. Sometimes, however, it is interesting to take differences not between data corresponding to consecutive periods, but with a greater interval. For example, it may be appropriate to take differences with an interval of 12 periods when using monthly data. This kind of difference is usually connected with the problem of the seasonality of the data. Particularly, in the methodology of ARIMA models is often use the terminology: "*seasonal difference operator of period s*", where "s" is the interval with which differences are taken, 4 in quarterly data, 12 in monthly data, etc. The s appears in this case as a subscript of the symbol ∇, this is, ∇_s, to distinguish them from the differences between consecutive periods.

A.2 Time lag operator

The time lag operator B applied to a time series means that, given a time series z_t, this is lagged in 1 period. It is defined as

$$Bz_t = z_{t-1}$$

The meaning of $B^2 z_t$ is immediate, since

$$B^2 z_t = B[Bz_t] = B[z_{t-1}] = z_{t-2}$$

In general

$$B^k z_t = z_{t-k}$$

The B operator can be manipulated as if it were any algebraic amount. In this way,

$$B(az_t) = aB(z_t) = az_{t-1}$$

$$B^k B^d(z_t) = B^{k+d}(z_t) = z_{t-k-d}$$

The identity time lag operator is $B^0 = I = 1$. The application of the identity time lag operator does not alter the reference period

$$B^0 z_t = z_t$$

The relationship between difference operator and time lag operator is immediate, since $\nabla = 1 - B$.

Similarly, the relationship between the seasonal difference operator and the time lag operator can be established

$$\nabla_d = 1 - B^d$$

Operator B can be used to express a lagged model. For example, given the model

$$z - \phi_1 z_{t-1} - \phi_2 z_{t-2} - ... - \phi_p z_{t-p} = a_t$$

applying the operator B

$$[1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p] = a_t$$

The expression that appears in square brackets is called *lag-polynomial* and it is represented as $\phi(B)$. Then, the model can be formulated in a compact form as

$$\phi(B) z_t = a_t$$

Appendix B
Software

B.1 R Scripts to generate structural time series and study time-varying treatment of the trend in state space models

```
#############################################################
#               FUNCTIONS FOR THE KALMAN FILTER             #
#   DEPENDING ON THE STOCHASTIC OR DETERMINISTC TREATEMENT  #
#                   OF THE LEVEL AND SLOPE                  #
#############################################################

Determinista <- function(Z){
  ssKi = SSModel(Z ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                   matrix(0))),
                   H = matrix(NA))

  fitSSKI = fitSSM(ssKi, inits = c(0,0,0), method = "BFGS")
  var <- fitSSKI$optim.out$par

  # Compute the filtered estimates of the states
  # Smoothed results
  out <- KFS(fitSSKI$model, filtering = "state")
  # Estimates values
  dfSSKSIStates = KFS(fitSSKI$model, filtering = "state")$alphahat

  # Compute the log-likelihood
  logLik(fitSSKI$model)

  # Value of level and slope
  dfSSKSIStates[1,]

  # Value of the variance of the level and slope
  fitSSKI$model$Q
  D_L <- as.matrix(fitSSKI$model$Q)[1,]
  D_S <- as.matrix(fitSSKI$model$Q)[4,]

  # Value of the variance of the error
```

```
(H <- fitSSKI$model$H)
D_Error <- as.numeric(fitSSKI$model$H)

############################
# AIC, AICc and BIC values #
############################

nW <- 0
for (i in 1:length(fitSSKI$model$Q)){
  if (fitSSKI$model$Q[i] != 0){
    nW <- nW+1
  }
  else {
    nW <- nW
  }
}
nW

(npar <- length(H)+nW)
(2*npar)
Ld <- -logLik(fitSSKI$model)

(n <- length(Z))

(aic <- 2*Ld + 2*npar)
(aicc <- aic +(2*(npar^2+npar))/(n-npar-1))
(bic <- 2*Ld + log(n)*npar)

######## MAPE and Theils U values with the smoothed ########

temp <- dfSSKSIStates[,1]

# MAPE values for the estimates
MAPE1 <- mean(abs(temp-Z)/abs(Z))
(MAPE1*100)


# Theils U values for the estimates
(Ut1 <- sqrt(mean((((temp-Z))/Z)^2)/
              mean(((diff(Z))/Z[1:(n-1)])^2)))


###################### q-ratios ######################
qratioE <- D_Error/D_Error
qratioN <- D_L/D_Error
qratioP <- D_S/D_Error


###########################################################
# Saving of the results obtained
Table_Values <- cbind(D_Error,D_L,D_S, aic, aicc ,bic, MAPE1*100,
                      Ut1, qratioE, qratioN, qratioP,
                      dfSSKSIStates[1,1],dfSSKSIStates[1,2])
colnames(Table_Values) <- c("D_error","D_nivel","D_pend","AIC","AICc",
                            "BIC","MAPE","UTheil","q-ratioE",
                            "q-ratioN","q-ratioP","mu1","beta1")

```

```
  ##############################################################
  # Saving of time series, level and slope estimates
  Estimates <- cbind(temp,dfSSKSIStates[,1],dfSSKSIStates[,2])
  colnames(Estimates) <- c("Serie_est","mu_est","beta_est")

  mylist <- list(Table_Values,Estimates)
  names(mylist) <- c("Table_Values","Estimates")
  return(mylist)
}





Estocastico <- function(Z){
  ssKi = SSModel(Z ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                   matrix(NA))),
                   H = matrix(NA))

  fitSSKI = fitSSM(ssKi, inits = c(0,0,0), method = "BFGS")
  var<-fitSSKI$optim.out$par

  # Compute the filtered estimates of the states
  # Smoothed results
  out <- KFS(fitSSKI$model, filtering = "state")
  # Estimates values
  dfSSKSIStates = KFS(fitSSKI$model, filtering = "state")$alphahat

  # Compute the log-likelihood
  logLik(fitSSKI$model)

  # Value of level and slope
  dfSSKSIStates[1,]

  # Value of the variance of the level and slope
  fitSSKI$model$Q
  D_L <- as.matrix(fitSSKI$model$Q)[1,]
  D_S <- as.matrix(fitSSKI$model$Q)[4,]

  # Value of the variance of the error
  (H <- fitSSKI$model$H)
  D_Error <- as.numeric(fitSSKI$model$H)

  ###############################
  # Values of AIC, AICc and BIC #
  ###############################

  nW <- 0
  for (i in 1:length(fitSSKI$model$Q)){
    if (fitSSKI$model$Q[i] != 0){
      nW <- nW+1
    }
    else {
      nW <- nW
    }
  }
```

```r
  nW

  (npar <- length(H)+nW)
  (2*npar)
  Ld <- -logLik(fitSSKI$model)

  (n <- length(Z))

  (aic <- 2*Ld + 2*npar)
  (aicc <- aic +(2*(npar^2+npar))/(n-npar-1))
  (bic <- 2*Ld + log(n)*npar)

  ######## MAPE and Theils U values with the smoothed ########

  temp <- dfSSKSIStates[,1]

  #MAPE values for the estimates
  MAPE1 <- mean(abs(temp-Z)/abs(Z))
  (MAPE1*100)


  #Theils U values for the estimates
  (Ut1 <- sqrt(mean((((temp-Z))/Z)^2)/
               mean(((diff(Z))/Z[1:(n-1)])^2)))


  ###################### q-ratios ######################
  qratioE <- D_Error/D_Error
  qratioN <- D_L/D_Error
  qratioP <- D_S/D_Error


  ################################################################
  # Saving of the results obtained
  Table_Values <- cbind(D_Error,D_L,D_S, aic, aicc ,bic, MAPE1*100,
                        Ut1, qratioE, qratioN, qratioP,
                        dfSSKSIStates[1,1],dfSSKSIStates[1,2])
  colnames(Table_Values) <- c("D_error","D_nivel","D_pend","AIC","AICc",
                              "BIC","MAPE","UTheil","q-ratioE",
                              "q-ratioN","q-ratioP","mu1","beta1")

  ###########################################################
  # Saving of time series, level and slope estimates
  Estimates <- cbind(temp,dfSSKSIStates[,1],dfSSKSIStates[,2])
  colnames(Estimates) <- c("Serie_est","mu_est","beta_est")

  mylist <- list(Table_Values,Estimates)
  names(mylist) <- c("Table_Values","Estimates")
  return(mylist)
}



LsSd <- function(Z){
  ssKi = SSModel(Z ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                   matrix(0))),
                 H = matrix(NA))
```

```
fitSSKI = fitSSM(ssKi , inits = c(0,0,0), method = "BFGS")
var <- fitSSKI$optim.out$par

# Compute the filtered estimates of the states
# Smoothed results
out <- KFS(fitSSKI$model , filtering = "state")
# Estimates values
dfSSKSIStates = KFS(fitSSKI$model , filtering = "state")$alphahat

# Compute the log-likelihood
logLik(fitSSKI$model)

# Value of level and slope
dfSSKSIStates[1,]

# Value of the variance of the level and slope
fitSSKI$model$Q
D_L <- as.matrix(fitSSKI$model$Q)[1,]
D_S <- as.matrix(fitSSKI$model$Q)[4,]

# Value of the variance of the error
(H <- fitSSKI$model$H)
D_Error <- as.numeric(fitSSKI$model$H)

###############################
# Values of AIC, AICc and BIC #
###############################

nW <- 0
for (i in 1:length(fitSSKI$model$Q)){
  if (fitSSKI$model$Q[i] != 0){
    nW <- nW+1
  }
  else {
    nW <- nW
  }
}
nW

(npar <- length(H)+nW)
(2*npar)
Ld <- -logLik(fitSSKI$model)

(n <- length(Z))

(aic <- 2*Ld + 2*npar)
(aicc <- aic +(2*(npar^2+npar))/(n-npar-1))
(bic <- 2*Ld + log(n)*npar)

######## MAPE and Theils U values with the smoothed ########

temp <- dfSSKSIStates[,1]

#MAPE values for the estimates
MAPE1 <- mean(abs(temp-Z)/abs(Z))
(MAPE1*100)
```

```


  #Theils U values for the estimates
  (Ut1<- sqrt(mean((((temp-Z))/Z)^2)/
              mean(((diff(Z))/Z[1:(n-1)])^2)))


  ###################### q-ratios ######################
  qratioE <- D_Error/D_Error
  qratioN <- D_L/D_Error
  qratioP <- D_S/D_Error


  ################################################################
  #Saving of the results obtained
  Table_Values <- cbind(D_Error,D_L,D_S, aic, aicc ,bic, MAPE1*100,
                        Ut1, qratioE, qratioN, qratioP,
                        dfSSKSIStates[1,1],dfSSKSIStates[1,2])
  colnames(Table_Values) <- c("D_error","D_nivel","D_pend","AIC","AICc",
                              "BIC","MAPE","UTheil","q-ratioE",
                              "q-ratioN","q-ratioP","mu1","beta1")

  #############################################################
  # Saving of time series, level and slope estimates
  Estimates <- cbind(temp,dfSSKSIStates[,1],dfSSKSIStates[,2])
  colnames(Estimates) <- c("Serie_est","mu_est","beta_est")

  mylist <- list(Table_Values,Estimates)
  names(mylist) <- c("Table_Values","Estimates")
  return(mylist)
}



LdSs <- function(Z){
  ssKi = SSModel(Z ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                   matrix(NA))),
                   H = matrix(NA))

  fitSSKI = fitSSM(ssKi, inits = c(0,0,0), method = "BFGS")
  var <- fitSSKI$optim.out$par

  # Compute the filtered estimates of the states
  # Smoothed results
  out <- KFS(fitSSKI$model, filtering = "state")
  # Estimates values
  dfSSKSIStates = KFS(fitSSKI$model, filtering = "state")$alphahat

  # Compute the log-likelihood
  logLik(fitSSKI$model)

  # Value of level and slope
  dfSSKSIStates[1,]

  # Value of the variance of the level and slope
  fitSSKI$model$Q
  D_L <- as.matrix(fitSSKI$model$Q)[1,]
```

```
D_S <- as.matrix(fitSSKI$model$Q)[4,]

# Value of the variance of the error
(H <- fitSSKI$model$H)
D_Error <- as.numeric(fitSSKI$model$H)

###############################
# Values of AIC, AICc and BIC #
###############################

nW <- 0
for (i in 1:length(fitSSKI$model$Q)){
  if (fitSSKI$model$Q[i] != 0){
    nW <- nW+1
  }
  else {
    nW <- nW
  }
}
nW

(npar <- length(H)+nW)
(2*npar)
Ld <- -logLik(fitSSKI$model)

(n <- length(Z))

(aic <- 2*Ld + 2*npar)
(aicc <- aic +(2*(npar^2+npar))/(n-npar-1))
(bic <- 2*Ld + log(n)*npar)

######## MAPE and Theils U values with the smoothed ########

temp <- dfSSKSIStates[,1]

#MAPE values for the estimates
MAPE1<- mean(abs(temp-Z)/abs(Z))
(MAPE1*100)


#Theils U values for the estimates
(Ut1<- sqrt(mean((((temp-Z))/Z)^2)/
              mean(((diff(Z))/Z[1:(n-1)])^2)))


###################### q-ratios ######################
qratioE <- D_Error/D_Error
qratioN <- D_L/D_Error
qratioP <- D_S/D_Error


#################################################################
#Saving of the results obtained
Table_Values <- cbind(D_Error,D_L,D_S, aic, aicc ,bic, MAPE1*100,
                      Ut1, qratioE, qratioN, qratioP,
                      dfSSKSIStates[1,1],dfSSKSIStates[1,2])
colnames(Table_Values) <- c("D_error","D_nivel","D_pend","AIC","AICc",
```

```
                              "BIC","MAPE","UTheil","q-ratioE",
                              "q-ratioN","q-ratioP","mu1","beta1")

  ##############################################################
  # Saving of time series, level and slope estimates
  Estimates <- cbind(temp,dfSSKSIStates[,1],dfSSKSIStates[,2])
  colnames(Estimates) <- c("Serie_est","mu_est","beta_est")

  mylist <- list(Table_Values,Estimates)
  names(mylist) <- c("Table_Values","Estimates")
  return(mylist)
}
```

```
########################################################
#     SCRIPT THAT GENERATES TIME SERIES SIMULATIONS    #
#           BASED ON THE SPECIFIED Q-RATIOS            #
########################################################

# PACKAGES
library(KFAS)
library(MLmetrics)

# THE ERROR VARIANCE VALUES, THE Q-RATIOS AND
# THE SIZE OF THE TIME SERIES ARE INITIALIZED
Desviacion <- 5
qratioN0 <- 0.0
qratioP0 <- 0.0
increment <- 1.0
npuntos <- 21
n <- 144
set.seed(-257689769) # A seed is fixed

# An array is created that saves all generated simulated series
Series<-matrix(NA, ncol=npuntos*npuntos, nrow=n-1)
# A matrix is created that saves all levels of the simulated
# time series generated
mu<-matrix(NA, ncol=npuntos*npuntos, nrow=n-1)
# A matrix is created that saves all the slopes of the simulated
# time series generated
beta<-matrix(NA, ncol=npuntos*npuntos, nrow=n-1)

# LOOP GENERATING THE NPOINTS * NPOINTS TIME SERIES SIMULATED
# IN FUNCTION OF THE VARIATION OF THE Q-RATIO
# A counter is initialized to save the series generated in the Series matrix
count<-1

qratioN <- vector()
qratioP<- vector()
# The first loop iterates the level q-ratio
for (i in 1:npuntos){
  if (i == 1){
    # The first q-ratio is initialized to the first value
    qratioN[i] <- qratioN0
  }
  else{
    # In each iteration we add increment
    qratioN[i] <- qratioN[i-1]+increment
```

```
  }

  # The second loop iterates the slope q-ratio
  for (j in 1:npuntos){

    if (j == 1){
      # The first q-ratio is initialized to the first value
      qratioP[j] <- qratioP0
    }
    else{
      # In each iteration we add increment
      qratioP[j] <- qratioP[j-1]+increment
    }

    # The variances of the level and slope errors are calculated
    # based on the q-ratios as the root of the q-ratio * variancia_error^2
    D_N <- sqrt(qratioN[i]*Desviacion^2)
    D_P <- sqrt(qratioP[j]*Desviacion^2)

    model <- SSModel(matrix(NA,144,1) ~ SSMtrend(degree = 2,
                                               Q = list(D_N,D_P)),
                     H = Desviacion)
    states <- simulateSSM(model, "states")
    epsilon <- simulateSSM(model, "epsilon")

    # The series obtained is saved
    Series[,count]<-states[2:144,1,1]+epsilon[2:144]+3000
    mu[,count]<-states[2:144,1,1]
    beta[,count]<-states[2:144,2,1]
    # Itera the counter of the time series to calculate
    count<-count+1
  }
}

######################################################################
# Models depending on the deterministic or stochastic behaviour of   #
#                     the level and slope                            #
######################################################################

# A matrix is created that saves all the results of each of
# the simulations and the models applied
# For this, an auxiliary matrix is created that saves the results of
# each time series and adds them to the final results matrix
# Auxiliary Matrix
m_resul<- matrix(NA, ncol = 13, nrow = 4)
# Final results matrix
Resul <- matrix(NA, ncol = 13, nrow = npuntos*npuntos*4)
# Estimates matrix
Estimaciones <- matrix(NA,ncol = 1,nrow = nrow(Series))

for (i in 1:ncol(Series)){

  # Determinist Model Results
  m_resul[1,] <- Determinista(Series[,i])$Varianzas
  # Stochastic Level - Deterministic Slope Model Results
  m_resul[2,] <- LsSd(Series[,i])$Varianzas
  # Deterministic Level - Stochastic Slope Model Results
```

```
  m_resul[3,] <- LdSs(Series[,i])$Varianzas
  # Stochastic Model Results
  m_resul[4,] <- Estocastico(Series[,i])$Varianzas

  Est1 <- as.data.frame(Determinista(Series[,i])$Estimaciones)
  Est2 <- as.data.frame(LsSd(Series[,i])$Estimaciones)
  Est3 <- as.data.frame(LdSs(Series[,i])$Estimaciones)
  Est4 <- as.data.frame(Estocastico(Series[,i])$Estimaciones)
  Estimaciones<-cbind(Estimaciones,Est1,Est2,Est3,Est4)

  if (i == 1){
    Resul <- m_resul

  }

  else {
    Resul<-rbind(Resul,m_resul)
  }
}

MM <- matrix(NA, ncol = 4, nrow = npuntos*npuntos*4)
tparN <- vector()
tparP <- vector()
S<-c(rep(1,4))
for (i in 2:(npuntos*npuntos)){
  S<-rbind(c(S,rep(i,4)))
}
model<-c("LdSd","LsSd","LdSs","LsSs")
MM[,4]<-rbind(rep(model,npuntos*npuntos))

for (i in 1:npuntos)
{
  cqratioN <- c(rep(qratioN[i],4))

  for (j in 1:npuntos){
    tparN<- rbind(tparN,cbind(cqratioN))
    tparP <- rbind(tparP,cbind(rep(qratioP[j],4)))
  }
}
MM[,2]<-tparN
MM[,3]<-tparP
MM[,1]<-S

Resul <- cbind(MM,Resul)
colnames(Resul)<-c("Serie","q-ratioN","q-ratioP", "Modelo","D_error",
                   "D_nivel","D_pend", "AIC","AICc","BIC","MAPE",
                   "UTheil","q-ratioE","q-ratioN","q-ratioP","mu1","beta1")


# The Series matrix is the one that contains all the time series generated
# in the simulation
View(Series)
# The Resul matrix contains the results obtained for each
# time series and model
View(Resul)
# The mu matrix is the one that contains all the levels of
# the time series generated in the simulation
```

```
View(mu)
# The mu matrix is the one that contains all the slopes of
# the time series generated in the simulation
View(beta)
# The Estimates matrix contains the estimates of
# the time series, mu and beta
View(Estimaciones)
```

```
################################################
#   SCRIPT GENERATING THE CONTOUR GRAPHICS     #
#        FOR SENSITIVITY ANALYSIS              #
################################################

library(plot3D)

# Save the cases of each model
# Deterministic Model
Modelo1 <- subset(Resul,Resul[,4]=="LdSd")
# Stochastic Level - Deterministic Slope Model
Modelo2 <- subset(Resul,Resul[,4]=="LsSd")
# Deterministic Level - Stochastic Slope Model
Modelo3 <- subset(Resul,Resul[,4]=="LdSs")
# Stochastic Model
Modelo4 <- subset(Resul,Resul[,4]=="LsSs")

# Define the x and y axes (the q-ratios move from 0 to 20)
x <- seq(0,20,1)
y <- seq(0,20,1)

# Save the values of the estimated error deviation for each model
z1 <- matrix(as.numeric(Modelo1[,5]),21,21)
z2 <- matrix(as.numeric(Modelo2[,5]),21,21)
z3 <- matrix(as.numeric(Modelo3[,5]),21,21)
z4 <- matrix(as.numeric(Modelo4[,5]),21,21)

# Graphics
# Deterministic Model
contour(x,y,z1)
persp(x,y,z1)
image2D(z1,x,y)
persp3D(x,y,z1)
ribbon3D(x,y,z1)
hist3D(x,y,z1)

# Stochastic Level - Deterministic Slope Model
contour(x,y,z2)
persp(x,y,z2)
image2D(z2,x,y)
persp3D(x,y,z2)
ribbon3D(x,y,z2)
hist3D(x,y,z2)

# Deterministic Level - Stochastic Slope Model
contour(x,y,z3)
persp(x,y,z3)
image2D(z3,x,y)
persp3D(x,y,z3)
```

```
ribbon3D(x,y,z3)
hist3D(x,y,z3)

# Stochastic Model
contour(x,y,z4)
persp(x,y,z4)
image2D(z4,x,y)
persp3D(x,y,z4)
ribbon3D(x,y,z4)
hist3D(x,y,z4)
```

B.2 R scripts to forecast tourist movements in the border of the Canary Islands

```
# Data from JSON
#
# Author: Elisa M. Jorge Gonzalez <elisajg0@gmail.com>

#' @title Get data from JSON
#' @description Get data from ISTAC website of tourist movements in
#'              frontiers (FRONTUR) by islands
#' @param X Url
#' @param Y Type of data
#          "TOTAL"   -> tourist movements in frontiers in total
#          "ISLANDS" -> tourist movements in frontiers by islands
#' @return DATA Data obtained from url
#' @example
#  json_data(myurl, "TOTAL")
#  jdon_data(myurl, "ISLANDS")


json_data <- function(X,Y){

  # DATA
  # Read data from url
  Data <- fromJSON(url(X))
  # Get value, territory and date data
  Data <- Data$data
  # Get dimension codes
  t <- t(as.data.frame(Data$dimCodes))
  # Paste data and dimension codes in a data.frame
  Data <- cbind(Data,t)

  # SELECT DATA BY TERRITORY
  if (Y == "TOTAL")
  {# Select the total of tourist
    Data <- subset(Data, Data[,3] == "T")
    # Sort in ascending order
    Data <- arrange(Data, Data[,4])
    colnames(Data) <- c("Value","Code","Territory","Date")
    # Return data obtained from json
    return(Data)}
```

```
  if (Y == "ISLANDS")
  {# Select the total of tourist
    Data <- subset(Data, Data[,3] == "T" & Data[,4] == "T")
    # Sort in ascending order
    Data <- arrange(Data, Data[,5])
    # Get and save Lanzarote data
    ACE <- subset(Data, Data[,6] == "ES708")
    # Get and save Fuerteventura data
    FUE <- subset(Data, Data[,6] == "ES704")
    # Get and save Gran Canaria data
    LPA <- subset(Data, Data[,6] == "ES705")
    # Get and save Tenerife data
    TF  <- subset(Data, Data[,6] == "ES709")
    # Save data as data.frame
    DATA <- cbind(as.numeric_version(ACE[,1]), as.numeric_version(FUE[,1]),
                  as.numeric_version(LPA[,1]), as.numeric_version(TF[,1]))
    colnames(DATA) <- c("ACE", "FUE", "LPA", "TF")
    # Return data obtained from json
    return(DATA)}

}
```

```
# State Space Model forecast
#
# Author: Elisa M. Jorge Gonzalez <elisajg0@gmail.com>

#' @title Get forecast
#' @description Executes the Basic Structural Model (BSM) with the inclusion
#'              of exogenous variables and seasonal
#' @param X Times series or set of time series to be forecast
#' @param Y Exogenous variables
#' @param n Number of available data
#' @return mylist List of forecast of a time series or set of time series
#'         and the value of MAPE for the estimations
#' @example
#   get_forecast(myseries,myexogenous_va,n=24)

get_forecast <- function(X,Y,n){

  # CREATION OF OUTPUT VARIABLES
  # Variable for MAPE values
  MAPE <- matrix(NA, ncol = ncol(X), nrow = 1)
  # Variable for forecast values
  Prediction <- matrix(0, ncol = ncol(X), nrow = nrow(X))

  # Loop for executes BSM with the inclusion of exogenous variables
  # and seasonal

  if (ncol(X) > 1)
  {for (i in 1:ncol(X))
  {
    ssKi <- SSModel(log(as.numeric(X[,i])) ~ SSMtrend(degree = 2,
                        Q = list(matrix(NA),matrix(NA)))+
                        SSMseasonal(12, Q = matrix(NA))+
                        SSMregression(~Y, Q =matrix(0)),
                      H = matrix(NA))
    fitSSKI <- fitSSM(ssKi, inits = c(0,0,0,0), method = "BFGS")
```

```
    out <- KFS(fitSSKI$model, filtering = "state")
    dfSSKSIStates <- ts(out$alphahat, start = time(X)[1], frequency = 12)

    for (j in 1:ncol(Y))
    {
      Prediction[,i] <- Prediction[,i] + dfSSKSIStates[,j]*Y[,j]
    }
    Prediction[,i] <- Prediction[,i] + dfSSKSIStates[, ncol(Y) + 1] +
                      dfSSKSIStates[, ncol(Y) + 3]
    Prediction[,i] <- exp(Prediction[,i])

    MAPE[,i] <- (mean(abs((Prediction[1:n,i]) -
                            as.numeric(X[1:n,i]))/as.numeric(X[1:n,i])))*100
  }
  }
  else
  {
    ssKi <- SSModel(log(X) ~ SSMtrend(degree = 2,
                      Q = list(matrix(NA),matrix(NA)))+
                      SSMseasonal(12, Q = matrix(NA))+
                      SSMregression(~Y, Q =matrix(0)),
                    H = matrix(NA))
    fitSSKI <- fitSSM(ssKi, inits = c(0,0,0,0), method = "BFGS")
    out <- KFS(fitSSKI$model, filtering = "state")
    dfSSKSIStates <- ts(out$alphahat, start = time(X)[1], frequency = 12)

    for (j in 1:ncol(Y))
    {
      Prediction <- Prediction + dfSSKSIStates[,j]*Y[,j]
    }
    Prediction <- Prediction + dfSSKSIStates[, ncol(Y) + 1] +
                  dfSSKSIStates[, ncol(Y) + 3]
    Prediction <- exp(Prediction)

    MAPE <- (mean(abs(Prediction[1:n] -
                        as.numeric(X[1:n]))/as.numeric(X[1:n])))*100
  }



  mylist <- list(MAPE,Prediction)
  names(mylist) <- c("MAPE","Prediction")

  return(mylist)

}
```

```

# Title: Example of forecast using State Space Models
# Description: This script obtain forecast employing State Space Models
#              using data from xls or xlsx
# Version: 1.0


######################
# LOADING R PACKAGES #
######################
```

```r
# To read xls and xlsx files
library(readxl)
# To export xls and xlsx files
library(xlsx)
# To execute State Space Models
library(KFAS)


######################
# LOADING R FUNCTIONS #
######################
# Function to get forecast
source("R/forecast.R", encoding = 'UTF-8')


####################################
# READ DATA FROM XLS OR XLSX FILES #
####################################
# Data to forecast
Data     <- read.xlsx('data/Data.xls', sheetIndex = 1)
# Exogenous variables
Calendar <- read.xlsx("data/Calendar.xls", sheetIndex = 1)

# Number of available data
m <- nrow(Data)
# Number of available date
d <- nrow(Calendar)


#########################
# TIME SERIES STRUCTURE #
#########################
# Data to forecast
Data_ts <- as.data.frame(ts(Data[,3:ncol(Data)], start = c(Data[1,1],
                                                           Data[1,2]),
                            frequency = 12))
Data_ts <- ts(rbind(as.matrix(Data[,3:ncol(Data)]),matrix(NA,
                    ncol = ncol(Data) - 2, nrow = 6)),
                    start = c(Data[1,1],Data[1,2]),
                    frequency = 12)
# Exogenous variables
Calendar_ts <- ts(Calendar[,2:ncol(Calendar)], start = c(Data[1,1],
                                                         Data[1,2]),
                  end = time(Data_ts)[nrow(Data_ts)],
                  frequency = 12)


################
# GET FORECAST #
################
Z <- get_forecast(Data_ts,Calendar_ts,m)


#####################
# EXTRACT THE VALUES #
#####################
# Forecast values
```

```
Prediction <- round(Z$Prediction,0)
colnames(Prediction) <- colnames(Data)[3:ncol(Data)]
# MAPE values
MAPE <- Z$MAPE


##########
# OUTPUT #
##########
# Input data
write.xlsx(Data, file = "output/Forecast.xls", sheetName = "Data",
           append = FALSE)
# Forecast data
Date_f   <- as.character(Calendar[(nrow(Data) + 1):nrow(Data_ts),1])
Forecast <- cbind(Date_f,Prediction[(nrow(Data) + 1):nrow(Data_ts),])
write.xlsx(Forecast, file = "output/Forecast.xls", sheetName = "Forecast",
           append = TRUE)
# MAPE values
if (ncol(Prediction) == 1)
{names(MAPE) <- colnames(Data)[3:ncol(Data)]}
if (ncol(Prediction) > 1)
{colnames(MAPE) <- colnames(Data)[3:ncol(Data)]}
write.xlsx(MAPE, file = "output/Forecast.xls", sheetName = "MAPE",
           append = TRUE)
```

```

# Title: Forecast tourist movements in frontiers of Canary Islands (FRONTUR)
# Description: This script obtain forecast tourist movements in frontiers of
#              Canary Islands
#              employing State Space Models
# Version: 1.0


######################
# LOADING R PACKAGES #
######################
# To read xls and xlsx files
library(readxl)
# To export xls and xlsx files
library(xlsx)
# To execute State Space Models
library(KFAS)
# To read json
library(jsonlite)
# To database treatment
library(dplyr)
# To data treatment
library(varhandle)


#######################
# LOADING R FUNCTIONS #
#######################
# Function to get data from JSON
source("R/json_data.R", encoding = 'UTF-8')
# Function to get forecast
source("R/forecast.R", encoding = 'UTF-8')
```

```


#######################
# READ DATA FROM JSON #
#######################
# Url of Total tourist movements in frontiers at Canary Islands
CAN_json      <- "http://www.gobiernodecanarias.org/istac/jaxi-istac/
                  tabla.do?accion=jsonMtd&uuidConsulta=d17050d0-1e73-
                  450c-8648-62390c625b17"
# Url of tourist movements in frontiers by islands
ISLANDS_json <- "http://www.gobiernodecanarias.org/istac/jaxi-istac/
                  tabla.do?accion=jsonMtd&uuidConsulta=5fd35c5b-26dc-
                  4aaa-975d-9254c6dee11e"

# Get data from json
# Total tourist movements in frontiers at Canary Islands data
CAN <- json_data(CAN_json , "TOTAL")
# Tourist movements in frontiers by islands data
ISLANDS <- json_data(ISLANDS_json, "ISLANDS")

# Movements in frontiers at Canary Islands data
DATA <- cbind(as.character(CAN$Date),as.numeric_version(CAN$Value),ISLANDS)
colnames(DATA) <- c("DATE","CAN",colnames(ISLANDS))

# Number of available data
m <- nrow(DATA)


####################################
# READ DATA FROM XLS OR XLSX FILES #
####################################
# Exogenous variables
Calendar <- read.xlsx("data/Calendar.xls", sheetIndex = 1)
# Number of available date
d <- nrow(Calendar)


#########################
# TIME SERIES STRUCTURE #
#########################
# Data to forecast
Data_ts <- as.data.frame(ts(subset(DATA, select = -DATE), start = c(2010,01),
                            frequency = 12))
Data_ts <- ts(rbind(subset(DATA, select = -DATE), matrix(NA,
              ncol = ncol(subset(DATA, select = -DATE)), nrow = 6)),
              start = c(2010,01),
              frequency = 12)
# Exogenous variables
Calendar_ts <- ts(Calendar[,2:ncol(Calendar)], start = c(2010,01),
                  end = time(Data_ts)[nrow(Data_ts)],
                  frequency = 12)


################
# GET FORECAST #
################
Z <- get_forecast(Data_ts,Calendar_ts,m)
```

```


######################
# EXTRACT THE VALUES #
######################
# Forecast values
Prediction <- round(Z$Prediction,0)
colnames(Prediction) <- colnames(DATA)[2:ncol(DATA)]
# MAPE values
MAPE <- Z$MAPE


##########
# OUTPUT #
##########
# Input data
write.xlsx(DATA, file = "output/Forecast.xls", sheetName = "Data",
           append = FALSE)
# Forecast data
Date_f   <- as.character(Calendar[(nrow(DATA) + 1):nrow(Data_ts),1])
Forecast <- cbind(Date_f,Prediction[(nrow(DATA) + 1):nrow(Data_ts),])
write.xlsx(Forecast, file = "output/Forecast.xls", sheetName = "Forecast",
           append = TRUE)
# MAPE values
colnames(MAPE) <- colnames(DATA)[2:ncol(DATA)]
write.xlsx(MAPE, file = "output/Forecast.xls", sheetName = "MAPE",
           append = TRUE)
```

B.3 Package STSM

```
#' @title The Local Level Model (LLM)
#' @description Function LLM applies the Local Level Model.
#' @details The local level model allowing that the level of the series vary
#'          in time dependign
#'          of the level of the previous time plus
#'          a random element.
#' @param Serie Time series object containing the data.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param stochastic Logical. If TRUE the level will be treated as
#'                   stochastic.
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' LLM(data, anno = 1949, period = 1, freq = 12, num_f = 6,
#'     stochastic = TRUE)


```

```
LLM <- function(Serie, anno, period, freq, num_f = 0, stochastic = TRUE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }
  else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}

  if (stochastic == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")}
  else
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")}

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  estimates <- dfSSKSIStates[,1]

  return(estimates)
}





#' @title The Local Level Model with the inclusion of exogenous variables
#' @description Function LLM_X applies the Local Level Model with exogenous
#'              variables.
#' @details Function include exogenous variables in the local level model
#'          to improve the predictions.
#' @param Serie Time series object containing the data.
#' @param Exog  Time series object or set of time series containing the
#'              exogenous variables.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param L_stoch Logical. If TRUE the level will be treated as
#'                stochastic (TRUE default).
#' @param X_stoch Logical. If TRUE the exogenous variable will be treated
#'                as stochastic (FALSE default).
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
```

```r
#' data <- AirPassengers
#' vexog <- nottem[1:(length(AirPassengers)+6)]
#' forecast <- LLM_X(Serie = data, Exog = vexog, anno = 1949, period = 1,
#'                   freq = 12, num_f = 6,
#'                   L_stoch =TRUE, X_stoch = TRUE)


LLM_X <- function(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
                  X_stoch = FALSE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}
  # Exogenous variables
  Exog_ts <- ts(as.data.frame(Exog), start = c(anno, period),
                frequency = freq)

  if (L_stoch == TRUE & X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(NA))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  }else if (L_stoch == TRUE & X_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(NA))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }else if (L_stoch == FALSE & X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }else if (L_stoch == FALSE & X_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")}

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  # CREATION OF OUTPUT VARIABLE
  # Variable for forecast values
  estimates <- matrix(0, ncol = ncol(as.data.frame(Serie_ts)),
                      nrow = nrow(as.data.frame(Serie_ts)))
  for (j in 1:ncol(Exog_ts))
  {estimates <- estimates + dfSSKSIStates[,j]*Exog_ts[,j]}
  estimates <- estimates + dfSSKSIStates[, ncol(Exog_ts) + 1]

```

```
  return(estimates)
}
```

```
#' @title The Local Linear Trend Model (LLTM)
#' @description Function LLTM applies the Local Linear Trend Model.
#' @details The local linear trend model is the local level model adding
#'          a slope term, which is generated by
#'          a random wal.
#' @param Serie Time series object containing the data.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param L_stoch Logical. If TRUE the level will be treated
#'                as stochastic (TRUE default).
#' @param S_stoch Logical. If TRUE the slope will be treated
#'                as stochastic (TRUE default).
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' LLTM(data, anno = 1949, period = 1, freq = 12, num_f = 6)


LLTM <- function(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
                 S_stoch = TRUE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}

  if (L_stoch == TRUE & S_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                            matrix(NA))),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                            matrix(0))),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                            matrix(NA))),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
```

```
                                                              matrix(0))),
                     H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")
  }

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                        frequency = frequency(Serie_ts))

  estimates <- dfSSKSIStates[,1]

  return(estimates)
}





#' @title The Local Linear Trend Model with the inclusion of
#'        exogenous variables
#' @description Function LLTM_X applies the Local Linear Trend Model
#'              with exogenous variables.
#' @details Function include exogenous variables in the local linear
#'          trend model to improve the predictions.
#' @param Serie Time series object containing the data.
#' @param Exog  Time series object or set of time series containing the
#'              exogenous variables.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param L_stoch Logical. If TRUE the level will be treated
#'                as stochastic (TRUE default).
#' @param S_stoch Logical. If TRUE the slope will be treated
#'                as stochastic (TRUE default).
#' @param X_stoch Logical. If TRUE the exogenous variable will be treated
#'                as stochastic (FALSE default).
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' vexog <- nottem[1:(length(AirPassengers)+6)]
#' forecast <- LLTM_X(Serie = data, Exog = vexog, anno = 1949, period = 1,
#'                     freq = 12, num_f = 6,
#'                     L_stoch = TRUE, S_stoch = FALSE, X_stoch = FALSE)


LLTM_X <- function(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
                   S_stoch = TRUE, X_stoch = FALSE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
```

```
   Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                        matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                  start = c(anno, period), frequency = freq)
 }else
 {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}
 # Exogenous variables
 Exog_ts <- ts(as.data.frame(Exog), start = c(anno, period),
               frequency = freq)

 if (L_stoch == TRUE & S_stoch == TRUE & X_stoch == TRUE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                           matrix(NA)))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
 }else if (L_stoch == TRUE & S_stoch == TRUE & X_stoch == FALSE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                           matrix(NA)))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
 }else if (L_stoch == TRUE & S_stoch == FALSE & X_stoch == FALSE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                           matrix(0)))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
 }else if (L_stoch == FALSE & S_stoch == FALSE & X_stoch == FALSE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                           matrix(0)))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")
 }else if (L_stoch == FALSE & S_stoch == TRUE & X_stoch == FALSE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                           matrix(NA)))+
                            SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
 }else if (L_stoch == FALSE & S_stoch == FALSE & X_stoch == TRUE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                           matrix(0)))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
 }else if (L_stoch == FALSE & S_stoch == TRUE & X_stoch == TRUE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                           matrix(NA)))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
 }else if (L_stoch == TRUE & S_stoch == FALSE & X_stoch == TRUE)
 {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                           matrix(0)))+
                            SSMregression(~Exog_ts, Q = matrix(NA)),
                  H = matrix(NA))
 fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
```

```
  }

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  # CREATION OF OUTPUT VARIABLE
  # Variable for forecast values
  estimates <- matrix(0, ncol = ncol(as.data.frame(Serie_ts)),
                      nrow = nrow(as.data.frame(Serie_ts)))
  for (j in 1:ncol(Exog_ts))
  {estimates <- estimates + dfSSKSIStates[,j]*Exog_ts[,j]}
  estimates <- estimates + dfSSKSIStates[, ncol(Exog_ts) + 1]

  return(estimates)
}
```

```
#' @title The Basic Structural Model (BSM)
#' @description Function BSM applies the Basic Structural Model.
#' @details The Basic Structural Model is the local linear trend model with
#'          a seasonal term.
#' @param Serie Time series object containing the data.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param L_stoch Logical. If TRUE the level will be treated
#'                as stochastic (TRUE default).
#' @param S_stoch Logical. If TRUE the slope will be treated
#'                as stochastic (TRUE default).
#' @param Sea_stoch Logical. If TRUE the seasonal will be treated
#'                  as stochastic (TRUE default).
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' BSM(AirPassengers, anno = 1949, period = 1, freq = 12, num_f = 6)


BSM <- function(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
                S_stoch = TRUE, Sea_stoch = TRUE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}

  if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
```

```
                                                         matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                         matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                         matrix(0)))+
                              SSMseasonal(freq, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                         matrix(0)))+
                              SSMseasonal(freq, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                         matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  }else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                         matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }else if (L_stoch == TRUE & S_stoch == FALSE & Sea_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                         matrix(0)))+
                              SSMseasonal(freq, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  }else if (L_stoch == TRUE & S_stoch == FALSE  & Sea_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                         matrix(0)))+
                              SSMseasonal(freq, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  estimates <- dfSSKSIStates[,1] + dfSSKSIStates[,3]

  return(estimates)
}
```

```
#' @title The Basic Structural Model with the inclusion of time-varying
#'        exogenous variables (TVP)
#' @description Function BSM_X applies the Basic Structural Mosel with
#'              the inclusion of exogenous variables.
#' @details Function include time-varying exogenous variables in the
#'          basic structural model to improve the predictions.
#' @param Serie Time series object containing the data.
#' @param Exog  Time series object or set of time series containing
#'              the exogenous variables.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param L_stoch Logical. If TRUE the level will be treated
#'                as stochastic (TRUE default).
#' @param S_stoch Logical. If TRUE the slope will be treated
#'                as stochastic (TRUE default).
#' @param Sea_stoch Logical. If TRUE the seasonal will be treated
#'                  as stochastic (TRUE default).
#' @param X_stoch Logical. If TRUE the exogenous variable will be treated
#'                as stochastic (FALSE default).
#' @return Time series containing estimates results.
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' vexog <- nottem[1:(length(AirPassengers)+6)]
#' forecasts <- BSM_X(Serie = data, Exog = vexog, anno = 1949, period = 1,
#'                    freq = 12, num_f = 6,
#'                    L_stoch = TRUE, S_stoch = TRUE, Sea_stoch = TRUE,
#'                    X_stoch = TRUE)


BSM_X <- function(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
                  S_stoch = TRUE, Sea_stoch = TRUE, X_stoch = FALSE) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}
  # Exogenous variables
  Exog_ts <- ts(as.data.frame(Exog), start = c(anno, period),
                frequency = freq)

  if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == TRUE &
      X_stoch == FALSE)
```

```
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                      matrix(NA)))+
                             SSMseasonal(freq, Q = matrix(NA))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
} else if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == FALSE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                      matrix(NA)))+
                             SSMseasonal(freq, Q = matrix(0))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
} else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == FALSE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                      matrix(0)))+
                             SSMseasonal(freq, Q = matrix(0))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0), method = "BFGS")
} else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == TRUE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                      matrix(0)))+
                             SSMseasonal(freq, Q = matrix(NA))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
} else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == TRUE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                      matrix(NA)))+
                             SSMseasonal(freq, Q = matrix(NA))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
} else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == FALSE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                      matrix(NA)))+
                             SSMseasonal(freq, Q = matrix(0))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
} else if (L_stoch == TRUE & S_stoch == FALSE & Sea_stoch == TRUE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                      matrix(0)))+
                             SSMseasonal(freq, Q = matrix(NA))+
                             SSMregression(~Exog_ts, Q = matrix(0)),
                  H = matrix(NA))
fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
} else if (L_stoch == TRUE & S_stoch == FALSE  & Sea_stoch == FALSE &
            X_stoch == FALSE)
{sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
```

```
                                                        matrix(0)))+
                              SSMseasonal(freq, Q = matrix(0))+
                              SSMregression(~Exog_ts, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == TRUE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                        matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(NA))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == TRUE & Sea_stoch == FALSE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                        matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(0))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == FALSE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                        matrix(0)))+
                              SSMseasonal(freq, Q = matrix(0))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE & Sea_stoch == TRUE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                        matrix(0)))+
                              SSMseasonal(freq, Q = matrix(NA))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == TRUE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                        matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(NA))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == TRUE & Sea_stoch == FALSE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                        matrix(NA)))+
                              SSMseasonal(freq, Q = matrix(0))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == FALSE & Sea_stoch == TRUE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                        matrix(0)))+
```

```
                              SSMseasonal(freq, Q = matrix(NA))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == FALSE  & Sea_stoch == FALSE &
             X_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                              matrix(0)))+
                              SSMseasonal(freq, Q = matrix(0))+
                              SSMregression(~Exog_ts, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  }

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  # CREATION OF OUTPUT VARIABLE
  # Variable for forecast values
  estimates <- matrix(0, ncol = ncol(as.data.frame(Serie_ts)),
                      nrow = nrow(as.data.frame(Serie_ts)))
  for (j in 1:ncol(Exog_ts))
  {estimates <- estimates + dfSSKSIStates[,j]*Exog_ts[,j]}
  estimates <- estimates + dfSSKSIStates[, ncol(Exog_ts) + 1] +
               dfSSKSIStates[, ncol(Exog_ts) + 3]

  return(estimates)
}
```

```
#' @title The Cycle Plus Noise Model (CNM)
#' @description Function CNM applies the Cycle Plus Noise Model.
#' @details The Cycle Plus Noise Model assumed that the trend be constant
#'          and include the cycle component.
#' @param Serie Time series object containing the data.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param c_period A integer, which specify the length of the cycle.
#' @return Time series containing estimates results
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' CNM(data, anno = 1949, period = 1, freq = 12, num_f = 6, c_period = 12)



CNM <- function(Serie, anno, period, freq, num_f = 0, c_period) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
```

```
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }
  else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}

  sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0)) +
                             SSMcycle(c_period, Q = matrix(NA)),
                  H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")

  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  estimates <- dfSSKSIStates[,1] + dfSSKSIStates[,2]

  return(estimates)

}




#' @title The Cycle Plus Noise Model with the inclusion of exogenous
#'        variables
#' @description Function CNM_X applies the Cycle Plus Noise Model with
#'              the inclusion of exogenous variables.
#' @details Function include exogenous variables in the cycle plus noise
#'          model to improve the predictions.
#' @param Serie Time series object containing the data.
#' @param Exog  Time series object or set of time series containing the
#'              exogenous variables.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
#' @param c_period A integer, which specify the length of the cycle.
#' @param X_stoch Logical. If TRUE the exogenous variable will be treated
#'                as stochastic (FALSE default).
#' @return Time series containing estimates results
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' vexog <- nottem[1:(length(AirPassengers)+6)]
#' forecasts <- CNM_X(data, vexog, anno = 1949, period = 1, freq = 12,
#'                    num_f = 6,
#'                    c_period = 12, X_stoch = FALSE)


CNM_X <- function(Serie, Exog, anno, period, freq, num_f = 0, c_period,
                  X_stoch = FALSE) {

```

```
  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }
  else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}
  # Exogenous variables
  Exog_ts <- ts(as.data.frame(Exog), start = c(anno, period),
                frequency = freq)

  if (X_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0)) +
                              SSMcycle(c_period, Q = matrix(NA)) +
                              SSMregression(~Exog_ts, Q = matrix(0)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")
  }else
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 1, Q = matrix(0)) +
                               SSMcycle(c_period, Q = matrix(NA)) +
                               SSMregression(~Exog_ts, Q = matrix(NA)),
                    H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")}


  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
                      frequency = frequency(Serie_ts))

  # CREATION OF OUTPUT VARIABLE
  # Variable for forecast values
  estimates <- matrix(0, ncol = ncol(as.data.frame(Serie_ts)),
                      nrow = nrow(as.data.frame(Serie_ts)))
  for (j in 1:ncol(Exog_ts))
  {estimates <- estimates + dfSSKSIStates[,j]*Exog_ts[,j]}
  estimates <- estimates + dfSSKSIStates[, ncol(Exog_ts) + 1] +
               dfSSKSIStates[, ncol(Exog_ts) + 2]

  return(estimates)

}
```

```
#' @title The Trend Plus Cycle Model (TCM)
#' @description Function TCM applies the Trend Plus Cycle Model.
#' @details The Trend Plus Cycle Model assumed that the trend has a level
#'          and slope terms and
#'          include the cycle component.
#' @param Serie Time series object containing the data.
#' @param anno A integer, which specify the year of the first observation.
#' @param period A integer, which specify the period of the first
#'               observation.
#' @param freq  The number of observations per unit of time.
#' @param num_f Number of data to forecast.
```

```r
#' @param L_stoch Logical. If TRUE the level will be treated
#'                as stochastic (TRUE default).
#' @param S_stoch Logical. If TRUE the slope will be treated
#'                as stochastic (TRUE default).
#' @param c_period A integer, which specify the length of the cycle.
#' @return Time series containing estimates results
#' @export
#' @import KFAS
#' @examples
#' data <- AirPassengers
#' TCM(data, anno = 1949, period = 1, freq = 12, num_f = 6, c_period = 12)



TCM <- function(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
                S_stoch = TRUE, c_period) {

  # TIME SERIES STRUCTURE
  # Data to forecast
  if (num_f != 0) {
    Serie_ts <- as.data.frame(ts(Serie, start = c(anno, period),
                                 frequency = freq))
    Serie_ts <- ts(rbind(as.matrix(Serie_ts),
                         matrix(NA, ncol = ncol(Serie_ts), nrow = num_f)),
                   start = c(anno, period), frequency = freq)
  }
  else
  {Serie_ts <- ts(Serie, start = c(anno, period), frequency = freq)}

  if (L_stoch == TRUE & S_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                            matrix(NA))) +
                              SSMcycle(c_period, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0,0), method = "BFGS")
  } else if (L_stoch == TRUE & S_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(NA),
                                                            matrix(0))) +
                              SSMcycle(c_period, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == TRUE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                            matrix(NA))) +
                              SSMcycle(c_period, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0,0), method = "BFGS")
  } else if (L_stoch == FALSE & S_stoch == FALSE)
  {sski <- SSModel(Serie_ts ~ SSMtrend(degree = 2, Q = list(matrix(0),
                                                            matrix(0))) +
                     SSMcycle(c_period, Q = matrix(NA)),
                   H = matrix(NA))
  fitSSKI <- fitSSM(sski, inits = c(0,0), method = "BFGS")}


  out <- KFS(fitSSKI$model, filtering = "state")
  dfSSKSIStates <- ts(out$alphahat, start = time(Serie_ts)[1],
```

```
                        frequency = frequency(Serie_ts))

  estimates <- dfSSKSIStates[,1] + dfSSKSIStates[,3]

  return(estimates)

}
```

Package 'STSM'

September 2, 2019

Type Package

Title Structural Time Series Models

Version 0.1.0

Depends KFAS

Author Elisa Jorge González [aut, cre]

Maintainer Elisa Jorge González <elisajg0@gmail.com>

Description
R package that implement different structural models depending of unobserved components are combined and its characteristics.

License CPL (>=3)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

R topics documented:

`BSM` *The Basic Structural Model (BSM)*

Description

Function BSM applies the Basic Structural Model.

1

Usage

```
BSM(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
  S_stoch = TRUE, Sea_stoch = TRUE)
```

Arguments

Serie	Time series object containing the data.
anno	A integer, which specify the year of the first observation.
period	A integer, which specify the period of the first observation.
freq	The number of observations per unit of time.
num_f	Number of data to forecast.
L_stoch	Logical. If TRUE the level will be treated as stochastic (TRUE default).
S_stoch	Logical. If TRUE the slope will be treated as stochastic (TRUE default).
Sea_stoch	Logical. If TRUE the seasonal will be treated as stochastic (TRUE default).

Details

The Basic Structural Model is the local linear trend model with a seasonal term.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
BSM(AirPassengers, anno = 1949, period = 1, freq = 12, num_f = 6)
```

BSM_X *The Basic Structural Model with the inclusion of time-varying exogenous variables (TVP)*

Description

Function BSM_X applies the Basic Structural Mosel with the inclusion of exogenous variables.

Usage

```
BSM_X(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
  S_stoch = TRUE, Sea_stoch = TRUE, X_stoch = FALSE)
```

Arguments

Serie	Time series object containing the data.
Exog	Time series object or set of time series containing the exogenous variables.
anno	A integer, which specify the year of the first observation.
period	A integer, which specify the period of the first observation.
freq	The number of observations per unit of time.

`num_f`	Number of data to forecast.
`L_stoch`	Logical. If TRUE the level will be treated as stochastic (TRUE default).
`S_stoch`	Logical. If TRUE the slope will be treated as stochastic (TRUE default).
`Sea_stoch`	Logical. If TRUE the seasonal will be treated as stochastic (TRUE default).
`X_stoch`	Logical. If TRUE the exogenous variable will be treated as stochastic (FALSE default).

Details

Function include time-varying exogenous variables in the basic structural model to improve the predictions.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
vexog <- nottem[1:(length(AirPassengers)+6)]
forecasts <- BSM_X(Serie = data, Exog = vexog, anno = 1949, period = 1, freq = 12, num_f = 6,
                   L_stoch = TRUE, S_stoch = TRUE, Sea_stoch = TRUE, X_stoch = TRUE)
```

`CNM` *The Cycle Plus Noise Model (CNM)*

Description

Function CNM applies the Cycle Plus Noise Model.

Usage

```
CNM(Serie, anno, period, freq, num_f = 0, c_period)
```

Arguments

`Serie`	Time series object containing the data.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`c_period`	A integer, which specify the length of the cycle.

Details

The Cycle Plus Noise Model assumed that the trend be constant and include the cycle component.

Value

Time series containing estimates results

Examples

```
data <- AirPassengers
CNM(data, anno = 1949, period = 1, freq = 12, num_f = 6, c_period = 12)
```

`CNM_X` *The Cycle Plus Noise Model with the inclusion of exogenous variables*

Description

Function CNM_X applies the Cycle Plus Noise Model with the inclusion of exogenous variables.

Usage

```
CNM_X(Serie, Exog, anno, period, freq, num_f = 0, c_period,
  X_stoch = FALSE)
```

Arguments

`Serie`	Time series object containing the data.
`Exog`	Time series object or set of time series containing the exogenous variables.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`c_period`	A integer, which specify the length of the cycle.
`X_stoch`	Logical. If TRUE the exogenous variable will be treated as stochastic (FALSE default).

Details

Function include exogenous variables in the cycle plus noise model to improve the predictions.

Value

Time series containing estimates results

Examples

```
data <- AirPassengers
vexog <- nottem[1:(length(AirPassengers)+6)]
forecasts <- CNM_X(data, vexog, anno = 1949, period = 1, freq = 12, num_f = 6,
                  c_period = 12, X_stoch = FALSE)
```

`LLM` *The Local Level Model (LLM)*

Description

Function LLM applies the Local Level Model.

Usage

```
LLM(Serie, anno, period, freq, num_f = 0, stochastic = TRUE)
```

Arguments

`Serie`	Time series object containing the data.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`stochastic`	Logical. If TRUE the level will be treated as stochastic.

Details

The local level model allowing that the level of the series vary in time dependign of the level of the previous time plus a random element.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
LLM(data, anno = 1949, period = 1, freq = 12, num_f = 6, stochastic = TRUE)
```

`LLM_X` *The Local Level Model with the inclusion of exogenous variables*

Description

Function LLM_X applies the Local Level Model with exogenous variables.

Usage

```
LLM_X(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
  X_stoch = FALSE)
```

Arguments

`Serie`	Time series object containing the data.
`Exog`	Time series object or set of time series containing the exogenous variables.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`L_stoch`	Logical. If TRUE the level will be treated as stochastic (TRUE default).
`X_stoch`	Logical. If TRUE the exogenous variable will be treated as stochastic (FALSE default).

Details

Function include exogenous variables in the local level model to improve the predictions.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
vexog <- nottem[1:(length(AirPassengers)+6)]
forecast <- LLM_X(Serie = data, Exog = vexog, anno = 1949, period = 1, freq = 12, num_f = 6,
                  L_stoch =TRUE, X_stoch = TRUE)
```

`LLTM` *The Local Linear Trend Model (LLTM)*

Description

Function LLTM applies the Local Linear Trend Model.

Usage

```
LLTM(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
  S_stoch = TRUE)
```

Arguments

`Serie`	Time series object containing the data.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`L_stoch`	Logical. If TRUE the level will be treated as stochastic (TRUE default).
`S_stoch`	Logical. If TRUE the slope will be treated as stochastic (TRUE default).

Details

The local linear trend model is the local level model adding a slope term, which is generated by a random wal.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
LLTM(data, anno = 1949, period = 1, freq = 12, num_f = 6)
```

`LLTM_X` *The Local Linear Trend Model with the inclusion of exogenous variables*

Description

Function LLTM_X applies the Local Linear Trend Model with exogenous variables.

Usage

```
LLTM_X(Serie, Exog, anno, period, freq, num_f, L_stoch = TRUE,
  S_stoch = TRUE, X_stoch = FALSE)
```

Arguments

`Serie`	Time series object containing the data.
`Exog`	Time series object or set of time series containing the exogenous variables.
`anno`	A integer, which specify the year of the first observation.
`period`	A integer, which specify the period of the first observation.
`freq`	The number of observations per unit of time.
`num_f`	Number of data to forecast.
`L_stoch`	Logical. If TRUE the level will be treated as stochastic (TRUE default).
`S_stoch`	Logical. If TRUE the slope will be treated as stochastic (TRUE default).
`X_stoch`	Logical. If TRUE the exogenous variable will be treated as stochastic (FALSE default).

Details

Function include exogenous variables in the local linear trend model to improve the predictions.

Value

Time series containing estimates results.

Examples

```
data <- AirPassengers
vexog <- nottem[1:(length(AirPassengers)+6)]
forecast <- LLTM_X(Serie = data, Exog = vexog, anno = 1949, period = 1, freq = 12, num_f = 6,
                  L_stoch = TRUE, S_stoch = FALSE, X_stoch = FALSE)
```

TCM *The Trend Plus Cycle Model (TCM)*

Description

Function TCM applies the Trend Plus Cycle Model.

Usage

```
TCM(Serie, anno, period, freq, num_f = 0, L_stoch = TRUE,
  S_stoch = TRUE, c_period)
```

Arguments

Serie	Time series object containing the data.
anno	A integer, which specify the year of the first observation.
period	A integer, which specify the period of the first observation.
freq	The number of observations per unit of time.
num_f	Number of data to forecast.
L_stoch	Logical. If TRUE the level will be treated as stochastic (TRUE default).
S_stoch	Logical. If TRUE the slope will be treated as stochastic (TRUE default).
c_period	A integer, which specify the length of the cycle.

Details

The Trend Plus Cycle Model assumed that the trend has a level and slope terms and include the cycle component.

Value

Time series containing estimates results

Examples

```
data <- AirPassengers
TCM(data, anno = 1949, period = 1, freq = 12, num_f = 6, c_period = 12)
```

Bibliography

[Andrawis et al. (2011)] Andrawis RR, Atiya AF and El-Shishiny H (2011) Combination of long term and short term forecast, with application to tourism demand forecasting. International Journal of Forecasting 27(3): 870–886.

[AENA (2017)] AENA (2017) Available at: http://www.aena.es/es/pasajeros/pasajeros.html (accessed 06 July 2018).

[Akaike (1974)] Akaike H (1974) A new look at statistical model identification. IEEE Transactions on Automatic Control 19(6): 716–723.

[Antoniou and Yannis (2013)] Antoniou C and Yannis G (2013) State-space based analysis and forecasting of macroscopic road safety trends in Greece. Accident Analysis & Prevention 60: 268–76. doi:10.1016/j.aap.2013.02.039.

[Armstrong and Collopy (1992)] Armstrong JS and Collopy F (1992) Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting 8(1): 69–80.

[Athanasopoulos and de Silva (2012)] Athanasopoulos G and de Silva A (2012) Multivariate exponential smoothing for forecasting tourist arrivals. Journal of Travel Research 51(5): 640–652.

[Athanasopoulos et al. (2011)] Athanasopoulos G, Hyndman RJ, Song H, et al. (2011) The tourism forecasting competition. International Journal of Forecasting 27: 822–844.

[Baker (2014)] Baker DMcA (2014) The effects of terrorism on the travel and tourism industry. International Journal of Religious Tourism and Pilgrimage 2(1): Article 9. doi: 10.21427/D7VX3D

[Box and Jenkins (1970)] Box G and Jenkins G (1970) Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

[Bok et al. (2016)] Box GEP, Jenkins GM, Reinsel GC, et al. (2016) Time Series Analysis: Forecasting and Control. 5th ed. Hoboken: Wiley.

[Brockwell and Davis (2002)] Brockwell PJ and Davis RA (2002) Introduction to Time Series and Forecasting. 2nd ed. New York: Springer.

[Burger et al. (2001)] Burger CJSC, Dohnal M, Kathrada M, et al. (2001) A practitioners guide to time-series methods for tourism demand forecasting – a case study of Durban, South Africa. Tourism Management 22(4): 403–409.

[Burnham and Anderson (2002)] Burnham KP and Anderson DR (2002) Model Selection and Inference: A practical Information-Theoretic Approach. 2nd ed. New York: Springer-Verlag.

[Casals et al. (2016)] Casals J, Garcia-Hiernaux A, Jerez M, et al. (2016) State-Space Methods for Time Series Analysis. New York: Chapman and Hall/CRC.

[Chan et al. (2005)] Chan F, Lim C and McAleer MJ (2005) Modelling multivariate international tourism demand and volatility. Tourism Management 26(3): 459–471.

[Chen et al. (2019)] Chen JL, Li G, Wu DC, et al. (2019) Forecasting seasonal tourism demand using a multi-series structural time series method. Journal of Travel Research 58(1): 92–103.

[Cho (2003)] Cho V (2003) A comparison of three different approaches to tourist arrivals forecasting. Tourism Management 24(3): 323–330.

[Chu (2004)] Chu FL (2004) Forecasting tourism demand: a cubic polynomial approach. Tourism Management 25: 209–218.

[Clements and Hendry (1993)] Clements MP and Hendry DF (1993) On the limitations of comparing mean square forecast errors. Journal of Forecasting 12(8): 617–637.

[Commandeur et al. (2013)] Commandeur JJF, Bijleveld FD, Bergel-Hayat R, Antoniou C, Yannis G, Papadimitriou E (2013) On statistical inference in time series analysis of the evolution of road safety. Accident Analysis and Prevention 60: 424-434.

[Commandeur and Koopman (2007)] Commandeur JJF and Koopman SJ (2007) An Introduction to State Space Time Series Analysis. New York: Oxford University Press.

[Commandeur et al. (2011)] Commandeur JJ F, Koopman SJ, and Ooms M (2011) Statistical software for state space methods. Journal of Statistical Software 41 (1):1–18.

[Conte et al. (1986)] Conte SD, Dunsmore HE and Shen VY (1986) Software engineering metrics and models. Menlo Park, CA: Benjamin/Cummings.

[Crouch (1994)] Crouch GI (1994) The study of international tourism demand: a review of practice. Journal of Travel Research 33: 41–54.

[De Gooijer and Hyndman (2006)] De Gooijer JG and Hyndman RJ (2006) 25 years of time series forecasting. International Journal of Forecasting 22 (3): 443–73. doi:10.1016/j.ijforecast.2006.01.001.

[Montgomery et al. (2015)] Montgomery DC, Jennings CL and Kulahci M (2015) Introduction to Time Series Analysis and Forecasting. Canada: Wiley Series in Probability and Statistics.

[Du Preez and Witt (2003)] Du Preez J and Witt SF (2003) Univariate versus multivariate time series forecasting: an application to international demand. International Journal of Forecasting 19: 435–451.

[Durbin (2000)] Durbin J (2000) The state space approach to time series analysis and its potential for official statistics. Australian and New Zealand J. of Statistics 42: 1-23.

[Durbin and Koopman (2012)] Durbin J and Koopman SJ (2012) Time Series Analysis by State-Space Methods. Oxford: Oxford University Press.

[EUROSTAT (2017a)] EUROSTAT (2017a) DATABASE. Available at: http://ec.europa.eu/eurostat/data/database (accessed 02 September 2018).

[EUROSTAT (2017b)] EUROSTAT (2017b) Calendar Adjustment. Available at: $http : //ec.europa.eu/eurostat/statistics - explained/index.php/\ Glossary : Calendar - adjustment$ (accessed 10 September 2018).

[Fildes (1992)] Fildes R (1992) The evaluation of extrapolative forecasting methods. International Journal of Forecasting 8(1): 81–98.

[$FRONTUR - forecast$ (2019)] $FRONTUR - forecast$ (2019) Forecast tourist movements in the border of the Canary Islands (FRONTUR). Available at: $https : //github.com/ISTAC/FRONTUR_forecast$ (accessed 9 September 2019).

[Garcia-Ferrer and Queralt (1997)] Garcia-Ferrer A and Queralt RA (1997) A note on forecasting international tourism demand in Spain. International Journal of Forecasting 13(4): 539–549.

[Garín-Muñoz (2011)] Garín-Muñoz T (2011) La demanda de turismo británico en España. Boletín Económico del ICE 3010: 49–62.

[Garín-Muñoz (2006)] Garín-Muñoz T (2006) Inbound international tourism to Canary Islands: a dynamic panel data model. Tourism Management 27(2): 281–291.

[Gil-Alana (2005)] Gil-Alana LA (2005) Modelling international monthly arrivals using seasonal univariate long-memory processes. Tourism Management 26(6): 867–878.

[Gil-Alana et al. (2008)] Gil-Alana LA, Cunado J and Perez F (2008) Tourism in the Canary Islands: forecasting using several time series models. Journal of Forecasting 27(7): 621–636.

[GNTB (2016)] GNTB (2016) German National Tourist Board Annual Report. German National Tourist Board. Available at: http://www.germany.travel/en/germany/about-us/annual-report/annual-report.html (accessed 03 July 2018).

[González and Moral (1995)] González P and Moral P (1995) An analysis of tourism the international tourism demand in Spain. International Journal of Forecasting 11: 233–251.

[González and Moral (1996)] González P and Moral P (1996) Analysis of tourism trend in Spain. Annals of Tourism Research 23: 739–754.

[Greenidge (2001)] Greenidge K (2001) Forecasting tourism demand: an STM approach. Annals of Tourism Research 28(1): 98–112.

[Greenidge and Jackman (2010)] Greenidge, K Jackman, M (2010). Modelling and Forecasting Tourist Flows to Barbados Using Structural Time Series Models. Tourism and Hospitality Research, January.

[Guizzardi and Mazzocchi (2010)] Guizzardi A and Mazzocchi M (2010) Tourism demand for Italy and the business cycle. Tourism Management 31: 367–377.

[Gunter and Önder (2015)] Gunter U and Önder I (2015) Forecasting international city tourism demand for Paris: accuracy of uni-and multivariate models employing monthly data. Tourism Management 46: 123–135.

[Hannan and Deistler (1988)] Hannan EJ and Deistler M (1988). The Statistical Theory of Linear Systems. John Wiley, New York

[Harrison and Stevens (1971)] Harrison PJ and Stevens CF (1971) A Bayesian approach to short-term forecasting. Journal of the Operational Research Society 22(4): 341–362.

[Harrison and Stevens (1976)] Harrison PJ and Stevens CF (1976) Bayesian forecasting. Journal of the Royal Statistical Society 38(3): 205–247.

[Harvey (1989)] Harvey AC (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge: Cambridge University Press.

[Harvey (1985)] Harvey AC (1985) Trends and cycles in macroeconomic time series. Journal of Business and Economic Statistics 3 (3): 216–27. doi: 10.2307/1391592.

[Harvey and Koopman (1996)] Harvey AC and Koopman SJ (1996) Multivariate structural time series model. London: Suntory-Toyota International Centre for Economics and Related Disciplines.

[Harvey et al. (1997)] Harvey DI, Leybourne SJ and Newbold P (1997) Testing the equality of prediction mean squared errors. International Journal of Forecasting 13: 281–291.

[Harvey and Shepard (1993)] Harvey AC and Shepard N (1993) Structural time series models. In Vol. 11 of handbook of statistics, ed. Maddala GS, Rao CR and Vinod HD. Amsterdam: Elsevier Science Publishers BV, 261–302.

[Harvey et al. (2004)] Harvey AC, Shepard N and Koopman SJ (2004) State space and unobserved component models: Theory and applications. Cambridge: Cambridge University Press.

[Hassani et al. (2017)] Hassani H, Silva ES, Antonakakis N, et al. (2017) Forecasting accuracy evaluation of tourist arrivals. Annals of Tourism Research 63: 112–127.

[Hoti et al. (2004)] Hoti S, Leon C and McAleer M (2004) International tourism demand and volatility models for the Canary Islands. In: 2nd meeting of the international environment modelling and software society, vol. 3, pp.1345–1350.

[INE (2016)] INE (2016) España en cifras. Instituto Nacional de Estadística. Available at: $http : //www.ine.es/ss/Satellite?L = es_E S\&c = INEPublicacion_C\&cid =$ 1259924856416&p = 1254735110672&$pagename = ProductosYServicios\%2FPYSLayout\&param1 = PYSDetalleGratuitas$ (accessed 10 September 2018).

[IMPACTUR (2016] IMPACTUR (2016) Estudio del impacto económico del turismo sobre la economía y el empleo de las Islas Canarias. Exceltur. Available at: http://www.exceltur.org/impactur/ (accessed 01 September 2018).

[ISTAC (2017a)] ISTAC (2017a) Encuesta de Alojamiento Turístico. Instituto Canario de Estadística. Available at: $http://www.gobiernodecanarias.org/istac/temas_estadisticos/sectorservicios/$ hosteleriayturismo/oferta/C00065A.html (accessed 10 September 2018).

[ISTAC (2017b)] ISTAC (2017b) Recopilación de Estadísticas de Transporte Aéreo. Instituto Canario de Estadística. Available at: $http://www.gobiernodecanarias.org/istac/temas_estadisticos/sectorservicios/$ hosteleriayturismo/demanda/C00017A.html (accessed 10 September 2018).

[Jiao and Chen (2019)] Jiao EX and Chen JL (2019) Tourism forecasting: a review of methodological developments over the last decade. Tourism Economics 25: 469–492.

[Jorge-González et al. (2019a)] Jorge-González E, González-Dávila E, Martín-Rivero R and Lorenzo-Díaz D (2019) Stochastic and deterministic trend in state space models. Communications in Statistics - Simulation and Computation. doi: 10.1080/03610918.2019.1606240.

[Jorge-González et al. (2019b)] Jorge-González E, González-Dávila E, Martín-Rivero R and Lorenzo-Díaz D (2019) Univariate and multivariate forecasting of tourism demand using state-space models. Tourism Economics. doi: 10.1177/1354816619857641.

[Kalman (1960)] Kalman RE (1960) A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, Transactions ASMA. Series D, 82, 35-45.

[Kim et al. (2018)] Kim CS, Bai BH, Kim PB, et al. (2018) Review of reviews: A systematic analysis of review papers in the hospitality and tourism literature. International Journal of Hospitality Management 20: 49–58.

[Kim and Moosa (2001)] Kim JH and Moosa I (2001) Seasonal behaviour of monthly international tourist flows: Specifications and implications for forecasting models. Tourism Economics 7(4): 381–396.

[Kitagawa (1998)] Kitagawa G (1998) A self-organizing sate-space model. Journal of the American Statistical Association 93 (443): 1203–15. doi:10.2307/2669862.

[Koc and Altinay (2007)] Koc E and Altinay G (2007) An analysis of seasonality in monthly per person tourist spending in Turkish inbound tourism from a market segmentation perspective. Tourism Management 28: 227–237.

[Koopman (1992)] Koopman SJ (1992) Diagnostic checking and intra-daily effects in times series models. Tinbergen Institute research series 27.

[Koopman et al. (2017)] Koopman SJ, Commandeur JF, Bijleveld FD and Vujic S (2017) Continuous time state space modelling with an application to high-frequency road traffic data. Workink paper, July 29, Vrije Universiteit Amsterdam.

[Koopman et al. (2010)] Koopman SJ, Harvey AC, Doornik JA, et al. (2010) Stamp 8.3: Structural Time Series Analyser, Modeller and Predictor. London: Timberlake Consultants.

[Koopman and Ooms (2006)] Koopman SJ and Ooms M (2006) Forecasting daily time series using periodic unobserved components time series models. Computational Statistics and Data Analysis 51 (2): 885–903. doi:10.1016/j.csda.2005.09.009.

[Kulendran and Shan (2002)] Kulendran N and Shan J (2002) Forecasting China's monthly inbound travel demand. Journal of Travel and Tourism Marketing 13(1/2): 5–19.

[Kulendran and Witt (2001)] Kulendran N and Witt SF (2001) Cointegration versus least squares regression. Annals of Tourism Research 28: 291–311.

[Kulendran and Witt (2003)] Kulendran N and Witt SF (2003) Forecasting the demand for international business tourism. Journal of Travel Research 41(3): 265–271.

[Kulendran and Wong (2005)] Kulendran N and Wong KKF (2005) Modelling seasonality in tourism forecasting. Journal of Travel Research 44(2): 163–170.

[Landaluce-Calvo (2017)] Landaluce-Calvo MI (2017) Multivariate exploratory approach for the analysis of temporal structures. An application to the evolution of the tourist demand in Spain. Estudios de Economía Aplicada 35(2): 465–491.

[Lawson et al. (2011)] Lawson AR, Ghosh B, and Broderick B (2011) Prediction of traffic-related nitrogen oxides concentrations using Structural Time-Series models. Atmospheric Environment 45 (27): 4719–27. doi:10.1016/j.atmosenv.2011.04.053.

[Li et al. (2005)] Li G, Song H and Witt SF (2005) Recent developments in econometric modelling and forecasting. Journal of Travel Research 44(1): 82–99.

[Lim (1999)] Lim C (1999) A meta analysis review of international tourism demand. Journal of Travel Research 37: 273–284.

[Lim (1997)] Lim C (1997). Review of International Tourism Demand Models. Annals of Tourism Research, 24, 835-849.

[Lim and McAleer (2001)] Lim C and McAleer M (2001) Monthly seasonal variations: Asian tourism to Australia. Annals of Tourism Research 28: 68–82.

[López et al. (2017)] López AM, Flores MA and Sánchez JI (2017) Forecasting of Spanish passenger air traffic based on time series models with aggregated and disaggregated approaches. Estudios de Economía Aplicada 35(2): 395–418.

[Martin and Murillo (2005)] Martin G and Murillo C (2005) High-frequency health data and spline functions. Statistics in Medicine 24: 967–81.

[Martin and Witt (1987)] Martin CA, Witt SF (1987). Tourism Demand Forecasting Models: Choice of Appropriate Variable to Represent Tourists' Cost of Living. Tourism Management 8: 233-246.

[Medeiros et al. (2008)] Medeiros MC, McAleer M, Slottje D, et al. (2008) An alternative approach to estimating demand: neural network regression with conditional volatility for high frequency air passenger arrivals. Journal of Econometrics 147: 372–383.

[Nghiem et al. (2016)] Nghiem S, Commandeur JF and Connelly LB (2016) Determinants of road traffic safety: New evidence from Australia using state-space analysis. Accident Analysis and Prevention 94: 65–72. doi:10.1016/j.aap.2016.05.012.

[ONS (2016)] ONS (2016) Travel trends. Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/articles/traveltrends/2016 (accessed 01 September 2018).

[Peña (2010)] Peña D (2010) Análisis de series temporales. Madrid: Alianza Editorial.

[Petri et al. (2009)] Petris G, Petrone S and Campagnoli P (2009) Dynamic linear models with R. New York: Springer-Verlag.

[Pfeffermann and Allon (1989)] Pfeffermann D and Allon J (1989) Multivariate exponential smoothing: methods and practice. International Journal of Forecasting 5: 83–98.

[Poinelli et al. (2015)] Poinelli E, Buffa P and Pavlidis N (2015) Spain − the tourism sector is heading the recovery. HVS consulting. Available at: $http://www.hvs.com/article/7226/in-focus-spain-the-tourism-sector-is-heading-the-recovey$ (accessed 01 September 2018).

[Prideaux et al. (2003)] Prideaux B, Laws E and Faulkner B (2003) Events in Indonesia: exploring the limits to formal tourism trends forecasting methods in complex crisis situations. Tourism Management 24(4): 475–487.

[Rossello-Nadal (2001)] Rossello-Nadal J (2001) Forecasting turning points in international visitor arrivals in the Balearic Islands. Tourism Economics 7: 365–380.

[Saayman and Botha (2017)] Saayman A and Botha I (2017) Non-linear models for tourism demand forecasting. Tourism Economics 23(3): 594–613.

[Silva et al. (2017)] Silva ES, Ghodsi Z, Ghodsi M, et al. (2017) Cross country relations in European tourist arrivals. Annals of Tourism Research 63: 151–168.

[Song (2011)] Song H (2011) Tourism forecasting: an introduction. International Journal of Forecasting 27: 817–821.

[Song et al. (2012)] Song H, Dwyer L, Li G, et al. (2012) Tourism economics research: a review and assessment. Annals of Tourism Research 39(3): 1653–1682.

[Song and Li (2008)] Song H and Li G (2008) Tourism demand modelling and forecasting: a review of recent research. Tourism Management 29(2): 203–220.

[Song et al. (2010)] Song H, Li G, Witt SF, and Fei B (2010) Tourism Demand Modelling and Forecasting: How Should Demand Be Measured? Tourism Economics 16: 63-81. doi: 10.5367/000000010790872213

[Song et al. (2011)] Song H, Li G, Witt SF, et al. (2011) Forecasting tourist arrivals using time-varying parameter structural time series models. International Journal of Forecasting 27: 855–869.

[Song and Witt (2000)] Song H and Witt SF (2000) Tourism Demand Modelling and Forecasting: Modern Econometric Approaches. Oxford: Pergamon. Journal of Travel Research 42: 57-64.

[Song and Wong (2003)] Song H, Wong KF (2003) Tourism demand modeling: A time-varying parameter approach. Journal of Travel Research 42: 57-64.

[Song et al. (2003a)] Song H, Wong KF and Chon KS (2003a) Modelling and forecasting demand for Hong Kong tourism. International Journal of Hospitality Management 22: 435–451.

[Song et al. (2003b)] Song H, Witt SF and Jensen TC (2003b) Tourism forecasting: Accuracy of alternative econometric models. International Journal of Forecasting 19: 123–141.

[Schwarz G (1978)] Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6 (2): 461–4. doi: 10.1214/aos/1176344136.

[Travel and Tourism (2018)] Travel and Tourism (2018) Travel & Tourism Economic Impact World 2018. World Travel & Tourism Council. Available at: $https : //www.wttc.org/economic - impact/country - analysis/regional - reports/\#undefined$ (accessed 10 September 2018).

[Tsui et al. (2014)] Tsui WHK, Balli HO, Gilbey A, et al. (2014) Forecasting of Hong Kong airport's passenger throughput. Tourism Management 42: 62–76.

[Tsui and Balli (2017)] Tsui WHK and Balli F (2017) International arrivals forecasting for Australian airports and the impact of tourism marketing expenditure. Tourism Economics 23: 403–428.

[Turner and Witt (2001)] Turner LW and Witt SF (2001) Forecasting tourism using univariate and multivariate structural time series model. Tourism Economics 7(2): 135–147.

[Wang and Getz (2007)] Wang G and Getz LL (2007) State-space models for stochastic and seasonal fluctuations of vole and shrew populations in east-central Illinois. Ecological Modelling 207 (2–4): 189–96. doi: 10.1016/j.ecolmodel.2007.04.026.

[West and Harrison (1997)] West M and Harrison PJ (1997) Bayesian forecasting and dynamic models. 2nd ed. New York: Springer.

[Witt and Witt (1995)] Witt SF and Witt CA (1995) Forecasting tourism demand: a review of empirical research. International Journal of Forecasting 11: 447–475.

[Wong et al. (2007)] Wong K, Song H, Witt SF, et al. (2007) Tourism forecasting: to combine or not combine? Tourism Management 28: 1068–1078.

[World Tourism Organization (2019] World Tourism Organization (2019) International Tourism Highlights, 2019 Edition. UNWTO. Madrid. doi: https://doi.org/10.18111/9789284421152.

[Wu et al. (2012)] Wu DC, Li G and Song H (2012) Economic analysis of tourism consumption dynamics: a time-varying parameter demand system approach. Annals of Tourism Research 39(2): 667–685.

[Wu et al. (2017)] Wu DC, Song H and Shen S (2017) New developments in tourism and hotel demand modeling and forecasting. International Journal of Contemporary Hospitality Management 29(1): 507–529.

[Young et al. (1991)] Young PC, Ng CN, Lane K and Parker D (1991) Recursive forecasting, smoothing and seasonal adjustment of nonstationary environmental data. Journal of Forecasting 10 (1–2): 57–89. doi: 10.1002/for.3980100105.

[Zhou-Grundy and Turner (2014)] Zhou-Grundy Y and Turner LW (2014) The challenge of regional tourism demand forecasting: the case of China. Journal of Travel Research 53: 747–759.